

PROFILING PATTERNS IN HEALTHCARE USING AN ENSEMBLE
MODEL FRAMEWORK TO PREDICT EMPLOYEE HEALTH RISKS

NICHOLAS CHAN KHIN WHAI

DISSERTATION SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN
INFORMATION SYSTEMS

SCHOOL OF ENGINEERING AND TECHNOLOGY
SUNWAY UNIVERSITY
MALAYSIA

2024

PROFILING PATTERNS IN HEALTHCARE USING AN ENSEMBLE MODEL FRAMEWORK TO PREDICT EMPLOYEE HEALTH RISKS

ABSTRACT

In the current evolution of the digital world, data has become the cornerstone of decision-making processes, shaping industries and societies alike. The exponential growth of data, commonly referred to as big data, has sparked a surge in interest in advanced analytics techniques to harness its potential. Among these techniques, big data analytics, particularly in healthcare, holds immense promise for understanding overall population health and predicting high-risk and high-cost individuals. This thesis delves into the realm of healthcare analytics in Malaysia, focusing on the analysis of extensive medical data to identify patterns and insights that can aid in the identification of high-risk and high-cost individuals. The objectives of this research are: first, to uncover and comprehend usage patterns within healthcare claims data, elucidating factors contributing to the identification of high-risk individuals; second, to propose an innovative ensemble stacking model approach; and third, to demonstrate the efficacy of this approach in enhancing predictive accuracy. The proposed ensemble stacking model integrates the Stacking technique with hybrid feature selection and feature engineering methodologies. By amalgamating multiple predictive models into a cohesive framework, the ensemble model offers superior predictive accuracy compared to traditional single-model approaches. Furthermore, the model's versatility enables its application across various classification tasks within the healthcare domain. Through empirical analysis, this research highlights the enhanced predictive accuracy and efficacy of the ensemble model framework. Notably, key features such as ICD Category, TotalRemainingAmt, and TotalAmtInsured emerge as significant contributors to determining an individual's risk profile based on their medical claim

patterns and behaviours. By leveraging big data analytics and ensemble modelling techniques, this research contributes to the advancement of predictive analytics in healthcare, offering valuable insights for decision-makers and stakeholders in the industry.

Keywords: Big Data Analytics, Healthcare, Ensemble Model, Stacking Model, Predictive Analysis

(278 words)

Acknowledgements

The completion of this thesis would not have been possible without the dedicated guidance and assistance of several individuals who generously contributed to its preparation and culmination. First and foremost, I extend my heartfelt gratitude to Prof. Dr. Angela Lee for her unwavering support and mentorship throughout my thesis journey. Her boundless energy and enthusiasm served as a constant source of motivation, inspiring me to strive for excellence. I am grateful for her accessibility and willingness to address my queries and challenges, which significantly enriched my research experience. I am also indebted to Assoc. Prof. Dr. Zuraini Binti Zainol from the National Defense University of Malaysia for her invaluable comments, suggestions, and feedback during the review process of my thesis. Her expertise and guidance played a pivotal role in validating the rigor of my research, for which I am deeply appreciative. To my parents, Alan, and Jaime, I extend my sincere appreciation for their unwavering support and encouragement throughout this academic endeavour. Their constant presence and words of encouragement were instrumental in keeping me focused and motivated. I am profoundly grateful for their belief in my abilities and unwavering support. Finally, I would like to express my heartfelt gratitude to my partner, Claudia, for her unwavering support and understanding. Her unwavering presence and encouragement sustained me through the challenges and triumphs of this journey. I am deeply grateful for her unwavering support and companionship. To each of these individuals, I extend my sincerest thanks for their invaluable contributions and unwavering support, without which this thesis would not have been possible.

Table of Contents

ABSTRACT	i
<i>Acknowledgements</i>	iii
List of Figures	vi
List of Tables	ix
1. Introduction	1
1.1 Background of Research.....	1
1.2 Problem Statement.....	3
1.2.1 Research Questions	4
1.2.2 Research Objectives	5
1.2.3 Research Contribution	5
1.3 Research Motivation.....	7
1.4 Thesis Outline.....	8
2. Literature Review	10
2.1 Background of Research.....	10
2.2 Limitation of Previous Research	14
2.3 Overview of Big Data Analytics in Healthcare	15
2.3.1 Overview of Big Data Analytics in Healthcare within the context of Malaysia.....	17
2.4 Data Mining.....	21
2.4.1 Techniques in Data Mining	23
2.5 Ensemble Model Frameworks	28
2.6 Systematic Review of Healthcare Data Analytics	33
3. Research Methodology	41
3.1 Proposed Framework for Ensemble Predictive Modelling	42
3.2 Phase 1: Objective	42
3.3 Phase 2: Data Preparation.....	46
3.4 Phase 3: Model Development.....	58
3.5 Phase 4: Model Blending.....	61
4. Analysis (Descriptive and Predictive)	63
4.1 Descriptive Analysis - Healthcare Data.....	63
4.2 Predictive Analysis	95
4.3 Robustness Testing (Healthcare Data and 3 Case Studies).....	107
5. Summary and Discussion	119
6. Conclusion	130
6.1. Research Contribution	131
6.2. Research Outcome	133
6.3. Limitations and Future Work.....	135
6.4. Recommendations	136
References	138

Appendix	147
<i>List of Publications and Papers Presented</i>	<i>147</i>
<i>Research Ethics Approval Letter.....</i>	<i>148</i>
<i>Description of Datasets</i>	<i>149</i>
<i>Starting up SAS Enterprise Miner</i>	Error! Bookmark not defined.
<i>Creating New Project.....</i>	Error! Bookmark not defined.
<i>Creating Library</i>	Error! Bookmark not defined.
<i>Creating Data Source.....</i>	Error! Bookmark not defined.
<i>Creating Diagram</i>	Error! Bookmark not defined.
<i>Drill Down Analysis by Age Group (SP).....</i>	<i>159</i>

List of Figures

Figure 1: Malaysia Healthcare Industry Increasing Medical Expenditure (March 2018).....	8
Figure 2: Malaysia Healthcare Industry Increasing Medical Expenditure (December 2018).....	8
Figure 3: CRISP-DM (Cross-Industry Standard Process for Data Mining) Methodology	22
Figure 4: Prisma Flow Diagram of the Literature Screening Process	35
Figure 5: Proposed Practical Framework for Ensemble Model Building.....	42
Figure 6: Feature Selection Process	54
Figure 7: Ensemble Method Framework (Combination of Regression + Decision Tree).....	60
Figure 8: Ensemble Method Framework (Combination of Decision Tree + Decision Tree)	60
Figure 9: Basic Demographics of GP Patients 2018.....	63
Figure 10: Basic Demographics of GP Patients by Relationship 2018.....	64
Figure 11: Top 10 Diagnoses under GP Claims 2018	65
Figure 12: Diagnoses of GP Claims based on Patient Relationship (E) 2018.....	66
Figure 13: Diagnoses of GP Claims based on Patient Relationship (SP) 2018.....	67
Figure 14: Diagnoses of GP Claims based on Patient Relationship (C) 2018.....	68
Figure 15: GP Claim Frequency by Month 2018'	69
Figure 16: Highest Diagnoses based on GP Claims 2018	70
Figure 17: Chronic Condition Diagnosis based on GP Claims 2018	70
Figure 18: Chronic Condition Diagnosis 2 based on GP Claims 2018	71
Figure 19: Demographics of Patients (Employees) GP Claims (M) 2018	72
Figure 20: Demographics of Patients (Employees) GP Claims (F) 2018	73
Figure 21: Number of Patients (Employees) based on Total Remaining Amount < 1000	77
Figure 22: Business Industry and Top 20 Diagnosis based on Total Remaining Amount < 1000	78
Figure 23: Total Amount Insured Range based on Top 20 Diagnosis under GP 2018	79
Figure 24: Basic Demographics of SP Patients 2018	79
Figure 25: Basic Demographics of SP Patients by Relationship 2018	80
Figure 26: Top 10 Diagnoses under SP Claims 2018	81
Figure 27: Demographics of Patients (Employees) SP Claims 2018	81
Figure 28: Demographics of Patients (Employees) SP Claims (M) 2018	82
Figure 29: Demographics of Patients (Employees) SP Claims (F) 2018	83
Figure 30: Basic Demographics of IP Patients 2018	84
Figure 31: Diagnoses under IP Claims 2018	85
Figure 32: Demographics of Patients (Employees) IP Encounters 2018	86
Figure 33: Diagnoses of IP Encounters (M) 2018	87
Figure 34: Diagnoses of IP Encounters (F) 2018	87
Figure 35: Comparison of Age Group between (Below 40) and (Above 40) - GP	88
Figure 36: Comparison of (M) Age Group between (Below 40) and (Above 40) - GP	89
Figure 37: Comparison of (F) Age Group between (Below 40) and (Above 40) - GP	89
Figure 38: Comparison of Age Group between (Below 40) and (Above 40) - Employee; GP	90
Figure 39: Comparison of (M) Age Group between (Below 40) and (Above 40) - Employee; GP	91

Figure 40: Comparison of (F) Age Group between (Below 40) and (Above 40) - Employee; GP	91
Figure 41: Segment Size of Clustering	92
Figure 42: Profile Segment of Clustering	93
Figure 43: Segment 3 - Largest Segment.....	93
Figure 44: Predictive Model Framework.....	95
Figure 45: Fit Statistics - Default Tree	96
Figure 46: Subtree Assessment Plot - Default Tree.....	96
Figure 47: Variable Importance - Default Tree	97
Figure 48: Default Tree - Tree Overview	98
Figure 49: Node Rules - Node 5	99
Figure 50: Node Rules - Node 6	99
Figure 51: Node Rules - Node 14	100
Figure 52: Fit Statistics - Regression	100
Figure 53: Type 3 Analysis of Effects - Regression	101
Figure 54: Fit Statistics - Stacking Ensemble Model (Base Tree + Meta Tree)	102
Figure 55: Subtree Assessment Plot - Stacking Ensemble Model (Base Tree + Meta Tree).....	102
Figure 56: Variable Importance - Stacking Ensemble Model (Base Tree + Meta Tree)	103
Figure 57: Tree Overview - Stacking Ensemble Model (Base Tree + Meta Tree)	104
Figure 58: Fit Statistics - Model Comparison.....	105
Figure 59: Lift Chart - Model Comparison	105
Figure 60: Orange - Test and Score	107
Figure 61: Orange - ROC Chart	108
Figure 62: Orange - Tree Overview	108
Figure 63: Orange - Test and Score (Customer Churn, Loyalty Program)	110
Figure 64: Orange - ROC Chart (Customer Churn, Loyalty Program = "No")	111
Figure 65: Orange - ROC Chart (Customer Churn, Loyalty Program = "Yes")	112
Figure 66: Orange - Test and Score (Loan Risk Defaulted).....	113
Figure 67: Orange - ROC Chart (Loan Risk Defaulted = "No")	114
Figure 68: Orange - ROC Chart (Loan Risk Defaulted = "Yes")	114
Figure 69: Orange - Test and Score (Customer Attrition)	116
Figure 70: Orange - ROC Chart (Customer Attrition = "No")	117
Figure 71: Orange - ROC Chart (Customer Attrition = "Yes")	118
Figure 75: SAS Enterprise Miner Main Page.....	Error! Bookmark not defined.
Figure 76: Creating a New Project 1	Error! Bookmark not defined.
Figure 77: Creating a New Project 2	Error! Bookmark not defined.
Figure 78: Creating a Library 1	Error! Bookmark not defined.
Figure 79: Creating a Library 2.....	Error! Bookmark not defined.
Figure 80: Creating a Library 3.....	Error! Bookmark not defined.
Figure 81: Creating a Data Source 1	Error! Bookmark not defined.
Figure 82: Creating a Data Source 2	Error! Bookmark not defined.

<i>Figure 83: Creating a Data Source 3</i>	Error! Bookmark not defined.
<i>Figure 84: Creating a Data Source 4</i>	Error! Bookmark not defined.
<i>Figure 85: Creating a Data Source 5</i>	Error! Bookmark not defined.
<i>Figure 86: Creating a Data Source 6</i>	Error! Bookmark not defined.
<i>Figure 87: Creating a Data Source 7</i>	Error! Bookmark not defined.
<i>Figure 88: Creating a Data Source 8</i>	Error! Bookmark not defined.
<i>Figure 89: Successful Creation of Data Source</i>	Error! Bookmark not defined.
<i>Figure 90: Creating a Diagram 1</i>	Error! Bookmark not defined.
<i>Figure 91: Creating a Diagram 2</i>	Error! Bookmark not defined.
<i>Figure 92: Creating a Diagram 3</i>	Error! Bookmark not defined.
Figure 93: Comparison of Age Group between (Below 40) and (Above 40) - SP	159
Figure 94: Comparison of (M) Age Group between (Below 40) and (Above 40) - SP	160
Figure 95: Comparison of (F) Age Group between (Below 40) and (Above 40) - SP	161
Figure 96: Comparison of Age Group between (Below 40) and (Above 40) - Employee; SP	161
Figure 97: Comparison of (M) Age Group between (Below 40) and (Above 40) - Employee; SP	162
Figure 98: Comparison of (F) Age Group between (Below 40) and (Above 40) - Employee; SP	163

List of Tables

<i>Table 1: Search Strategies</i>	19
<i>Table 2: Feature Comparison between Ensemble Methods</i>	32
<i>Table 3: Search Strategies</i>	33
<i>Table 4: Characteristics of Included Studies</i>	35
<i>Table 5: Data Understanding - Healthcare Data</i>	46
<i>Table 6: Feature Engineering - Healthcare Data</i>	49
<i>Table 7: Feature Selection - Healthcare Data</i>	55
<i>Table 8: Top 10 Business Industry with Highest Patient (Employee) GP Claims 2018</i>	73
<i>Table 9: Top 10 Business Industry with Highest GP Claims 2018</i>	74
<i>Table 10: Top 10 Diagnoses based on # of GP Claims within Top 10 Business Industry</i>	75
<i>Table 11: Patients (Employees) based on Total Remaining Amount Range</i>	76
<i>Table 12: Customer Churn Dataset</i>	110
<i>Table 13: Loan Risk Default Dataset</i>	112
<i>Table 14: Employee Attrition Dataset</i>	116
<i>Table 15: Platform Testing (SAS Enterprise Miner and Orange)</i>	127
<i>Table 16: Robustness (3 differing industries - Retail / HR / Financial Institution)</i>	128
<i>Table 17: Past Research Comparison</i>	128
<i>Table 18: Research Outcome</i>	133
<i>Table 19: Data Understanding</i>	149
<i>Table 20: SP - Data Understanding</i>	150
<i>Table 21: IP - Data Understanding</i>	153

1. Introduction

1.1 Background of Research

In the 21st century, the global landscape has undergone a profound digital transformation, with big data emerging as one of the defining evolutions (Gore, 2012) (Kar, 2015). This exponential growth in data has triggered a fundamental shift in how businesses operate and make decisions (Rahm, 2016). The ability to capture, store, process, analyse and visualize an unprecedented volume of data to uncover meaningful insights and patterns which may be impactful towards business processes, business operations, create business opportunities while reducing time and monetary resource that were previously hidden within data deluge (Gore, 2012) (Rahm, 2016). Nowhere is the impact of big data more prominent than in the healthcare sector, which stands as one of the most data intensive industries (Eapen, 2004) (Sippe, 2015). Healthcare providers are swamped with a myriad of data sources including electronic medical records (EMRs), medical claims data, pharmaceutical data, patient behaviour, and other historical patient data (Sippe, 2015). These data serve as a critical platform for clinical decision support, disease prediction and prevention and understanding of population health (Sippe, 2015). Given the exponential growth of data in healthcare, there is an urgent need to leverage these vast amounts of data to extract meaningful insights and patterns which could benefit healthcare decision support and enhance healthcare delivery (Eapen, 2004).

To date, data analytics has emerged as a powerful analytical tool, offering a suite of techniques and methodologies for extracting meaningful insights and patterns from data. For instance, predictive analysis can be performed using data mining techniques such as Decision Tree, Regression, Support Vector Machine (SVM), Clustering and many others. Another such technique gaining prominence in today's expansive data analytics environment is Ensemble Learning. Ensemble Learning refers to a combination of learners or a process of running two or more analytical models trained to solve the same problem (Rouse, 2015) (Tuysuzoglu,

Birant, & Pala, 2017). It is a machine learning technique whereby predictions are combined from multiple learners into a single output that would potentially churn out a better performance as compared to a single learner; sometimes also known as synthesizing it into a single combined predictive outcome with improved predictive accuracy (Rouse, 2015) (Tuysuzoglu, Birant, & Pala, 2017).

For healthcare providers, the ability to effectively manage and analyse the vast amounts of data at their disposal is crucial (Guo & Chen, 2023). In addition to EMRs, medical claims data, medical diagnosis, and other claim records / patient behaviour significantly contribute to the healthcare data landscape. By harnessing data and data analytics, it unlocks an opportunity to analyse and identify patterns in these claims which could potentially assist them in making specific decisions such as to better understand the overall population health and enhance clinical decision-making. In Malaysia, employers provide employee medical coverages and benefits - these coverages and benefits may be extended to employee spouses, child / children and sometimes employee's parents (Malaysian Reserve, 2017) (Beh, 2019). The data generated from these medical coverages and benefits, commonly referred to as medical claims, represents a valuable repository of data for employers (Beh, 2019). Typically managed and stored by Human Resources (HR) department and third-party insurance panel appointed by the company, these medical claims data encompass details such as medical claim history, diagnoses, incurred claim amounts, and other pertinent information. However, due to the ever-increasing healthcare costs, it has presented employers with the urge to leverage on data analytics to optimize resource allocation and better manage healthcare expenditures (Malaysian Reserve, 2017). Predictive analysis emerges as a powerful analytical technique to proactively manage company resources and mitigate increasing healthcare costs (Singh, Agrawal, Sahu, & Kazancoglu, 2023). By identifying high-cost and high-risk employees early on, employers can implement targeted interventions to address potential health issues before they escalate into more serious

health problems. Early identification of high-risk individuals allows for the implementation of proactive measures, reducing the need for reactive and costly interventions down the line. However, in Malaysia, there remains inadequate research surrounding this area of interest. This gap in knowledge presents an opportunity for research to provide a breakthrough in valuable insights that could benefit employers in navigating healthcare management. The aim of this research is to analyse patterns within employee healthcare data to gain a comprehensive understanding of the overall population health and utilization of premium coverages provided by employers. To achieve this objective, a proposed ensemble stacking model approach will be applied, offering a simplified yet robust framework accessible to practitioners who may not have expertise in analytics (Abdunabi, 2016). By focusing on a simplified and practical predictive model, the research seeks to provide actionable insights that can inform strategic decision-making and resource allocation, ultimately benefiting employers and contributing to the advancement of healthcare management practices in Malaysia.

1.2 Problem Statement

What is the usage pattern and insights of employee healthcare claims? What do the claims data show? What are the factors contributing to a high-cost/-risk employee? There is a lack of understanding towards the overall employee population health. Employers need to show concern in the wellbeing of their employees because it affects the sustainability of medical premium coverages as well as productivity and focus on work (Institute of Medicine (US) Committee, 2002). Early detection presents an opportunity to be proactive instead of reactive and allows employers to prepare targeted approaches to address any health problems which may be triggered due to working conditions. Additionally, due to the ever-increasing medical costs (Zin, Rahman, Nazar, Kurdi, & Godman, 2023) (Mardhiah, 2023), it has triggered an urge from employers to discover and understand the usage patterns of healthcare claims to better understand high-cost/-risk employees (patients) while identifying factors which

contribute to high-cost/-risk employees (patients) - enabling employers to prepare proactive measures and strategies. The value healthcare data may carry is unknown if it is not harnessed, then the data captured and stored would be meaningless, thus, transforming it into information then knowledge would help employers to act. This is a collaborative research with one of the largest conglomerates in Malaysia.

Current research in healthcare prediction such as medical conditions and healthcare coverages, uses different machine learning algorithms such as - Naïve Bayes, Bayesian Network, Decision Tree, Neural Network and other boosting and bagging techniques (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007) (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012) (Jain, Predictive Modeling for Chronic Conditions, 2015). However, how will an ensemble stacking model approach compare against existing techniques? Will the proposed ensemble stacking model approach outperform existing literatures and increase the predictive accuracy? Stacking refers to combining multiple predictive models to churn a single predictive outcome. Ensemble stacking models were proposed because it has the potential to enhance predictive outcomes by increasing the predictive performance and accuracy while providing a more robust model as compared to boosting and bagging techniques. This research aims to uncover and verify if ensemble stacking model approach would increase the predictive accuracy.

1.2.1 Research Questions

- a. What is the usage pattern of employee healthcare claims and what are the factors contributing to a high-cost/-risk employee (patient)?
- b. How does an ensemble stacking model approach compare against the existing bagging / boosting techniques applied in existing literatures?
- c. Will the proposed ensemble stacking model approach increase predictive accuracy as compared to a single predictive model?

1.2.2 Research Objectives

- a. To discover and understand the usage pattern of healthcare claims to better understand high-cost/-risk employees (patients) while identifying factors which contribute to high-cost/-risk employees (patients).
- b. To propose an ensemble stacking model approach as it provides advantages such as simplicity; improved performance; and capability of a combined model induced by various models over bagging and boosting techniques.
- c. To verify and validate that ensemble stacking model approach would increase the predictive accuracy and can be used by practitioners who are non-experts in the field of analytics while being a more robust model which can be applied across a wide range of classification applications.

1.2.3 Research Contribution

- a. With minimal research and application of ensemble modelling techniques applied within the context of Malaysia, this research presents an opportunity to delve into data analytics within Malaysia using techniques such as ensemble model.
- b. Focus revolves around clinical identifications and predictions which require an expert in the field of analysis to perform analysis, while little focus has been put into proposing an ensemble stacking model which can be applied across a wide range of classification applications.
- c. Ensemble models have been applied across various fields such as weather forecasting, finance, manufacturing, security, and medicine. However, there has been minimal focus on proposing an ensemble stacking model with the combination of feature selection and feature engineering. Stacking technique also increases the robustness of the ensemble model as compared to bagging and boosting. An ensemble model aims to produce better predictive outcomes and accuracy.

- d. Data Mining, Predictive Analysis and Machine Learning techniques which were applied in the models are based on mathematical formulations, statistical calculations and technically too complex to be understood by the others who are not experts in the field. Hence, using simpler models such as Decision Tree or Clustering Analysis can provide easier understanding of the analysis which has been performed. The application of data mining techniques and other predictive models would result in classifications which may be difficult to understand (Olofsson, 2017). Moreover, there has been a growing concern regarding the interpretability of predictive models which will be taken into consideration in this research, which is rarely addressed in data mining prediction studies (Olofsson, 2017). It can be deemed as a problem of general interest within the field of analytics as well (Olofsson, 2017). By using an ensemble stacking model approach, it provides advantages such as simplicity; increases robustness; improved performance; and capability of a combined model induced by various models. It adds the flexibility which is lacking in other ensemble methods such as boosting and bagging. By combining the base- and meta-learner, there is a flexibility to use different models to improve the prediction accuracy. In fact, EU legislation which is the main aspects of the European legislation, policies and activities have requested that machine learning models be more interpretable under the act of General Data Protection Regulation (GDPR) and it adds that citizens have the right to obtain explanations for algorithmic decisions; for example, credit or risk assessments (Olofsson, 2017). This shows that the interpretability in general requires a subject matter expert to be involved in the analysis process and an individual without any knowledge in predictive analysis may not fully grasp the concept hence, there is the growing

interest in ensuring interpretability is done with individuals with minimal knowledge as well.

1.3 Research Motivation

This research will enable the researcher to improve and test the proposed ensemble stacking model which could be applied by practitioners across various classification applications. The motivation behind this research is to firstly, better understand the current healthcare claims while performing predictive analysis to identify potential patients (employees) who may be high cost/risk and to potentially reduce medical expenditure and cost and secondly, to improve and test the proposed ensemble stacking model while testing the assumption of ensemble stacking model being more accurate as compared to single predictive models.

Moreover, due to the ever-increasing medical cost (Birruntha, 2024) (Khoo, 2024), it has accelerated the urgency within employers to better understand their overall employee health population which could allow them to potentially prepare proactive measures to sustain the medical premium coverages incurred. In Figure 1, Frost & Sullivan stated in *The Edge Markets* in March 2018 that “Malaysia’s healthcare industry is expected to experience a growth to RM80b by 2020”. Fitch Research in December 2018 also stated in *The Star Online* (Figure 2) that “Malaysia’s healthcare market will reach RM127.9b by 2027”. These news articles show that medical expenditure will continue to rise in a tremendous rate, and it has become a cause of concern for employers who are consistently paying a premium for the employee medical insurance. That is why the need to explore and identify the current employee medical claim trend to better understand and to propose potential recommendations to employers to help sustain the premium insurance.



Figure 1: Malaysia Healthcare Industry Increasing Medical Expenditure (March 2018)



Figure 2: Malaysia Healthcare Industry Increasing Medical Expenditure (December 2018)

1.4 Thesis Outline

This thesis has 6 chapters organized as follows:

Chapter 1 provides the background of the research. It provides an overview of the thesis, problem statement, research questions, research objectives, research contribution and research motivation.

Chapter 2 provides a comprehensive literature review and related works performed within the scope of healthcare analysis and predictive analysis while it covers the area of ensemble models. It begins with the explanation of studies performed by previous researchers, then it covers the various topics which are included within this research and the section is concluded with a systematic review.

Chapter 3 explains the research methodology which have been applied in the research. It explains each section in detail and what are the processes which have been performed such as data preparation, feature selection and feature engineering.

Chapter 4 provides the holistic overview of the analysis which have been performed broken down into 2 major sections which are Descriptive Analysis and Predictive Analysis. This chapter shows the various predictive techniques which have been applied and how the ensemble stacking model produces a better predictive outcome.

Chapter 5 explains and discusses about the findings from the healthcare analysis and the ensemble stacking model which have been applied.

Chapter 6 provides a conclusion and future work opportunities. It concludes the work done by summarizing the research while including the limitations, recommendations, and future work.

2. Literature Review

2.1 Background of Research

Chapter 2: Literature Review starts off by diving into past research which have been carried out in healthcare analytics. These are research which have underlying similarities with the research that is being performed. Next, limitations were extracted from the literatures to perform gap analysis to identify any gaps of research. Big data analytics were further explored to better understand what has been implemented in the past and which areas can be explored further. In which, the focus is on Malaysia as a country to better understand how healthcare analytics can be applied within Malaysia's context and what are the limitations in the current research within Malaysia. Data mining and the techniques which have been applied is the core of this research. An approach which is common among the experts in the field of analytics would be ensemble. Stacking ensemble model approach has been explored in this research, a comparison between stacking, boosting, and bagging were performed. Finally, a systematic review of the scope of research was included in the last section of the literature review.

2.1 Analysis of Healthcare Analytics using Data Mining Techniques

A study by Tekieh, Mohammad Hossein, explores healthcare coverage disparity using quantitative analysis on a large dataset from the United States (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). One of the objectives is to build supervised models including decision tree and neural network to study the efficient factors in healthcare coverage (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). Groups of people with health coverage problems and inconsistencies were discovered by employing unsupervised modelling including K-Means clustering algorithm (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). The predictive modelling is based on the dataset retrieved from Medical Expenditure Panel Survey with 98,175 records in the original dataset (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques,

2012). After pre-processing the data, including binning, cleaning, dealing with missing values, and balancing, it contained 26,932 records and 23 variable (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). 50 classification models were built on IBM SPSS Modeler employing decision tree and neural networks (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). The accuracy of the models varies between 76% and 81% (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). The models can predict the healthcare coverage for a new sample based on its significant attributes (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). It was demonstrated that the decision tree models provide higher accuracy than the models based on neural networks (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). Also, having extensively analyzed the results, the most efficient factors in healthcare coverage are access to care, age, poverty level of family, and race/ethnicity (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). Another study focused on chronic diseases, where it was mentioned that chronic diseases contributed to 7 out of 10 deaths in the United States - it is one of the major causes of mortality around the world (Jain, Predictive Modeling for Chronic Conditions, 2015). Because of its adverse effect on the quality of life, it has become a major problem globally (Jain, Predictive Modeling for Chronic Conditions, 2015). Health care costs involved in managing these diseases are also very high (Jain, Predictive Modeling for Chronic Conditions, 2015). The study focused on 2 major chronic diseases which are Asthma and Diabetes which are among the leading causes of mortality around the globe. It involves design and development of a predictive analytics-based decision support system which uses five supervised machine learning algorithms to predict the occurrence of Asthma and Diabetes (Jain, Predictive Modeling for Chronic Conditions, 2015). This system helps in controlling the disease well in advance by selecting its best indicators and providing necessary feedback (Jain, Predictive Modeling for Chronic Conditions, 2015). Based

on several risk factors such as blood pressure, BMI, age, ethnicity, smoking status etc, the system would be able to predict the vulnerability of a person to a particular disease which helps in taking necessary action to avoid the disease well in advance (Jain, Predictive Modeling for Chronic Conditions, 2015). Lastly it was a study looking at health care data from patients in the Arizona Health Care Cost Containment System, Arizona's Medicaid program, they provide a unique opportunity to exploit state-of-the-art data processing and analysis algorithms to mine the data and provide actionable results that can aid cost containment (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). This work addresses specific challenges in this real-life health care application to build predictive risk models for forecasting future high-cost users (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). Such predictive risk modelling has received attention in recent years with statistical techniques being the backbone of proposed methods (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). The literature were surveyed, and a novel data mining approach was proposed to customize for this potent application (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). Our empirical study indicates that this approach is useful and can benefit further research on cost containment in the health care industry (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007).

Future enhancements and work which can be performed on the following research have been highlighted. Firstly, to compare the outcomes of intuitive heuristic approach and optimal approach in attribute reduction (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). The intuitive heuristic approach is the method used in this study, whereas the optimal approach is to experiment all combinations of attributes in each stage of attribute reduction (e.g. all combinations of 4 attributes out of 22 for stage 18) and selecting the best

combination which has the highest accuracy rate (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). Specify whether a cross-sectional study will have a better result in predicting healthcare coverage (just cross a specific part of the time, e.g. end of calendar years) or a longitudinal study (add the change of situation by passing time to the study) (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). Regarding the results of threshold analysis in this study, longitudinal study can provide dynamics of health insurance and also the results of being uninsured (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012). With the increase in the development of many health management strategies, the research presented here can be extended in a variety of directions. Some of the suggested extension include, (i) Diabetes prediction there were a smaller number of records with Borderline Diabetes hence one future scope would be to add a greater number of records with Borderline Diabetes, it will help the system to improve learning algorithms for Borderline cases, (ii) For Asthma prediction, clinical data can be included for training purpose. It will improve the overall accuracy of the system because clinical data have a significant effect on the predictions, (iii) the system can be extended to build models for other chronic conditions such as CKD, COPD and Heart Diseases (Jain, Predictive Modeling for Chronic Conditions, 2015). There is further scope to improve the interpretation of these results. It is commonly observed that a considerable percentage of high-cost patients do not remain that way every year (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). Also, two patients could share very similar profiles with only one of them being high cost (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). Studying these seemingly anomalous patients could provide a better understanding of how a high-cost patient is different from other patients (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007). Working with key partners and data owners, the focus is to provide a

reasonable and patient-specific answer to this question that will have a significant impact on cost containment in the health care industry (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007).

2.2 Limitation of Previous Research

Many applications were applied in Foreign Countries such as USA (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012) (Jain, Predictive Modeling for Chronic Conditions, 2015) (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007), more research and application can be applied within the context of Malaysia. As stated in a recent news article, there is a need to do more, and harness large amount of health data generated (Brar, 2018). In a journal article written by Nurul-Ain Mohd-Tahir, Malaysia has implemented health technology assessments, but the results have not been optimal, while more research and applications can be done (Mohd-Tahir, 2015). The Former Director-General of Health of Malaysia stated that information and communication technology can be employed to analyse healthcare data (Merican, 2018). Focus have been too clinical, there are too many clinical identifications and predictions while little focus has been put into potentially proposing an ensemble stacking model which can be applied by practitioners who are non-experts in the field of analytics (Alharti, 2018). Data Mining, Predictive Analysis and Machine Learning techniques which were applied in the models are based on mathematical formulations, statistical calculations and technically too complex to be understood by the others who are non-experts in the field of analytics (Alharti, 2018). There are literatures and journals which applied the ensemble method known as bagging and boosting (Abdunabi, 2016) (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real World Risk Modeling Application, 2007) (Moreira & Namen, 2018) (Li, Bai, & Reddy, 2016). However, another ensemble method known as stacking can be applied and a proposed ensemble stacking model approach would be an alternative to increase predictive accuracy,

while it could potentially be a key predictive model to be applied across various industries in different fields of interest. In the past, the focus revolved around identifying an algorithm or predictive model that can best stratify data and perform accurate predictions (Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014). However, this does not solve the underlying issue of the practicality and interpretability of the model (Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014).

2.3 Overview of Big Data Analytics in Healthcare

Big data analytics is historically inevitably connected to that of data science (Wang, Kung, & Byrd, 2018) (Guo & Chen, 2023). Michael Cox and David Ellsworth first used the word "big data" in 1997 in a paper presented at an IEEE conference to describe the representation of the data and the problems it posed to computer systems (Wang, Kung, & Byrd, 2018). Rapid innovations at the end of the 1990s in Information Technology enabled large volumes of data to be generated, however, there were little useful information (Wang, Kung, & Byrd, 2018). Concepts of Business Intelligence (BI) were developed to echo the importance of collecting, integrating, analysing, and interpreting business information and how the process can assist businesses in making appropriate business decisions through the understanding of market trends and behaviours (Wang, Kung, & Byrd, 2018). In the early 2000s, an evolution of big data development broke through where it was defined by 3Vs: Volume, Velocity and Variety (Wang, Kung, & Byrd, 2018). Big data can be described as "large volumes of high velocity, complex and varied data that require advanced analytical techniques to capture, store, distribute, manage, and analyse raw data into valuable pieces of information (Raghupathi & Raghupathi, 2014). The 3Vs encompasses the general description of big data where volume refers to the sheer amount of data being generated or with reference to the scale and size of data, while velocity refers to the speed at which data is being generated, stored, processed, and

analyzed - real-time, batch or periodic and lastly, variety refers to the types of data - structured or semi-structured (Emmanuel & Stanier, 2016).

Data is being generated at such exponential rate in almost every sector especially in healthcare where it is described as one of the most data extensive industries (Raghupathi & Raghupathi, 2014) (Guo & Chen, 2023). Healthcare industry have been generating enormous amounts of data through record keeping, patient detailed information and compliance & regulations requirements (Raghupathi & Raghupathi, 2014). Traditionally these data were in hard copy, the trend has begun to digitize medical records and aggregate clinical data into electronic databases (Wang, Kung, & Byrd, 2018) (Guo & Chen, 2023) (Ibeh, et al., 2024). This sparked major development in the ability to ensure healthcare data are captured, stored, managed, and analyzed to convert data into actionable and searchable information to assist healthcare providers in making better decisions (Wang, Kung, & Byrd, 2018) (Jeremiah Olawumi Arowoogun 1, Chidi, Adeniyi, & Okolo, 2024). By definition, big data in healthcare refers to the large and complex electronic healthcare data which is virtually impossible to be managed and analyzed by traditional software as the volume of data being generated is overwhelming but also because of the diversity of data types and speed at which the data must be managed (Raghupathi & Raghupathi, 2014). The potential for big data analytics in healthcare to lead to improved results exists in several contexts, for example: by evaluating patient preferences and the quality and outcomes of care to determine the most appropriate and cost-effective interventions, and by providing analysis and tools to influence the actions of providers; applying advanced analytics to patient populations (e.g. segmentation and predictive modelling) to proactively classify individuals benefiting from preventive treatment or improvements in lifestyle; broad-scale disease profiling to detect predictive events and promote prevention initiatives; gathering and publishing data on medical procedures; Helping patients to identify treatment procedures or schemes that provide the best value; detecting, anticipating

and mitigating fraud by introducing advanced fraud detection analytical systems and testing the quality and integrity of claims; and enforcing claim authorization far closer to real-time; create new revenue sources by aggregating and synthesizing clinical reports of patients and collections of statements to offer data and services to third parties, such as licensing data to assist pharmaceutical companies in selecting patients for inclusion in clinical trials (Raghupathi & Raghupathi, 2014). These are just some potentials for big data analytics in healthcare. One of the key areas in big data analytics in healthcare would revolve around patient profile analytics, where advanced analytical techniques such as segmentation and predictive modelling will be applied to identify patients who are at risk of developing specific diseases while providing preventative care or proactive care (Raghupathi & Raghupathi, 2014). However, it is said that electronic healthcare data are underutilized and wasted while there is a greater sense of urgency to convert raw data into meaningful and actionable information (Mehta & Pandit, 2018). Through identification of hidden patterns, it would lead to an improvement in healthcare quality while potentially making more cost-effective and timely decisions (Mehta & Pandit, 2018).

2.3.1 Overview of Big Data Analytics in Healthcare within the context of Malaysia

As the healthcare industry continues the exponential growth, the large volumes of data generated is expanding dramatically as well (Gunasekar & Kayalvizhi, 2019). Most of the data and information are kept in a hardcopy document and maintained manually, however, due to ever growing data being collected, many are transitioning into digitization of these data and information (Gunasekar & Kayalvizhi, 2019). Big data analytics (BDA) is seen as a vital aspect in healthcare and research has been performed to facilitate better services to patients - with more effort being put into the utilization of BDA to assist in the process of disease diagnostics and care delivery (Gunasekar & Kayalvizhi, 2019) (Azmi, Noor, Shukri, & Aidalina Mahmud, 2022). It can be considered as one of the fastest growing technology in Malaysia (Marjudi,

Setik, Ahmad, Harun, & Ismail, 2020) ((MAMPU), 2020). However, there are still a lot to be done as the adoption and research development has been halted by the fundamental problems present within the big data and data analytics paradigm - more analysis and information can be explored to better understand the readiness of Malaysia to incorporate BDA into healthcare analysis (Gunasekar & Kayalvizhi, 2019) (Marjudi, Setik, Ahmad, Harun, & Ismail, 2020). BDA involves complex processes that requires the expertise and knowledge of practitioners within the field of analytics to help decipher healthcare analysis (Gunasekar & Kayalvizhi, 2019). Some of the potential applications of BDA in healthcare include the utilization of machine learning algorithms to predict potential admission into health facilities; real-time alert / notification system to send an alert to doctors / nurses when there are any anomalies which have been detected in patients; to identify high cost / risk patients through predictive modelling (Lai, Mai, Sulaiman, & Lim, 2019) (Ghaleb, Dominic, Singh, & Naji, 2023). There are several areas of interests in Malaysia healthcare which have been voiced out in an article from IMU University, stating that, data analytic tools are vital in healthcare and more focus is required to create value-based care and to reduce overall medical cost while attaining a better risk prediction (Lai, Mai, Sulaiman, & Lim, 2019). Another area would be to create a paradigm shift among healthcare professional to openly embrace and accept analytics, this could be resulted in the jargons and expertise which may be required to perform and understand analytics while fundamentally focus could be to minimize and reduce the complexity by driving towards a simplified model - this has been further supported by an article written by Surenthiran Krishnan where he mentioned that specialized techniques and experts in analytics are often required to process and interpret these large healthcare data sets but analytics can be used to perform predictions of diagnostic services (Lai, Mai, Sulaiman, & Lim, 2019) (Krishnan, Magalingam, & Ibrahim, 2018). In another research by Quek Kia Fatt and Amutha Ramadas, they mentioned on how Malaysia is driving focus towards the use of big data analytics on

patients medical records and how it could potentially predict the outcome of disease prevention of co-morbidities and mortality (Fatt & Ramadas, 2018). Although, there are several concerns which have been raised, one of it being miscommunications gap, whereby practitioners who are non-experts and data scientist have a major knowledge gap as the understanding of the practitioners are minimal which affects the effectiveness and usage of analytics (Fatt & Ramadas, 2018).

There are research which have been conducted within the context of Malaysia on specific topics such as disease detection and healthcare resource utilization (Marjudi, Setik, Ahmad, Harun, & Ismail, 2020) (Krishnan, Magalingam, & Ibrahim, 2018) (A, et al., 2020). A recent research performed among white-collar workers in Malaysia on Cardiovascular Disease (CVD) risk factors, they mentioned that the total prevalence had almost doubled between the year 2015 to 2018 (Marjudi, Setik, Ahmad, Harun, & Ismail, 2020). The research focuses on the risk factors which may be present and may have a direct contribution to Cardiovascular Disease (CVD) (Marjudi, Setik, Ahmad, Harun, & Ismail, 2020). In another research by Surenthiran Krishnan, he proposed the use of a big data framework to predict heart diseases (Krishnan, Magalingam, & Ibrahim, 2018). In a research performed in University of Malaya Medical Centre, the time-series projections were explored to better understand the trend analysis and forecast the Covid-19 virus while at the same time to estimate the number of patients who might require care and estimate the number of resources which may be required such as aprons, sterile and non-sterile isolation gowns, face masks and face shields (A, et al., 2020). Furthermore, one of the most useful applications would be the utilization of simple mathematical calculations and approaches for healthcare and medical predictions - while the focus should be driven towards simple algorithms leading to more successful implementations of analytics in healthcare (A, et al., 2020).

Table 1: Search Strategies

Author	Year	Title	Conclusion
Suziyanti Marjudi; et al. (Marjudi, Setik, Ahmad, Harun, & Ismail, 2020)	2020	Cardiovascular Disease Risk Factors among White-Collar Workers towards Healthy Communities in Malaysia	a lot to be done as the adoption and research development has been hindered by the fundamental problems present within the big data and data analytics paradigm - more analysis and information can be explored to better understand the readiness of Malaysia to incorporate BDA into healthcare analysis
Pei Kuan Lai; et al. (Lai, Mai, Sulaiman, & Lim, 2019)	2019	Healthcare Big Data Analytics: Re-engineering Healthcare Delivery through Innovation	to openly embrace and accept analytics, hindered by jargons and expertise which may be required to perform and understand analytics while fundamentally focus could be to minimize and reduce the complexity by driving towards a simplified model
Surenthiran Krishnan; et al. [20]	2018	Review on Data Analytics Framework in Heart Disease	analytics can be used to perform predictions of diagnostic services. good understanding of the data will lead to the best approach and assist future patients.
Fatt, Quek Kia; et al. (Fatt & Ramadas, 2018)	2018	The Usefulness and Challenges of Big Data in Healthcare	potential to predict outcome of diseases, disease prevention. some concerns on miscommunication, there is a gap between practitioners who are non-experts and data scientist as the understanding of the practitioners are minimal which affects the effectiveness and usage of analytics
Azzeri A; et al. (A, et al., 2020)	2020	Prediction of Disease Burden and Healthcare Resource Utilization through Simple Predictive Analytics using Mathematical Approaches	while the focus should be driven towards simple algorithms leading to more successful implementations of analytics in healthcare.

In conclusion, some of the gaps which have been identified in Malaysia's context show that there are more to be done towards the adoption of data analytics as the fundamental problems are not looked into and more can be explored to better understand the healthcare data available in Malaysia. Often times, the jargons and expertise which may be required to incorporate data analytics has hindered the progression as there is a lack of focus on minimizing and reducing complexity through a simplified model. There is also a miscommunication and a gap between practitioners and data scientists leading to a view that data analytics has minimal effective in healthcare which is not the case. The gaps identified in Malaysia's context shows that there are more room for improvement and to be explored on the implementation of data analytics in Malaysia.

2.4 Data Mining

Data mining is known as the process to extract and discover meaningful insights from data, through a combination of statistical analysis, machine learning and database technology (Alonso, Díez, Rodrigues, Hamrioui, & López-Coronado, 2017). Data mining can also be described as a process which uses query tools and techniques to discover previously unknown patterns and trends within massive databases while using that information to perform predictive analysis (Kincade, 1998) (Xiong, et al., 2024). Data mining would usually include tools and techniques such as classification, regression, clustering, and association to perform analysis (Alonso, Díez, Rodrigues, Hamrioui, & López-Coronado, 2017) (Feng & Fan, 2024). Each data mining technique would be used for different purposes depending on the objective. Classification and prediction are the most common modelling objectives - classification refers to prediction of categorical labels (discrete or binary) while prediction refers to continuous value functions (Alonso, Díez, Rodrigues, Hamrioui, & López-Coronado, 2017). As mentioned, data mining aims at learning from data through two general methods known as supervised and unsupervised learning methods (Obenshain, 2004). Supervised learning

methods are applied when input variables used to make predictions of a target with a known outcome while unsupervised learning methods are applied more commonly on a target without a known outcome (Obenshain, 2004). For instance, an example of supervised learning method would be to predict customer churn by identifying the characteristics that distinguish churn while an unsupervised learning method would be to identify segments of individuals based on their spending behaviours and patterns.

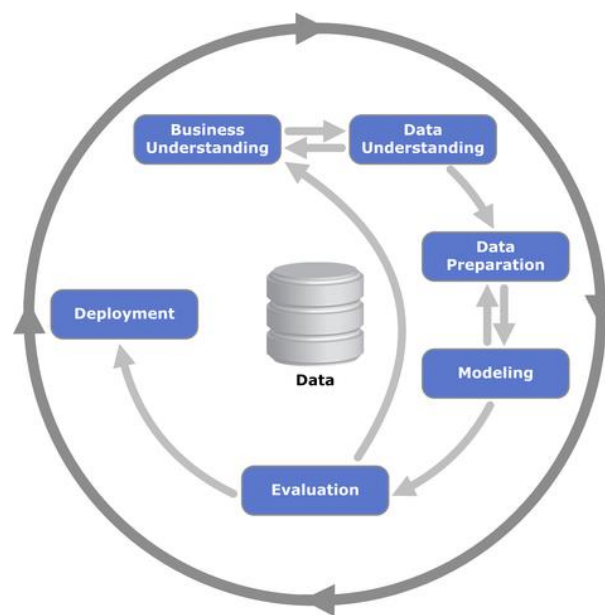


Figure 3: CRISP-DM (Cross-Industry Standard Process for Data Mining) Methodology

Cross-Industry Standard Process for Data Mining also known as CRISP-DM (Figure 3) proposes a 6-step process methodology for data mining: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Chye & Tan, 2011). Business understanding is considered as the most important step where the business objectives, problem statements and the success criterion would be defined (Chye & Tan, 2011). Moreover, as data mining implies, data would also be a crucial component where the 2nd and 3rd step would ensure a thorough understanding of the data while the data would be prepared for analysis - some would suggest ETL (Extract, Transform and Load), data transformation and sampling - these are essential antecedents for data modelling (Chye & Tan, 2011). Modelling is the 4th step where data analysis would be performed - classification models, regression models,

association, clustering are some analytical techniques which are applied (Chye & Tan, 2011). Evaluation step would allow for a comparison of models based on their predictive accuracy prior to model selection (Chye & Tan, 2011). Once the model has been evaluated and selected, deployment can be proceeded with actual implementation of the selected model (Chye & Tan, 2011).

2.4.1 Techniques in Data Mining

There are several techniques which are being used commonly in Data Mining / in the world of Data Analysis. Some of the common techniques include Decision Tree, Regression and Clustering. Decision Tree is commonly referred to as a classification machine-learning algorithm, it resembles a tree-like structure. It uses an if-else top-down approach to split the test variable (target variable) that is characterized by the root node, internal nodes, and leaf nodes. Regression is a supervised learning algorithm based on statistical methods, there are 2 common types of regression: Logistic and Linear Regression. Regression aims to identify the correlation between a target or dependent variable and predictor(s) or independent variables. Clustering is an unsupervised machine learning algorithm used to split groups based on their distinct characteristics. It organizes objects into groups based on their similarities or common features.

Clustering analysis is an unsupervised machine-learning algorithm that divides groups of objects with common characteristics into distinct groupings (Zhong & Xiao, 2017). In other words, it would organize similar characteristic objects together into a cluster and dissimilar objects would belong in another cluster (Bertsimas, et al., 2008). The objective of Clustering analysis is to identify structure within the data set (Zhong & Xiao, 2017). It is commonly applied in exploratory data mining while it can be said that it is part of descriptive analysis as it does not possess any predictive capability (Deshmukh & Gulhane, 2016). The groups in Clustering are not known in advance and the goal is not to define how the grouping should be

generated but rather, it uses every single variable in a dataset to consider the clusters to be generated (Chua, Clustering Analysis Concepts). When using Clustering, it is not possible to measure the accuracy of the model, rather, it is determined based on the usefulness of the clusters which have been defined (Chua, Clustering Analysis Concepts). Because of this, Clustering is considered as an open-ended solution to explore, understand, and formulate questions about data in the exploratory data analysis (Chua, Clustering Analysis Concepts). Moreover, if it is applied in healthcare, Clustering would potentially be useful in identifying association between risk-factors and health (Zhong & Xiao, 2017). Clustering can be divided into 2 categories, Hard Clustering or Soft Clustering (Kaushik, An Introduction to Clustering and different methods of Clustering, 2016). In Hard Clustering, a data point will either belong to a cluster completely or it does not (Kaushik, An Introduction to Clustering and different methods of Clustering, 2016) - for instance, a customer is put into one out of 10 groups present (Kaushik, An Introduction to Clustering and different methods of Clustering, 2016). In Soft Clustering, instead of divided each data point into separate clusters, a probability or likelihood is calculated to identify if a data point will be put into which clusters is assigned (Kaushik, An Introduction to Clustering and different methods of Clustering, 2016) - for example, a customer is assigned a probability to be in either of the 10 clusters which are presented (Kaushik, An Introduction to Clustering and different methods of Clustering, 2016). Several algorithms can be applied when using Clustering technique such as Agglomerative Hierarchical Clustering and K-Means Clustering or K-Nearest Neighbour Clustering (Deshmukh & Gulhane, 2016). Agglomerative Hierarchical Clustering uses a “bottom-up” approach, it begins with each data point and progressively creates clusters by merging every data point together until the last data point available (Chua, Clustering Analysis Concepts). K-Means, on the other hand, is a non-hierarchical method for grouping where it uses a “top-down approach” where it begins with a pre-defined number of clusters which assigning each data point to the them (Chua, Clustering

Analysis Concepts) - one key aspect to note is that there are no duplicates as every data point is belonging to only a single cluster (Chua, Clustering Analysis Concepts). As compared to Agglomerative Hierarchical Clustering, K-Means is computationally faster and is able to handle a large dataset (Chua, Clustering Analysis Concepts). With the massive amounts of data, Clustering analysis would present an opportunity to discover hidden patterns previously unknown as well as detecting anomalies (Zhong & Xiao, 2017). As mentioned, when Clustering analysis is applied into a given dataset, it will automatically detect and identify patterns hidden in the dataset and group objects who are similar (Bertsimas, et al., 2008).

Decision Tree is a supervised machine-learning algorithm that resembles a tree-like structure (Yuvaraj & SriPreethaa, 2017). It follows a flowchart if-else structure in a top-down approach where an internal node or a non-leaf node represents a test on a selected variable (Yuvaraj & SriPreethaa, 2017). Decision Tree can be characterized by its specific properties such as it contains a root node, internal nodes, leaf node and it is defined by the rules and conditions of the splits (Chua, Decision Tree Concepts). The root node is the first node at the top which begins the tree structure, it is determined by how pure the attribute is while the following nodes are internal nodes, and the final node is leaf node (Chua, Decision Tree Concepts). Commonly, the pureness of an attribute is calculated through a series of methods such as Logworth, Entropy or Gini (Chua, Decision Tree Concepts). There are 2 ways in which a split can occur, multi-way split or binary split as the name suggests multi-way splits meaning splits which are more than 2 values while binary is just 2 (Chua, Decision Tree Concepts). Every branch in the tree model represents the outcome of a test while the final node of the model refers to as leaf node indicates the class label which denotes the predicted outcome value (Yuvaraj & SriPreethaa, 2017). Decision Tree is also referred to as a classification model. Decision Tree model is a structure where the process begins at the top also known as the root node and the arguments would flow down until it reaches the last node which is known as the leaf node - where as

previously mentioned, a predicted outcome is shown, or a decision can be made (Raul, Patil, Raheja, & Sawant, 2016). It is also interpreted as a unique set of rules form which is characterized and denoted by its hierarchical organization rules (Raul, Patil, Raheja, & Sawant, 2016) - this hierarchy will allow for simple but powerful outcomes to make strategic decisions (Raul, Patil, Raheja, & Sawant, 2016). In decision making, Decision Tree is used to visually and explicitly represent the flow of decisions (Gupta, 2017). As mentioned previously, as the name suggests, it presents decisions in a tree-like model (Gupta, 2017). Decision Tree is a simple and fast learning classification model where the objective is to construct an optimal tree model based on the specified target variable (Yuvaraj & SriPreethaa, 2017). An advantage of Decision Tree is because of the non-parametric nature, it has the ability to handle large, complicated datasets without imposing a complex parametric nature (Yuvaraj & SriPreethaa, 2017). Moreover, because of its simple nature, Decision Tree can be often said to be mimicking human level of thinking hence, it is easily understandable and interpretable (Sanjeevi, 2017).

Regression is a supervised learning algorithm which is based on statistical methods (Rathi, 2010). In other words, Regression is a data mining predictive technique which is used to predict a range of numerical values (also known as continuous values), given a particular problem and dataset to solve (Chapple, 2018). The aim of Regression would be to identify the relationship between a target (dependent) variable and predictor(s) (independent) variables (Ray, 2015). There are 2 basic form of Regression technique which are most commonly used and known, Linear Regression and Logistic Regression (TechDifferences, 2018). To differentiate between Linear and Logistic Regression, the nature of Linear is used when the target (dependent) variable is continuous whereas the nature of Logistic is used when the target (dependent) variable is binary (TechDifferences, 2018). A simple Linear Regression model can be used when it is an individual predictor (independent) variable, but it is not ideal because an individual predictor variable would not reveal meaningful discoveries as compared to analysing

multiple contributing or influencing factors at the same time (Stoltzfus, 2011). Linear Regression would generate a best fit straight line between the target (dependent) variable and the predictor(s) (independent) variables (TechDifferences, 2018). This would reveal the distinct and unique contributions of each predictor variable, this technique is known as multivariate Linear Regression (Stoltzfus, 2011). The equation for a Linear Regression model with multiple predictor (independent) variables is as follows: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i$ the component of the equation is as follows; the left side Y represents the estimated continuous outcome whereas the right side of the equation represents the linear regression equation for all selected predictor variables in the created model (Stoltzfus, 2011). As mentioned, even though Linear Regression is the most common Regression technique, when it comes to a binary target (dependent) variable, Logistic Regression would be the preferred choice of technique (Stoltzfus, 2011). While Linear Regression aims at predicting continuous or numerical value outcomes, Logistic Regression aims at identifying the probability of an event such as (success or failure), binary values such as (0/1, true/false, yes/no, etc.) (Ray, 2015). The equation for a Logistic Regression: Probability of outcome(\hat{Y}_i) = $\frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i}}$. Logistic Regression, on the left side Y, represents the estimated probability of one binary outcome category (i) versus the other, instead of presenting an estimated continuous outcome in Linear Regression (Stoltzfus, 2011). On the right side of the equation, in Logistic Regression, the predictor (independent) variables are expressed on a logit scale as compared to Linear Regression which is expressed on an original linear format (Stoltzfus, 2011). Logistic Regression implements a logit scale transformation because of the basic parameters of the model where the binary outcome expressed as a probability must be between 0 and 1 (Stoltzfus, 2011). As compared to Decision Tree, both Regression and Decision Tree are data mining techniques used to solve analytical problems, however, they defer in such a way whereby Regression is used to predict numerical

or continuous values whereas Decision Tree would assign data into discrete categories (Chapple, 2018).

2.5 Ensemble Model Frameworks

An example would be if you want to buy a new car, the probability of going to the first car showroom and purchasing it based on the dealer's advice without seeking a second advice is highly improbable (Singh A. , 2018). Prior to visiting the car showroom, you had likely browsed a few websites with reviews from owners and comparisons of different car models (Singh A. , 2018). You are also likely to ask your family and friends for their feedback as well. In short, you would not conclude explicitly, but instead come to a decision that also considers the views of others (Singh A. , 2018). Which brings the attention towards ensemble methods where it uses the same idea in machine learning and data mining (Singh A. , 2018). Through a combination of decisions from multiple models to improve the overall performance (Singh A. , 2018).

Ensemble methods were first mentioned back in 1977 where Tukeys Twicing used an ensemble of two liner regression models (Narayanan, 2014). However, it has been as a challenge to identify the origin of ensemble methods. Using two linear regression models with the first model fitted to the data and the second to correct the error of the first model, has the potential to enhance predictive outcomes (Narayanan, 2014). Moreover, in today's data mining society, it is well-known among the people within the field of analytics (Narayanan, 2014) (Srivathsan, et al., 2024). Ensemble learning also known as ensemble methods refers to a combination of learners which are trained to solve the same problem (Tuysuzoglu, Birant, & Pala, 2017) (Srivathsan, et al., 2024). It is a machine learning technique whereby the predictions are combined into a single output that potentially has a better performance than an individual model (Tuysuzoglu, Birant, & Pala, 2017) (Asif, Zhao, Tang, & Zhu, 2024). Sometimes also referred to as ensemble modelling, a process of running two or more analytical models and

synthesizing it into one prediction outcome to improve the accuracy of prediction in predictive analytics and data mining (Rouse, 2015). It can be defined as a machine learning process to obtain better prediction performance by strategically combining various learning algorithms for prediction (Abuassba, Zhang, Luo, Shaheryar, & Ali, 2017) (TalukdeR & Akter, 2024). Ensembles have a reputation of reducing the risks of selecting the wrong models by aggregating candidate models (Abuassba, Zhang, Luo, Shaheryar, & Ali, 2017). The fundamental idea of ensemble learning was to combine weak learners into a strong learner, with the ability to provide better generalization error while reducing the over-fitting of outputs (Tuysuzoglu, Birant, & Pala, 2017). Different classification models may interpret and misclassify patterns; hence, accuracy can be improved by combining multiple classifiers to enhance decisions (Tuysuzoglu, Birant, & Pala, 2017). “Alone we can do so little and together we can do much”, this is a quote by Helen Keller in the 50’s as a reflection of achievements and success stories in real life (Valiance Solutions, 2016). This reflects prominently on the idea behind ensemble learning whereby there is an increase interest in ensuring a more accurate prediction and classification, therefore, ensemble method has proven to provide this solution in one of the most convincing ways (Valiance Solutions, 2016).

Ensemble modelling is proposed in this research because in predictive modelling, a single model based on one dataset could potentially contain bias, high variance or anomalies which will affect the prediction outcome (Rouse, 2015). Even when applying specific modelling techniques, there is a risk of such drawbacks (Rouse, 2015). The solution to overcome these problems would be to combine different models with varying strengths to reduce the limitations of a single model and provide improved outcomes (Rouse, 2015). Most of the errors made by a model are due to three main factors: uncertainty, noise, and bias (DeFilippi, 2018). By using ensemble methods, it can increase the final model's stability and reduce the previously mentioned errors (DeFilippi, 2018). When combining multiple models, it can reduce the

uncertainty, even if they are not good individually, it will not suffer from random errors from a single source (DeFilippi, 2018). The principle of ensemble methods would be to combine weak learners together to create a strong learner with the idea of many joining and emerging as one (DeFilippi, 2018).

There are 3 main advantages when applying ensemble learning or modelling which include: (1) more accurate prediction outcomes, (2) a more stable and robust model because by aggregating the results into multiple models, it potentially reduces noisy data as compared to individual models, (3) capturing of linear and non-linear relationships in data through ensemble modelling of 2 different models (Juhi, 2018) (Ravanshad, 2018). Also as stated by Madhu Narayanan, a multi-model ensemble provides a way to combine results from various learning models to potentially increase the accuracy and prediction of the overall model (Narayanan, 2014). He states that there are one of the main reasons for an effective outcome; in a dataset, there might be patterns which are not or cannot be captured by a single learning algorithm, meaning, different learning models have their own strengths and weaknesses and by combining them, it provides a platform to mitigate the weaknesses while exploiting the individual strengths (Narayanan, 2014).

There are various types of ensemble methods which could be implemented such as Boosting, Bagging, and Stacking. But in this research, the focus is on Stacking. Stacking is a significantly different approach of combining models with the concept of meta-learning (Valiance Solutions, 2016). This approach does not have any empirical formula for the weight function or any similar functionality as bagging or boosting (Valiance Solutions, 2016). The main idea behind the ensemble method of stacking is to use a different (new) model to correct the errors of the previous model, which translates to one model stacking on top of the other (Ravanshad, 2018). Stacking involves training a learning algorithm and combining the predictions with various other learning algorithms (Ravanshad, 2018). How stacking works is that, when a 2nd algorithm

or learning model (combiner algorithm) is used as the “2nd stage” to combine the results from the “1st stage” (Narayanan, 2014). An example of stacking ensemble method would be to combine decision tree and regression model to predict an outcome.

Stacking was introduced back in 1992 by Wolpert (Charan, 2017). It is an ensemble method which uses a multi-model approach to build a new model for an enhanced prediction outcome (Singh A. , 2018). Stacking is when a new model is trained by combining the predictions from two (or more) previous models (DeFilippi, 2018). Stacking can also be referred to as meta-learning, which literally translates to learning about learning (Dzeroski & Zenko, 2004). Ensemble methods can be referred to as blending whereby the numbers are blended to produce a prediction (DeFilippi, 2018). However, by adding more layers and models to the base learner, it may not result in a better predictive model (DeFilippi, 2018). To build a stacking ensemble method, both the base learner and meta-learner would need to be trained (Menahem, Rokach, & Elovici, 2009). As mentioned, stacking has the basis of two learners which are the base learner and meta-learner (Charan, 2017). Base learners and meta-learners are normal machine learning/data mining algorithms (Charan, 2017). The base learner has to be trained prior to combining the meta-learner; once the base model has been trained, they produced an output which will be used as the input for the meta-learner (Menahem, Rokach, & Elovici, 2009). Predictions from a model is used as input for the following sequential layer and combined to form a collection of new predictions (DeFilippi, 2018). During prediction, the output from the base learner is taken and combined with the meta-learner’s outcome to produce a combined final prediction (Menahem, Rokach, & Elovici, 2009). Base learners would fit normal data sets while the meta-learner would take on the predictions of the base learner (Charan, 2017).

In ensemble methods, weak learners or base models are used as the building blocks to design more complex models which would enhance the prediction outcomes (Rocca, 2019). This is because most of the time, the base models do not perform well individually because of high

bias or too much uncertainty (variance) to be robust (Rocca, 2019). The idea of ensemble methods would be to reduce bias and variance in such base learners by using a combination approach to produce a stronger and more robust model that achieve higher predictive accuracy (Rocca, 2019).

Stacking differs from bagging and boosting because of two characteristics: (1) Stacking are often considered as heterogeneous learners (various learning algorithms are combined) while bagging and boosting are known as homogeneous learners; (2) Stacking combines base models using meta-model approach while bagging and boosting uses what is known as deterministic algorithms (Rocca, 2019). By using stacking ensemble method, the approach process to provide advantages such as simplicity; improved performance; and capability of a combined model induced by various models (Menahem, Rokach, & Elovici, 2009).

Table 2: Feature Comparison between Ensemble Methods

Features	Stacking	Bagging	Boosting
Heterogeneous	✓	✗	✗
Homogeneous	✗	✓	✓
Base Learner	✓	✗	✗
Meta Learner	✓	✗	✗
Empirical Formula	✗	✓	✓
Deterministic Algorithms	✗	✓	✓
Blending	✓	✗	✗
Random Subspace	✗	✓	✗
Gradient Descent	✗	✗	✓
Minimize Variance	✓	✓	✗
Increase Predictive Accuracy	✓	✗	✓

2.6 Systematic Review of Healthcare Data Analytics

The systematic review focuses on identifying articles with relation to the scope of work through databases such as IEEE, ACM and OneSearch. This work only involved secondary data retrieval and analysis; no ethical approval required.

Literature Search Strategies (see Table 3). Researcher searched IEEE, ACM, and OneSearch for potential studies up to 31 July 2020. The following search terms: (healthcare data analytics OR data analytics, healthcare) AND (healthcare pattern profiling OR pattern profiling, healthcare) AND (ensemble model framework OR framework, ensemble model) AND (predictive analysis OR predictive analytics) AND (data mining). The search strategies with the Boolean or phrase operators were performed. Studies in English, available in full-text and conducted among humans were searched. Then, duplicates were removed, title and abstracts were screened for its suitability. Finally, articles with their full text were assessed for eligibility to be recruited into the analysis.

Table 3: Search Strategies

Search	Search Items	IEEE	ACM	One Search
#1	healthcare data analytics OR data analytics, healthcare	877	32670	2954
#2	healthcare pattern profiling OR pattern profiling, healthcare	66	6648	12
#3	ensemble model framework OR framework, ensemble model	1125	1125	2491
#4	predictive analysis OR predictive analytics	7512	37	7355
#5	data mining	33880	1300	51276
#6	#1 AND #4 AND #5	226	4	20
#7	#2 AND #4 AND #5	226	0	0
#8	#6 AND #7 AND #3	66	0	0

Inclusion criteria. Any studies that reported the approach of which predictive analysis was carried out and fulfilled the inclusion criteria were analysed. The inclusion criteria were as follows: (1) Predictive Analysis within the scope of healthcare. (2) Application and Predictive

Techniques used were stated clearly. (3) Studies were published within an English peer-review journal up to 31 July 2020. Other related studies were also included through careful review of the reference lists of related review articles and reverse-forward citation tracking. Studies were excluded if there were any duplicates, insufficient results, non-patient related or non-relevant studies upon further review.

Study selection. All relevant articles identified through the above databases were downloaded. Initially, de-duplication was performed. Researcher would screen each title and abstract for suitability based on the search strategies mentioned above. Then, full-text articles were assessed based on the inclusion criteria mentioned above.

Data extraction. The following data were extracted from every study: the name of author, year of publication, country, study design, participant characteristics, measures, confounding factors, and major findings. Researcher extracted the data and assessed the study quality of each article. Two investigators (NCKW, CN) individually extracted the data and assessed the study quality, while any discrepancies were resolved through a thorough discussion with (LXB) as the moderator.

Description of Studies

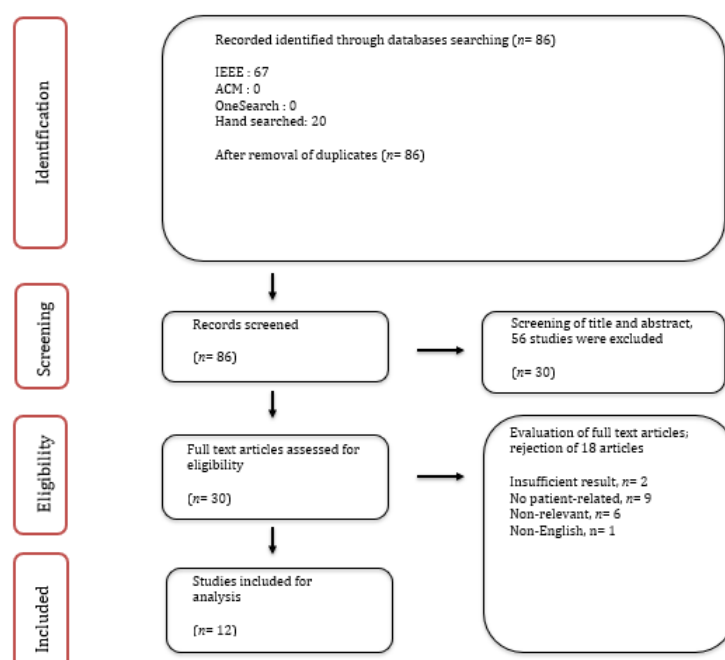


Figure 4: Prisma Flow Diagram of the Literature Screening Process

Description of included studies. Eighty-six manuscripts were identified in the initial screening as shown (see Figure 4). After removal of duplicate articles (n=86), a total of 86 studies were retrieved for further assessment. After screening for its suitability through title and abstract, 30 studies fulfilled both the inclusion and exclusion criteria. After careful evaluation of the 30 articles, only twelve studies were eligible for quantitative analysis in this study.

Table 4: Characteristics of Included Studies

Author	Year / Country	Measures	Factors
Giulia Bruno; et al. (Bruno, Cerquitelli, Chiusano, & Xiao, 2014)	2014, Italy	Clustering, multiple-level clustering approach, Classification model, decision tree	age, gender, HDLcholesterol
Bin Liu; et al. (Liu, et al., 2020)	2020, USA	A Bayesian Multi-Task and Feature Relationship Learning Approach	age, gender, ICD- codes, medications
Debby D. Wang; et al. (Wang, et al., 2017)	2017, Singapore	Cluster Analysis on Utilization Patterns, k-Medoids Clustering	total cost, length-of- stay
Ritesh Jain (Jain, Predictive Modeling for Chronic Conditions, 2015)	2015, USA	Predictive Modelling - Naive Bayes, Bayesian Network, Multilayer Perceptron Model, Logistic Regression, Decision Tree	blood pressure, BMI, age, ethnicity, smoking status
Tarek Abdunabi (Abdunabi, 2016)	2016, Canada	A Framework for Ensemble Predictive Modelling	N/A
Mohammad Hossein Tekieh (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012)	2012, Canada	Explores healthcare coverage disparity using a quantitative analysis on a large dataset from the United States (Decision Tree, Neural Networks)	access to care, age, poverty level of family, race/ethnicity
Prasan Kumar Sahoo; Suvendu Kumar Mohapatra (Sahoo, Mohapatra, & Wu, 2016)	2017, Taiwan	A Cloud-Enabled Big Data Analytic Platform (MapReduce)	N/A

David W. Bates; et al. (Bates, Sari, Ohno-Machado, Shah, & Escobar, 2014)	2014, USA	Using Analytics to Manage High Cost and High-Risk Patients	N/A
Yu-Kai Lin; et al. (Kai, Hsinchun, Brown, Hsing, & Jen, 2014)	2017, USA	Healthcare Predictive Analytics for Risk Profiling - Bayesian Multitask Learning Approach (BMLT)	age, bodyweight, male, smoking, ICD-codes, medications
Sai T. Moturu; et al. (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real-World Risk Modeling Application, 2007)	2007, USA	Predictive Future High-Cost Patients - Risk Modelling Approach	age, gender, county, race, marital status, disease, inpatient, outpatient, department, ICDcodes, mdc, bill charges
Hana Alharthi (Alharthi, 2018)	2018, Saudi Arabia	Healthcare Predictive Analytics focusing on Saudi Arabia	N/A
Pham, Hung N.; et al. (Pham, et al., 2019)	2019, Singapore	Predicting Hospital Readmission Patterns of Diabetic Patients using Ensemble Model and Cluster Analysis	time in hospital, total procedures, medications, outpatient, emergency, inpatient

This systematic review offers preliminary evidence regarding the scope of predictive analysis, pattern profiling within healthcare, healthcare analytics and ensemble modelling. The studies included within this systematic review focuses on the predictive techniques, methodological approach used to detect patient condition, and the exploration of a framework for ensemble predictive model. The main characteristics of the included studies are shown (see Table 2). Studies were conducted in Italy, UK, USA, Canada, Saudi Arabia, Taiwan, etc. Each study recorded an approximate of between 3,000 to 140,000 participants, respectively while the largest recorded participant was from the research using the database of Arizona Health Query (AZHQ) with 139,000 participants. As shown, some of the participant's characteristics include diabetic patients, national healthcare groups, medical expenditure surveys, Electronic Medical Records (EMRs) and Electronic Health Records (EHRs), while the predictive techniques which

were applied across the included studies were identified as well. Factors which had an influence in predicting patient's health were indicated as well. Finally, the major findings related to the studies were recorded.

Clustering Techniques. There are many predictive techniques which can be applied across different scenarios depending on the nature of the predictive outcomes. For instance, in the study by Giulia Bruno, et al. (Bruno, Cerquitelli, Chiusano, & Xiao, 2014), they used an approach known as clustering to identify groups of patients with similar characteristics and examination history in a dataset with a variable data distribution while applying a classification technique known as decision tree to perform prediction. Through this study, they identified that age, gender, and HDL (High-Density Lipoproteins) cholesterol were key drivers to determine diabetic patients. In comparison to the study by Bin Liu, et al. (Liu, et al., 2020), the approach they applied was a bayesian multi-task and feature relationship learning approach, which involves a complex bayesian technique while using association rules to identify feature relationship to identify diabetic patients. Similarly, the key drivers for diabetic patients were age and gender - additionally ICD-codes (International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO)) and medications a patient is taking were factors to determine a diabetic patient. In the study (Bruno, Cerquitelli, Chiusano, & Xiao, 2014), they used an approach known as clustering technique known as multi-level clustering but in another study by Debby D. Wang, et al. (Wang, et al., 2017), they applied a different technique known as k-Medoids clustering. In the study performed (Wang, et al., 2017), they aimed to identify patients of high-cost utilizers and lost-cost ones through clustering. Pham Hung N, et al. (Pham, et al., 2019) used cluster analysis to better understand the characteristics of patients who potentially might be readmitted, this allowed them to identify the patterns and trend of readmissions.

Predictive Techniques. Moving on, a study performed by Ritesh Jain (Jain, Predictive Modeling for Chronic Conditions, 2015), showed that he applied 5 predictive techniques, Naive Bayes, Bayesian Network, Multilayer Perceptron Model, Logistic Regression and Decision Tree to predict patients who were diagnosed with 2 chronic conditions (Asthma and Diabetes). Findings showed that by combining all 5 models, it managed to yield higher predictive accuracy. However, by combining these techniques, it would increase the complexity of the predictive model as well. Key factors which were mentioned include, blood pressure, BMI, age, ethnicity, and smoking status. Tarek Abdunabi (Abdunabi, 2016) performed a study on proposing a framework for ensemble predictive modelling by applying a technique known as fusion modelling - to create hybrid models through the proposed framework. This study is one of the few studies which focuses on proposing a framework for ensemble predictive modelling. Mohammad Hossein Tekieh (Tekieh, Analysis of Healthcare Coverage using Data Mining Techniques, 2012) explored healthcare coverage disparity in the United States by building two predictive models (decision tree and neural network) to study efficient factors in healthcare coverage. He managed to identify 4 factors which were access to care, age, poverty level of family and race/ethnicity as the key factors which would show disparity among healthcare coverages. Furthermore, he applied k-means clustering technique to discover groups of people with health coverage problems and inconsistencies. In his study, he demonstrated that the decision tree models provide higher accuracy than the models based on neural networks. Yu-Kai Lin, et al. (Kai, Hsinchun, Brown, Hsing, & Jen, 2014) proposed a bayesian multitask learning approach for healthcare predictive analytics for risk profiling within Electronic Health Records (EHR) which is like the study performed by Bin Liu, et al. (Liu, et al., 2020). In their study, they concluded that age, bodyweight, gender, smoking status, ICD-codes, and medications are key drivers when determining risk profiling among patients. Their analysis shows that the BMTL approach can create significant potential impacts on clinical practice in

reducing the failures and delays in preventive interventions. Sai T. Moturu, et al. (Moturu, Johnson, & Liu, Predicting Future High-Cost Patients: A Real-World Risk Modeling Application, 2007) performed a study to predict future high-cost patients within the Arizona Health Query dataset. He proposed solutions to statistical analytical techniques by applying non-random sampling (under-sampling or down-sampling, over-sampling, or up-sampling and a combination of both) and cost-sensitive learning. Through his analysis, he found that age, gender, county, race, marital status, disease, inpatient, outpatient, emergency department and pharmacy, ICD-codes, mdc (major diagnostic categories) and bill charges were factors influencing high-cost patients. Pham Hung N, et al. (Pham, et al., 2019) applied an ensemble method to predict readmission patterns of diabetic patients. By applying an ensemble model, they managed to achieve higher predictive accuracy as compared to constituent models and the baseline. These were the characteristics / factors which were used for the predictive analysis: time in hospital, total procedures, num medications, number outpatient log, number emergency log and number inpatient.

Big Data Analytics in Healthcare. An advanced analytical approach was applied in the study performed by Prasan Kumar Sahoo, et al. (Sahoo, Mohapatra, & Wu, 2016) where they applied a cloud-based analytical approach using MapReduce model on Electronic Health Records (EHR) as a solution to solving big data analysis problems. More important, the study focuses on the probabilistic data acquisition method is designed for the cloud-based healthcare system. David W. Bates, et al. (Bates, Sari, Ohno-Machado, Shah, & Escobar, 2014) performed an analytical study on how analytics can be applied to identify high-risk and high-cost patients. They proposed approaches and techniques such as decision tree or logistic regression to perform predictions while suggesting attributes to consider such as health problems, socioeconomic factors (poverty or racial minority) when associating with high-cost patients. Their focus also explores the efficient and effective use of predictive analytics. Hana Alharthi

(Alharthi, 2018) focuses her study on applying healthcare predictive analytics in Saudi Arabia - her argument revolves around health data analytics with the emphasis on predictive analytics as an emerging transformative tool to enable proactive and preventative treatment approach. She suggests that there is a lack of actionable knowledge towards a meaningful progress for better patient outcomes and improve quality of care.

Summary

Based on the review of the previous studies, which includes: Analysis of Healthcare Coverage using Data Mining Techniques, Predictive Modelling for Chronic Conditions and Predicting Future High-Costs Patients: A Real-World Risk Modelling Application, a few conclusions to be drawn would be (1) the application is in Foreign Countries (USA), minimal research and application has been applied in Malaysia's context (2) Focus is too clinical, there are too many clinical identifications and predictions while little focus has been put into potentially developing an ensemble framework which can be applied by practitioners who are not experts in the field (3) techniques applied are mathematical formulations, statistical calculations and technically too complex to be understood by the others who are not experts in the field (4) There are literatures and journals which applied the ensemble method known as bagging and boosting. However, another ensemble method known as stacking can be applied and a proposed ensemble framework would be an alternative to increase predictive accuracy, while it could potentially be a key predictive model to be applied across various industries in different fields of interest. It would be an alternative solution to increase predictive accuracy through ensemble model. Several research gaps which can be addressed – firstly, through the development of a practical framework to build ensemble model for classification predictions which can be applied across a wide range of applications, secondly, through the proposed framework, it would be tested to identify the effectiveness of the strategy to enhance and improve predictive accuracy and thirdly, a practical framework which can be applied by practitioners who are not

experts in the field. The analytics techniques which will be applied in this research include, clustering, decision tree, and regression. Besides that, there will be the addition of an ensemble learning method known as stacking where a combination of predictive models will be combined to perform a prediction. Also, a systematic review was done to enhance the literature review and to enhance the research quality.

3. Research Methodology

This chapter focuses on the Research Methodology. 3.1 describes the framework which has been proposed using the stacking ensemble model as a key feature in predictive analysis. It combines 3 aspects which includes: Feature Engineering, Feature Selection and Model Diversity & Flexibility to enhance the predictive outcomes and address the research gap. The Research Methodology can be broken down into 4 Phases, Phase 1: Objective, Phase 2: Data Preparation, Phase 3: Model Development and Phase 4: Model Blending. Each phase was explored and described in detail including the data cleaning processes, data processing techniques applied, and what predictive models were used to perform predictive analysis.

3.1 Proposed Framework for Ensemble Predictive Modelling

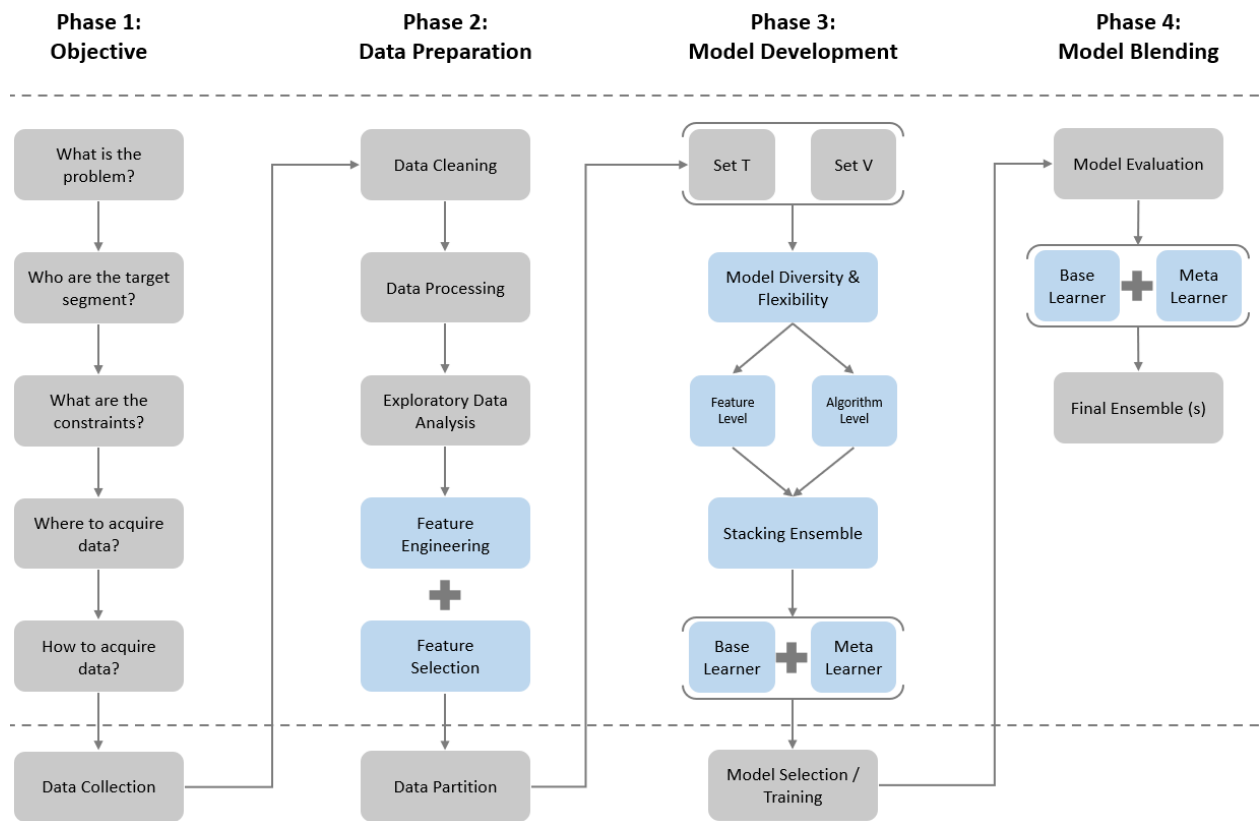


Figure 5: Proposed Practical Framework for Ensemble Model Building

3.2 Phase 1: Objective

- What is the problem?

As quoted by Albert Einstein, “if I have an hour to solve a problem, I will spend the first 55 minutes to analyse the problem and the last 5 minutes solving it” (Rose, 2021). The credibility of whether Einstein said this is up for debate, however, this approach might be the key and success factor to data analytics and data science. To reword it, maybe spend the first 55 minutes to ask the right questions (Rose, 2021). This approach is not restricted to data analytics or data science, but true to day-to-day interactions with people and problems. Just think of the numerous instances where one might have misunderstood or misread a situation or problem because one did not try to understand the problem and asking the right questions first? Questions are deemed to be the key to discovery and learning and asking the right questions would potentially reap breakthrough in business success (Rose, 2021) (Wunker, 2012). Asking

the right questions helps one to formulate and frame the right problems to look into as well. As stated by Harvard Business Review, asking the right questions, will allows one to tackle the right problems as well (Spradlin, 2012). With that, a clear and precise problem statement can be formulated. It can be described as a critical process to identify, discover, analyse, and solve problems (Annamalai, Azid, Kamaruddin, & Yeoh, 2013). It involves a detailed problem-solving process which includes discovering and identifying the problem, understanding of the problem, proposing a strategy to tackle the problem, researching alternative solutions, and actions to achieve the expected outcomes (Annamalai, Azid, Kamaruddin, & Yeoh, 2013). However, the most important step in solving any problem would be a clear, concise, and precise problem statement to help develop a better understanding of the problem and identify strategies to tackle these problems. The concept involves critical thinking and evaluation to formulate the problem statement. In this research, the problem revolves around the increasing medical costs which is an issue employer are looking extensively into, to reduce medical expenditure while potentially to identify employees (patients) with higher risk or cost to better understand the overall employee health population.

- Who is the target segment?

Approach the target segment through the funnel approach. This is one of the successful methods which many businesses and organizations apply called as a multiple staged funnel (Cohn, 2015). Identifying the right target segment allows one to drill down deeper into who are the audiences who to focus on when solving problem and who will be impacted the most. In this research, the focus is on the employers who are providing the insurance medical coverages to help them better understand the overall employee health population but at the same time to potentially reduce the ever-increasing medical expenditures through the prediction of high cost / risk employees.

- What are the constraints?

Constraints are often referred to as blockers or limitations in research. There will be constraints in terms of the data availability, tool availability, domain knowledge, or even time and scope constraints. These are often independent constraints and might affect the quality or outcome of a project. In some cases, there might be skill constraint or even data accessibility constraints. Below are some constraints pertaining to the research:

1. Data Constraints - the dataset provided in this research were lacking in detail such as patient medical health conditions such as BMI, blood pressure, etc which may further enhance the research outcome.
 2. Data Quality Constraints - there seems to be several data quality and completeness issue which were highlighted in the data preparation section. Data transformation, variable creation and other techniques were applied to ensure reliability and validity of producing research outcomes.
 3. Ethical Constraints - further considerations were put into protecting employees (patients) confidentiality, imposing constraints on study design, as this involves healthcare / medical related data. There were constraints in releasing too much data which may breach privacy acts / laws.
 4. Expertise Constraints - to ensure full model blending and state-of-the-art analytic algorithms which goes beyond a single stacking ensemble model, more knowledge, expertise, and experience in Machine Learning is required to produce more innovative solutions.
- Where to acquire data and how to acquire data?

In this research, data were obtained through a secondary source as data was prepared and masked by the HR department. These data were extracted from the HR's medical claim database which comprises of historical medical records / claims made by employees. These data were already in a structured format. Additionally, the data were discussed with the HR

department to ensure data completeness and concerns surrounding data quality issues were addressed.

There are several methods and tools which can be applied to extract / obtain data. It is largely dependent on the scope of research and what kind of data is required to perform the research. There are 2 categories of data types which are commonly referred to, qualitative or quantitative data. When referring to the scope of research, some research are considered as primary research, some are secondary research while others may be archival research. Primary research would require data to be obtained from primary sources, which is considered to be first-hand account of an event - they are often considered as raw information and evidence (Wales, 2021) (Streefkerk, 2018). They would generally represent original thinking, any reports on recent discoveries / events, or they might even share newfound information (Wales, 2021). They often provide direct and fresh evidence on the scope of research (Streefkerk, 2018). Secondary research would refer to secondary sources, these are often obtained through articles, scholar journals, past research, interpretation, analysis or even evaluation of primary sources (Wales, 2021). Some of the examples of primary sources include, logs, interviews, surveys, government documents, new research, or findings, etc. while secondary sources include, journal articles, textbooks, dissertations, etc. (Wales, 2021) (Streefkerk, 2018). Some data may be obtained through survey instruments, while some may require a tool called a crawler to crawl data from social media or the web. In some instances, data can also be obtained through an existing database.

- Data Collection

This research was in collaboration with a large conglomerate and the data was obtained from the Human Resource department. With respect to this research, historical medical claims data were given between the years of 2016 to 2018. These were provided in excel file format.

There were a total of 4 files which were provided, GP_2016_2018, SP_2016_2018, IP_2016_2018 and Industry Headcount_2016-2018.

3.3 Phase 2: Data Preparation

- Data Understanding - Healthcare Data

Description of Datasets

Below in Table 9 shows the description of the datasets which were used in this research. There are a total of 4 datasets which was used including: 1.GP_2016_2018, 2.SP_2016_2018, 3.IP_2016_2018 and 4.Industry_Headcount_2016_2018.

Table 5: Data Understanding - Healthcare Data

Dataset	Description
1. GP_2016_2018	This dataset has 39 variables and 161153 rows of data. It contains the patient's general information such as gender and age. It also contains industry information such as business industry, branch name and department name. The patient's claim activity and history (such as amount spent, amount insured, and date of visit) are also included together with the diagnosis and ICD codes. The dataset includes only General Practitioner (GP) visit type between the 3-year period of 2016 to 2018.
2. SP_2016_2018	This dataset has 74 variables and 20869 rows of data. It contains the patient's general information such as gender and age. It also contains industry information such as business industry, branch name and department name. The patient's claim activity and history (such as amount spent, amount insured, and date of visit) are also included together with the diagnosis and ICD codes. The dataset includes only Specialists (SP) visit type between the 3-year period of 2016 to 2018.
3. IP_2016_2018	This dataset has 119 variables and 14012 rows of data. It contains the patient's general information such as gender and age. It also contains industry information such as business industry, branch name and department name. The patient's claim activity and history (such as amount spent, amount insured, and date of visit) are also included together with the diagnosis and ICD codes. The dataset includes only In-Patient, Hospitalization (IP) type between the 3-year period of 2016 to 2018.

4. Industry Headcount_2016-2018	This dataset has 4 variables and 45 rows of data. It contains the industry headcount of each business industry and the total headcount for each year. The dataset includes data between the 3-year period of 2016 to 2018.
---------------------------------	---

- Data Cleaning and Data Processing

As observed, there are missing values, variations in wordings (some are highlighted, some are not, font colours variation), inconsistent naming conventions (some are capital letters, some are not), etc, these findings observed are usually present in raw data format, hence, data cleaning and pre-processing will be performed.

Capital letters were into a standardized format also known as normalization. Excel function, =PROPER, was used to capitalize only the first alphabet of each word. Once the results were standardized, it is more organized and structured. Minor replacements of capitalizing words such as Axa to AXA, Kl to KL, Jb to JB, and more were performed. When using the =PROPER function in excel it capitalizes only the first alphabet of each word hence, there are some words which had to be manually replace.

The =PROPER function was applied across 3 datasets: GP, SP, and IP. The following 6 variables (Entity Name, Corporate, Business Industry, BranchName, DeptName and Diagnosis) applied the =PROPER function to standardize and normalize the wording structure. The =DATE function was applied to convert date from text to actual date values. The data given in the initial dataset, DTDisability which refers to the date of visit or admission for each patient was stored as a text variable instead of date. Hence, text to actual dates were reformatted by using the =DATE function. This function was applied across the 3 datasets: GP, SP, and IP. Formatting date into an actual date format would present an opportunity to perform trend analysis of days and months as compared to date in text format.

Next, moving into identifying anomalies such as missing or null values, spelling errors, format errors, and data anomaly (e.g., negative age). As these data will affect the outcome of analysis, hence, it is best to remove, replace or transform these data. For missing or null values, a process called imputation or replacement was performed to replace the missing values with average/mean/median values, depending on the data structure. However, in some cases, if the percentage of missing values are too high or low, the entire column will be removed entirely - provided the column would not affect the outcome of the latter analysis. Data anomalies are abnormal data which are found in the dataset during data cleaning, as shown above, there is a patient with negative age (-75), this is not possible, hence it was assumed that there was a wrong input and replace the value with a positive age.

An example of missing or null value columns would be Start/Expiry Date and MC columns. For Start/Expiry Date, refer to DTDisability (date of claims) to replace the Start/Expiry date of the medical insurance year coverage. Similarly, under MC, there were many missing values which imputation or replacement had to be performed. In this case, missing values were replaced with 0. The solution was assumed given the fact that missing MC days would reflect that there were 0 MC taken. Another anomaly would be error in the data given as shown above under BranchName, Monash University Malaysia Sdn, it is supposed to be "Sdn Bhd". As this was just a missing "Bhd" value, it was replaced to standardize all anomalies.

Anomalies were found with relation to the variable format (format errors), which was set in the initial dataset. This resulted in an error in the values shown under Emp Annual Limit and Dep Annual Limit. A simple solution did the trick by converting the format to accounting in excel.

- Exploratory Data Analysis

Exploratory Data Analysis includes 2 sections which are Data Understanding and Descriptive Analysis. Data Understanding provides a complete overview of the collected data and variables which might be important and data to be included or excluded from the analysis. Descriptive

Analysis involves graphical representations such as bar charts, pie charts, histograms, line graphs, clustering (to better understand the behaviour of individuals with similar characteristics) – another form of descriptive analysis would involve the building of an analytical dashboard to display the various graphical representations on one canvas.

- Feature Engineering

Once removal of any irrelevant variables and columns were completed, the next phase involve variable creation. Variable creation is the process whereby new variables are created in accordance with the requirements of the analysis. In some cases, variable creation is done to minimize variation in a variable such as amount incurred, there could be a large variation of values and decimals. Thus, the approach taken to minimize variance would be grouping into ranges by creating a new variable. Other scenarios could be with reference to the target value, in order to perform prediction, a target value has to be selected. Some datasets do not include the target value and it has to be churned out by the analyst prior to running the predictive model.

Table 6: Feature Engineering - Healthcare Data

Variable Creation		
1.GP_2016_2018	2.SP_2016_2018	3.IP_2016_2018
Variable	Description	
Year	Indicating the year (2016, 2017 and 2018)	
DateofClaims	Indicating the actual date of claims	
ClaimFrequency	Indicating the number of claims a patient has made in a calendar year	
ClaimFrequencyGroup	Indicating patient's claim frequency based on groups	
PatientAgeGroup	Indicating patient's age based on groups	
AmtIncurredRange	Range of spending amount of a patient given a single visit	
TotalAmtIncurred	Total spending amount of a patient in a calendar year	
TotalAmtIncurredRange	Range of total spending amount of a patient in a calendar year	
AmtInsuredRange	Range of insured amount of a patient by the insurance provider given a single visit	
TotalAmtInsured	Total insured amount of a patient by the insurance provider in a calendar year	
TotalAmtInsuredRange	Range of total insured amount of a patient by the insurance provider in a calendar year	

TotalRemainingAmt	Total annual insurance coverage remaining limit in a given year upon deducting the total insured amount
TotalRemainingAmtRange	Range of total annual insurance coverage remaining limit in a given year upon deducting the total insured amount

Above in Table 6, shows the variables created to enhance the analysis process. A total of 10 variables were created which has been listed above. Year is to allow for an easier filter, when necessary, as the dates given in the dataset are generally in days; hence, year variable was created to represent those values. Date of Claims is a variable where the =DATE formula was applied to convert (DTDisability) text to date format. A variable in date format would enable better analysis such as identifying trend lines in days or months. Claim Frequency was created because there were no indicators to show the number of claims made by each patient (variable was created using formulas applied in Microsoft Power BI). Patient Age Group was created because there were too many varying ages, grouping the ages would minimize variance and provide better trend lines. Amount Incurred and Insured Range are created to minimize variance by grouping the amounts into ranges. During descriptive and predictive analysis, variables with high variance may affect the analysis outcomes; hence, the decision to group them into ranges. Total Amount Incurred and Insured are variables created to indicate the total amount spent or insured by the insurance provider as the general amount only indicates the amount spent or insured given a single visit to the clinic or hospital. To have a better indicator of the total amount spent and insured would be to create a variable to indicate the yearly spent and insured amount. This would also help in identifying the total annual insurance coverage remaining amount (another variable created) which will show how much remaining coverage a patient / an employee has remaining in a calendar year based on the deduction of the total insured amount. Total Amount Incurred, Insured and Remaining Range are all to minimize variance by grouping the values into distinct ranges.

- Feature Engineering (Target)

As discussed with the Group HR, no referencing or benchmark was used to categorized “RiskLevel” of patient. To identify “RiskLevel” of a patient, more exploration was needed to better understand the data and how different levels of conditions could be formulated to fulfil the Target. Firstly, there was a tagging of “LTM” which referred to Long Term Medication as shown next to the highlighted column with an indicator of “Y” or “N” - “Y” means yes to long term medication while “N” means no to long term medication. Patient with a “tagging” of LTM can be referred to as higher risk patients as they have certain chronic conditions. Secondly, based on a research by Daniela Koller (Koller, et al., 2014), he labelled 46 different conditions as chronic conditions hence, using the same labelling, patients who were diagnosed with similar conditions as higher risk were labelled, this is because there is a probability for patients as such to be diagnosed with chronic conditions and could potentially be a high-risk patient. Thirdly, a variable called “TotalRemainingAmt” was created to better understand the usage of a patient, if a patient has fully utilized their medical coverage provided by the employer. Annual Limit value was derived by deducting the TotalAmtInsured for the given year, which was how the remaining coverage value was generated. Once the Remaining Amount values were obtained, an 80-20 rule split to label patients who had below 20% of the remaining amount as higher risk - this is because a patient has almost fully utilized the medical coverage amount. The 80-20 ruling was derived by Vilfredo Federico Damaso Pareto (Kruse, 2016) based on the Pareto Principle where most observations in life are not distributed evenly (Better Explained, n.d.). The 80-20 rule goes by the saying as 20% of the sales reps generate 80% of total sales, 20% of customers account for 80% of total profits, etc. Hence, in this scenario, 80% of the medical coverage were used up with a remaining of 20%.

- Feature Selection

Due to the ever-increasing volume and variance in datasets, it has become an obligation to apply feature selection. It can be considered a good practice to choose the features (variables

or predictors) which will be useful at predicting (Y) as it is common to have features which are irrelevant and meaningless but noise (Prabhakaran, 2018). Some problems such as overfitting and the curse of dimensionality has resulted in the idea behind feature selection. Feature selection is also referred to as variable or attribute selection (Hoque, Singh, & Bhattacharyya, 2018). As mentioned by Mary Walowe Mwadulo (Mwadulo, 2016), recently, the need to apply feature selection have been increasing exponentially because of the large number of high dimensional features. The size and number of features can be reduced to manageable sizes through feature selection (Mwadulo, 2016). Feature selection is a process of selecting a subset of relevant/significant features (variables or predictors) from all features without any transformation (preserving the interpretation) while validating it with regards to the analysis objective and to build predictive models (Rawale, 2018) (Jovic, Brkic, & Bogunovic, 2015). By selecting a subset of relevant features, it performs a process of reducing the number of input variables before developing the predictive model (Brownlee, 2019). The process of removing irrelevant features would enable the predictive model and algorithms to concentrate primarily on the relevant data which are useful for the analysis and predictions (Hoque, Singh, & Bhattacharyya, 2018). It is intended to ensure the most useful features are used to perform prediction.

There are various feature selection methods which can be applied during the data pre-processing process to achieve an effective and efficient data reduction (Jovic, Brkic, & Bogunovic, 2015). Dataset size reduction can be achieved in two ways: reducing feature set or reducing sample set (Jovic, Brkic, & Bogunovic, 2015). However, in this research, the focus would be on reducing feature set. It is an important criterion since there are many features in the dataset which could potentially lead to model overfitting. There is a rule that states “garbage-in garbage-out” which is why data being fed to a predictive model must be relevant. An example would be the variables of “Name” or “ID”, poor quality input would lead to poor

quality output (Agarwal, 2019). Additional drawbacks of datasets with high dimensions include high computational demand while constructing models with datasets of many features, poor predictive accuracy, and difficulty in understanding data (Jovic, Brkic, & Bogunovic, 2015) (Chandrashekar & Sahin, 2014). Thus, the objectives to perform feature selection aims to solve the above drawbacks by reducing time and computational demand during analysis, improving predictive accuracy while removing irrelevant features to ensure better understanding of the data at hand. Also, a simpler predictive model which uses 10 variables would be easier to interpret and understand as compared to a predictive model which uses 100 variables (Charfaoui, 2020).

The 3 most common methods for feature selection are Filter, Wrapper and Embedded method. Filter methods generally used as data pre-processing. Filtering is a common process performed to remove features (variables or predictors) based on domain knowledge. However, within the scope of feature selection, features are filtered based on the statistical scores generated with relation to the correlation with the target variable. Some statistical formulas used include, Anova, Pearson's Correlation and LDA (Kaushik, Introduction to Feature Selection Methods with an Example (or How to Select the Right Variables?), 2016). Filter method uses a feature ranking technique to select the features based on the usefulness of each feature (Choudhury, 2019). Next, wrapper methods focus on using a subset of the features to train the model. Based on the outcomes, there is an option to add or remove features from the subset (Kaushik, Introduction to Feature Selection Methods with an Example (or How to Select the Right Variables?), 2016). But wrapper methods are very time consuming and computationally expensive. Common example of wrapper method includes forward selection, backward elimination, or recursive feature elimination (Kaushik, Introduction to Feature Selection Methods with an Example (or How to Select the Right Variables?), 2016). Last but not least, would be embedded methods which the method in itself is a combination of both filter and

wrapper methods (Kaushik, Introduction to Feature Selection Methods with an Example (or How to Select the Right Variables?), 2016). As the name suggests, the feature selection methods are embedded or built-in in the algorithms itself - an example would be Decision Tree model.

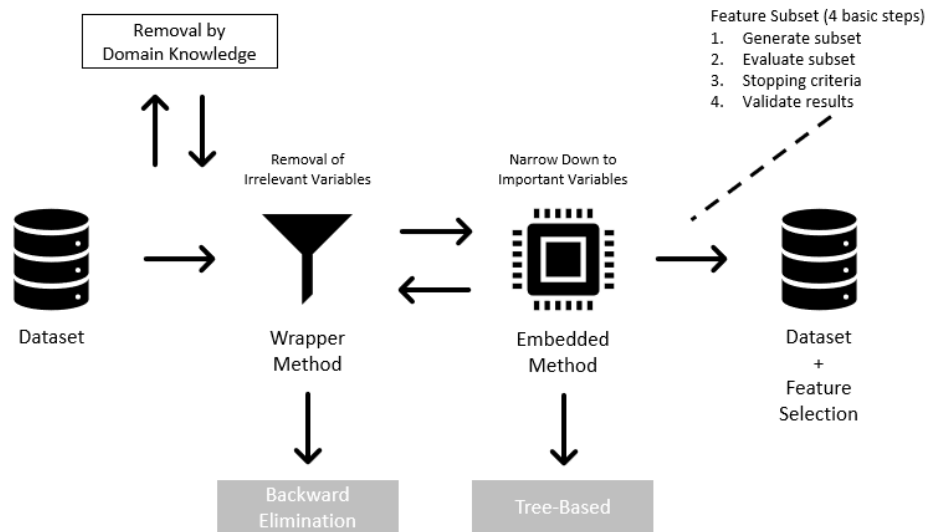


Figure 6: Feature Selection Process

Above (Figure 6) is a description of the feature selection process applied within this research. Once the first phase of data pre-processing, understanding of the data, etc. has been completed, the next phase would be to perform a feature selection process before moving into data analysis. The first step within the feature selection process would be to remove features based on domain knowledge, such as removal of “Name”, “ID”, etc. - features which have no impact or relevance to the objective. Next, the wrapper method was applied using the backward elimination technique, where it would start by using all the features then remove an insignificant feature at each iteration. This process would be repeated until the satisfactory performance have been achieved. With each varying analysis, the satisfactory performance may vary as well, hence, it is up to the analyst or domain knowledge required to justify the scoring. In this research, wrapper method was used to removed features with high variance or the set variance threshold (features with values which are the same), such as “Diagnosis” and “ICD Codes” where the

number of classes would exceed 512, hence these features should be removed to avoid overfitting or specificity where the analysis would yield too specific results. Upon the removal of all irrelevant, high variance, and insignificant features, tree-based embedded method was applied where a Decision Tree was selected to further help perform feature selection. Even though embedded methods are a combination of both filter and wrapper methods, in this research, wrapper method was applied prior to running the embedded method to further enhance the feature selection process as it would be able to minimize high feature dimensionality prior to applying the embedded method to achieve more efficient and effective feature selection. Embedded method as shown was used to scale down the features to achieve the research objective. Depending on the objective, the available resources and the desired degree of optimization, the selection process can be achieved in several ways. When performing feature selection, a feature subset process can be carried out accordingly as per the recommended steps: (i) generate a subset, (ii) evaluation the subset, (iii) create stopping criteria and (iv) to validate the results. Upon selected the feature subset, the evaluation process will begin in step 2, the process between step 1 and 2 will be repeated until it achieves the target set based on the stopping criteria. Then a validation of the results can be performed with relation to the objective of the research. When the process has been completed, the feature selected dataset would be ready to perform analysis.

Table 7: Feature Selection - Healthcare Data

GP Original	Removal based on Domain Knowledge	Wrapper Method	Embedded Method
EntityCode	BusinessIndustry	BusinessIndustry	BusinessIndustry
EntityName	IndustryHeadcount	PatientGender	PatientGender
BusinessIndustry	TotalHeadcount	PatientAge	PatientAge
IndustryHeadcount	EmpAnnualLimit	PatientAgeGroup	PatientAgeGroup
TotalHeadcount	PatientGender	Rel	Rel
Corporate	PatientAge	ClaimFrequency	ClaimFrequency
BranchName	PatientAgeGroup	ClaimFrequencyGroup	ClaimFrequencyGroup
DeptName	Rel	MCDays	MCDays

PatientID	ClaimFrequency	DateofClaims	ICDCategory
EmpAnnualLimit	ClaimFrequencyGroup	DTDISABILITY	AmtIncurred
PatientGender	MCDays	ICDCategory	AmtIncurredRange
PatientAge	Year	AmtIncurred	TotalAmtIncurred
PatientAgeGroup	DateofClaims	AmtIncurredRange	TotalAmtIncurredRange
Rel	DTDISABILITY	TotalAmtIncurred	AmtInsured
ClaimFrequency	Diagnosis	TotalAmtIncurredRange	AmtInsuredRange
ClaimFrequencyGroup	ICDCategory	AmtInsured	TotalAmtInsured
MCDays	ICDCode	AmtInsuredRange	TotalAmtInsuredRange
Year	AmtIncurred	TotalAmtInsured	TotalRemainingAmt
DateofClaims	AmtIncurredRange	TotalAmtInsuredRange	TotalRemainingAmtRange
DTDISABILITY	TotalAmtIncurred	TotalRemainingAmt	RiskLevel
Diagnosis	TotalAmtIncurredRange	TotalRemainingAmtRange	
ICDCategory	AmtInsured	RiskLevel	
ICDCode	AmtInsuredRange	LTM	
AmtIncurred	TotalAmtInsured	ExcessPaid	
AmtIncurredRange	TotalAmtInsuredRange	StartDate	
TotalAmtIncurred	TotalRemainingAmt	ExpiryDate	
TotalAmtIncurredRange	TotalRemainingAmtRange		
AmtInsured	RiskLevel		
AmtInsuredRange	LTM		
TotalAmtInsured	ExcessPaid		
TotalAmtInsuredRange	TypeOfClaims		
TotalRemainingAmt	StartDate		
TotalRemainingAmtRange	ExpiryDate		
RiskLevel			
LTM			
ExcessPaid			
TypeOfClaims			
StartDate			
ExpiryDate			

In the table above (Table 7), it shows the feature selection process to minimize the features in the dataset. GP Original represents the initial dataset then followed by Removal based on

Domain Knowledge to applying the selected feature selection methods of wrapper and embedded method. The initial dataset had 39 features, then it was narrowed down to 33 features after removal based on domain knowledge - subsequently, another 7 features were removed through the wrapper method and finally 6 features were removed upon applying the embedded method. A total of 19 features were removed through the feature selection process - the remaining 20 features (as shown in the embedded method) were selected for model building. These 20 features selected were evaluated and validated before a decision was made. They were confirmed based on the objectives of the research to better understand the claim patterns and behaviors of patients while also predicting risk of patients based on their past medical claims. The guidelines described above can be implemented by other research in the field of healthcare coverages for employers or insurance companies. Through these features, an understanding of the demographics of the patients, diagnoses, spent amount and amount insured together with the claim frequency and total remaining coverage would show an overall understanding of the claim behaviors of a patient while performing prediction based on past claim history.

- Data Partition

Data partition refers to allocation of data to perform various tasks such as model training, model evaluation and model testing. It is considered as one of the most crucial aspects of predictive modelling. For instance, the dataset should be split into training and validation subsets when building a predictive model, this is done to separate the samples into training and to perform validation on the trained data subset. The percentage of allocation and partition depends on the size of the available data. If the size of the available data is small, data partition might have a more significant impact on the quality of the model - furthermore, analysis might not be accurate as there might be bias in the model outcome. If the training data set is small, the predictive models and algorithms might not be able to capture and discover underlying patterns

within the dataset. Hence, data partition is a crucial part while building a predictive model. In this research, the data partition splitting was set at 70% training and 30% validation. Testing is generally not involved unless deployment of the model is involved, whereby the analyst might want to feed the model with actual data to test the performance and accuracy of prediction. In this case, deployment is not part of the scope of work, hence, testing was not involved.

3.4 Phase 3: Model Development

- Model Diversity and Flexibility

Feature Level

Model diversity can be achieved at feature level by using the same learning algorithm, with the same copies of training dataset, but the difference in features (Abdunabi, 2016) (Kuncheva, 2014) (Ranawana & Palade, 2006) (Rokach, 2009) (Sharkey, 2012). Features can be selected through manual selection or by applying feature selection algorithms. For each training iteration, the training subset may consist of features which overlaps or be completely different. Depending on the selected algorithms, some may have a built-in algorithm to incorporate model diversity at feature level such as Random Forest (Abdunabi, 2016). Differing algorithms may present various alternatives to diversity at feature level, such as a regression model can be achieved through manually selecting different subsets of the features, while in a classification model, some may choose to apply feature selection algorithms. By injecting the diversity and flexibility on the feature level, it presents the opportunity to achieve a more satisfactory predictive outcome.

Algorithm Level

The concept of model diversity at the algorithm level draws focus on selecting a variation of algorithms and techniques to train models by using the same training dataset and features selected to perform a comparison in terms of predictive performance and accuracy. Other alternatives to injecting diversity involve applying different parameters to the model such as in

a regression model to apply backward / forward / stepwise feature selection approach or in a classification model to adjust the maximum branch, maximum depth, or number of categorical split. Do bear in mind that any parameter adjustments may lead to increased complexity, under- or over-fitting. In this research, the selected algorithms to be used within the classification prediction would be regression and decision tree. As this research involves a classification prediction, hence, the chosen algorithms would reflect similarly. The model diversity of a predictive analysis depends on the nature of the prediction.

- Stacking Ensemble (Base Learner + Meta Learner)

The approach of selecting the suitable ensemble learning method also largely depends on the desired accuracy, the nature of the problem, as well as the computing resources (Abdunabi, 2016). Boosting and bagging techniques may be more complicated to implement, moreover, it may require an expert in the field to further analyse and explain the outcomes of these approaches. However, stacking is another ensemble method to use a combination of predictive models and algorithms to achieve higher predictive accuracy and performance while reducing the complexity of statistical, mathematical calculations. As mentioned in the gaps of knowledge, there are many research revolving around an increase in accuracy through these techniques but an easier predictive model and framework to perform a classification prediction using an ensemble approach is lacking. Stacking ensemble method focuses on the base and meta learner to be chosen.

The approach to select the learning models and algorithms, to build an ensemble predictive model generally largely depends on the nature of the problem, the experience, expertise and knowledge of the analyst or practitioner, computational cost, and scalability (Abdunabi, 2016). The proposed framework focuses on the practicality and simplicity approach where practitioners who are not experts in the field can apply an ensemble framework. Hence, the

choice to choose decision tree as the main algorithm for classification prediction as it is one of the easiest algorithm to understand due to the if-else algorithm approach.

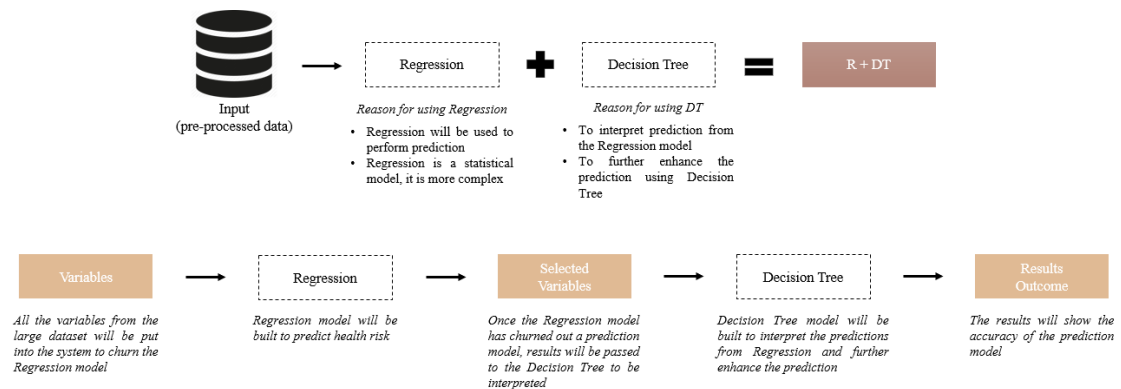


Figure 7: Ensemble Method Framework (Combination of Regression + Decision Tree)

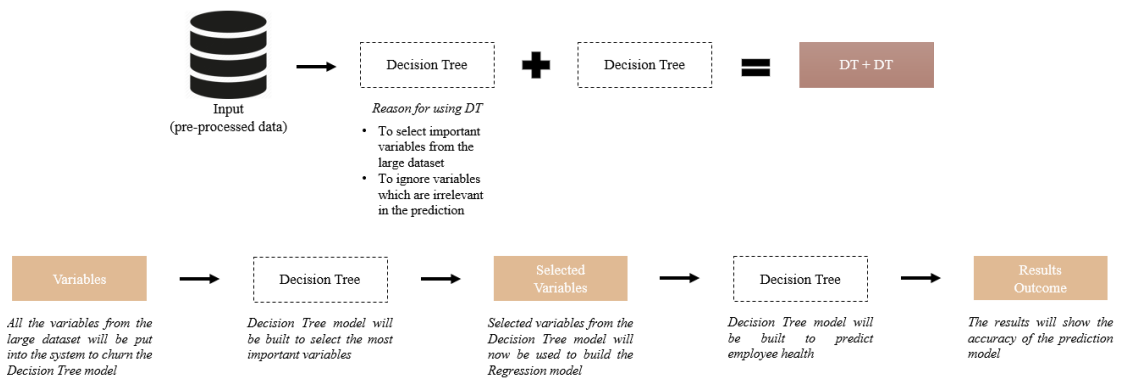


Figure 8: Ensemble Method Framework (Combination of Decision Tree + Decision Tree)

Above in Figure 7 shows the proposed ensemble method stacking model combining Regression as the base learner and Decision Tree as the meta-learner. This design is just a preliminary predictive framework. As shown in Figure 8, the predictive models also known as meta learner chosen would be Decision Tree model while the base learner would be Regression. In Figure 8, the meta learner and base learner would both be Decision Tree models. The learners have been chosen based on the idea of easy interpretation and usability because of the functionality it would benefit individuals who are not experts in the field which is why the proposed method to combine Regression and Decision Tree. Using Regression to perform prediction first then applying the Decision Tree model to interpret and enhance the prediction would allow for an easier interpretability as compared to using Decision Tree first as then Regression. Regression is a mathematical predictive model; hence, the nature would be more complex. A basic

overview would include the following: the dataset will be cleaned and prepared accordingly for prediction. Decision tree has the flexibility to handle missing values as well as selecting the important variables while ignoring irrelevant variables. Once decision tree has selected the variables, the output will be used to perform prediction using the decision tree model which is the meta-learner. There is a potential to test the ensemble model by combining decision tree as the meta-learner as well (Figure 8). This is the flexibility presented by an ensemble stacking method.

- Model Selection and Training

Model selection involves selecting the best model based on the nature of prediction. Once the ensemble method and model diversity has been decided, then model selection would be achieved by selecting the best performing models. Generally, in a predictive analytics scenario, many predictive models with varying settings will be adjusted to test the predictive outcomes - for example, in a decision tree predictive environment, the leaf size, split type (Gini or Entropy), split size (binary or multi), and other settings can be adjusted to test the predictive outcomes as well as the interpretability of the prediction. Do note that changes and adjustments to the parameters can have significant impact on the predictive performance of a model. Generally, a model is selected based on the predictive accuracy and the complexity of the model. A model that is too complex may result in under- or over-fitting which results in bias predictive outcomes, this is not ideal as it affects the overall predictive performance. On the other hand, model selection does not only involve the best performing model but the selection of the learning algorithms as well. This is large dependent on the nature of the problem, the domain knowledge and experience of the modeler, and the performance of each algorithm.

3.5 Phase 4: Model Blending

- Model Evaluation

Model evaluation is a significant process in predictive modelling (Abdunabi, 2016). It is more evident when performing ensemble predictive modelling, as the performance and diversity of a model/algorithm has to be evaluated completely - to assess the effectiveness and predictive accuracy of the ensemble model (Abdunabi, 2016). When the predictive models have been built, evaluation of the models must be performed to identify the best performing model. Best performing model means the model which produces the highest accuracy in terms of prediction. Through the evaluation of models, model comparison will be performed, and the highest accuracy prediction model can be selected. Model evaluation can include testing phase where the framework will be tested under various applications such as on different datasets and different predictive platforms. Moreover, if the results and predictive accuracy are not satisfactory, there is an alternative to restart at phase 1 (redefine objective(s), problem statement and to collect relevant data), 2 (to perform further data transformations and data processing to further prepare the data for prediction) or 3 (to relook into the chosen predictive models and algorithms or to apply different approaches depending on the nature of the prediction).

In this research, the aim was to test a practical ensemble framework using the ensemble method of stacking. In the analytics environment, robustness of a model is tested through applying different datasets in various scenarios to discover the robustness and specificity of the model. In addition, to test the tolerance of noisy data, anomalies, and errors to evaluate if the ensemble model is robust. A robust predictive model should achieve optimal performance by producing accurate and reliable results even with an increasing level of noisy data (Hu, Li, Wang, & Daggard, 2008). Moreover, the suggestion to apply the framework on different platforms to ensure that the framework can be applied on various platform which reduces the limitation of the framework. Platform testing was done on a proprietary software and an open-sourced software.

4. Analysis (Descriptive and Predictive)

Chapter 4 dives into the analysis performed in this research. The Analysis has been broken down into 2 sections, Descriptive Analysis and Predictive Analysis. 4.1 and 4.2 focuses on the healthcare data which has been described in detail in Chapter 3 while 4.3 focuses on the robustness testing through the healthcare data and 3 case studies which looked at how the framework and stacking ensemble model can be used in various classification scenarios. 4.1 drills down into the understanding of healthcare data and how employees (patients) are utilizing the medical insurance provided by the employer. Clustering was applied to better understand groups of individuals with similar health patterns / behaviours. 4.1 - descriptive analysis and clustering aims to address research objective 1 of understanding patterns in healthcare claims data while providing insights to overall employee population health. Predictive Analysis was broken 5 sections to further detail how each predictive model compares to each other. 4.2 - predictive analysis aims to address research objectives 2 and 3 of proposing an ensemble stacking model approach and validating the predictive accuracy and performance of the predictive model.

4.1 Descriptive Analysis - Healthcare Data

a) GP Overview

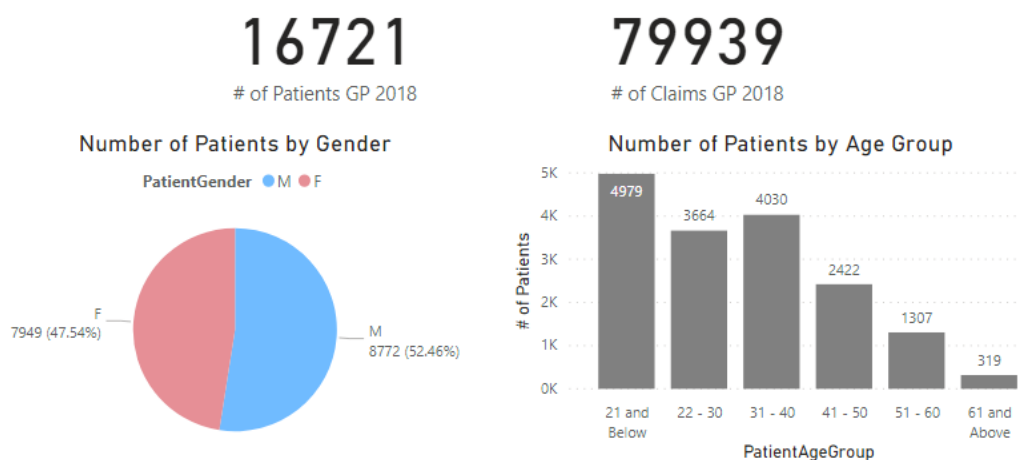


Figure 9: Basic Demographics of GP Patients 2018

Above shows a demographic overview of General Practitioner (GP) which includes all patients from employees to spouses and children who had made GP claims in year 2018. As shown above, there were a total of 16,721 individual patients who made GP claims in year 2018. A total of 79,939 claims were made by the 16,721 patients. 8,772 (52.46%) out of 16,721 patients were males while 7,949 (47.54%) out of 16,721 were females. Age of each patient have been categorized into groups under PatientAgeGroup. There are 6 categories which include, 21 and below, 22 - 30, 31 - 40, 41 - 50, 51 - 60 and 61 and Above. The highest age group who has the highest number of patients would be 21 and Below (4,979), followed by 31 - 40 (4030), 22 - 30 is the 3rd largest category which had 3,664 patients, while 4th, 5th and 6th were 41 - 50 (2422), 51 - 60 (1307) and 61 and Above (319) in a descending order, respectively.

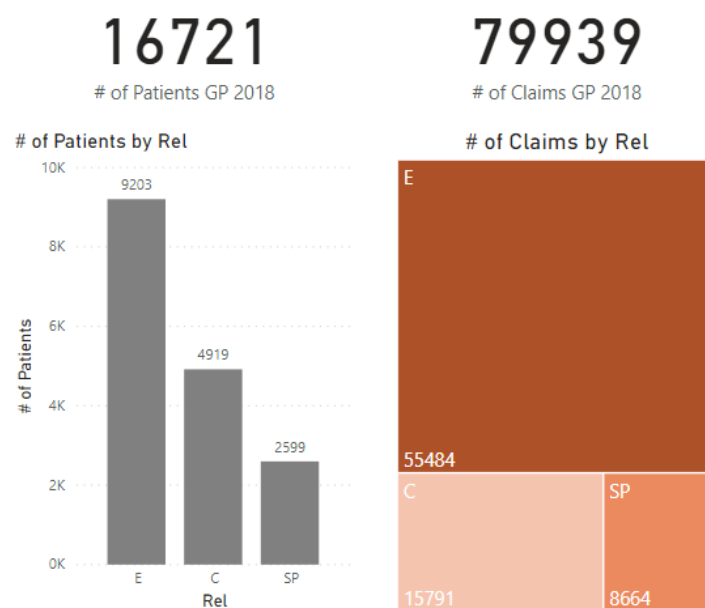


Figure 10: Basic Demographics of GP Patients by Relationship 2018

Next, the above illustration shows the demographics of the patients which were included under GP claims in year 2018. These illustrations would show the distribution of patients and claims made under GP in year 2018. On the left, the bar chart shows three categories of patients depending on their relationship (Rel); E represents employee, SP represents spouse and C represents child. Out of the total 16,721 patients, E occupies the highest percentage at 55% (9,203), while C is 2nd at 29.4% (4,919) and SP occupies the least at 15.6% (2,599). On the right

shows the number of claims made by each category of patient. E with the highest number of patients would have the highest number of claims at 55,484 claims made under GP in year 2018. C made 15,791 claims out of 79,939 while SP only made a total of 8,664 claims.

Diagnosis	# of Claims	% of Claims
Acute Upper Respiratory Infections Of Multiple And Unspecified Sites	8530	24.88%
Fever, Unspecified	6527	19.04%
Upper Respiratory Tract Infection	4510	13.16%
Diarrhoea And Gastroenteritis Of Presumed Infectious Origin(Acute Gastroenteritis /Dysentery)	3585	10.46%
Low Back Pain	2945	8.59%
Acute Pharyngitis	2368	6.91%
Atopic Dermatitis	1560	4.55%
Gastritis, Unspecified	1514	4.42%
Headache	1512	4.41%
Essential (Primary) Hypertension	1232	3.59%
Total	34283	100.00%

Figure 11: Top 10 Diagnoses under GP Claims 2018

Figure 11 shows the top 10 common diagnoses which were recorded by each respective clinician during a patient's visit to the clinic or hospital. A total of 34,283 claims were made within the top 10 diagnosis. "Acute Upper Respiratory Infections" was the most common diagnosis recorded with 8,530 claims while "Fever" was the 2nd most common diagnosis made at 6,527 claims and 3rd was "Upper Respiratory Tract Infection" at 4,510 claims recorded. "Acute Upper Respiratory Infections" and "Upper Respiratory Tract Infection" are both commonly known in layman terms as sore throat. There were two diagnoses which attracted attention, "Low Back Pain" and "Hypertension" - these conditions would be known as chronic conditions. Even though the number of claims made were only 2,945 and 1,232 respectively as shown under GP claims, it should be a cause of concern which would be further analyzed in the latter stages.

b) GP Overview - by Relationship

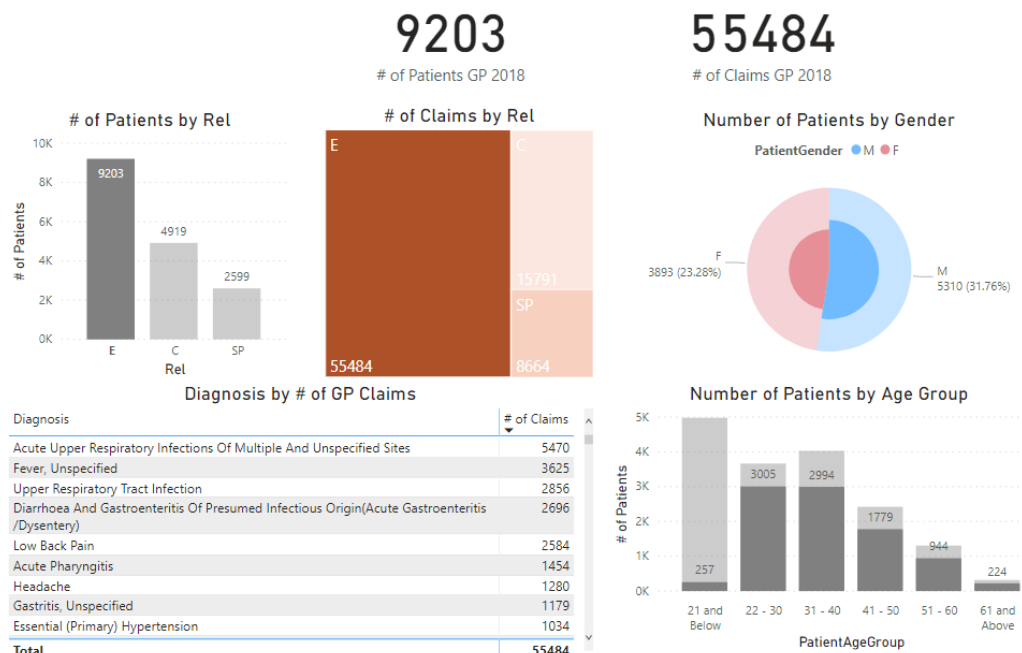


Figure 12: Diagnoses of GP Claims based on Patient Relationship (E) 2018

Here, patients were categorized based on relationships. Above is the largest category E, Employees. There were 9,203 patients who are tagged as employees. Out of these 9,203 patients, they made 55,484 claims in total under GP in year 2018. Based on number of patients by gender, M has 5,310 patients and F has 3,893 patients. The percentage as stated next to the number of patients signify the percentage over the total 16,721 patients. Looking at age group, the largest age group with the highest number of patients would be 22 - 30 with 3,005 patients, followed closely behind in 2nd would be 31 - 40 with 2,994 patients. 41 - 50 and 51 - 60 would be in 3rd and 4th with 1,779 and 944 patients, respectively. The last 2 groups would be 21 and below and 61 and Above, which have 257 and 224 patients. Similarly, based on the diagnoses by the number of claims, “Acute Upper Respiratory Infections” (5,470), “Fever” (3,625) and “Upper Respiratory Tract Infection” (2,856) while the 2 chronic conditions were recorded, “Low Back Pain” (2,584) and “Hypertension” (1,034).

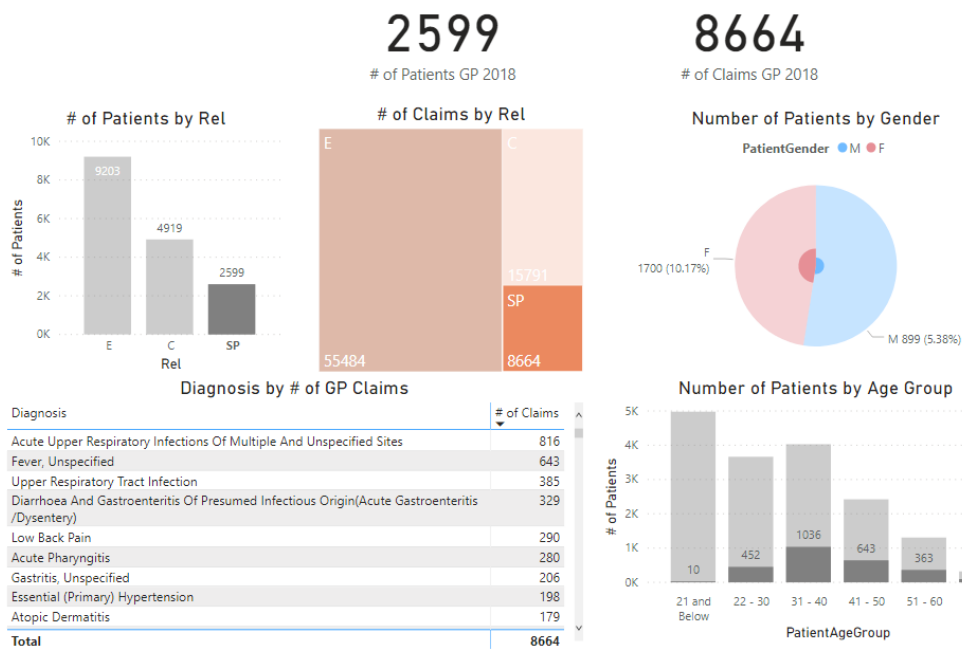


Figure 13: Diagnoses of GP Claims based on Patient Relationship (SP) 2018

Moving on in Figure 13 shows the patient relationship of Spouses (SP). There were 2,599 spouses who made a total of 8,664 claims under GP in year 2018. Based on number of patients by gender, there is a higher number of females (1,700) over males (899). Looking at the age group category, the largest segment would be 31 - 40 with (1,036) patients followed by 41 - 50 (643) then 22 - 30 (452). The last 3 groups in descending order would be 51 - 60 (363), 61 and Above (95) and 21 and Below (10). As shown, based on the diagnoses by the number of claims, “Acute Upper Respiratory Infections” (816), “Fever” (643) and “Upper Respiratory Tract Infection” (385) while the 2 chronic conditions were recorded, “Low Back Pain” (290) and “Hypertension” (198).

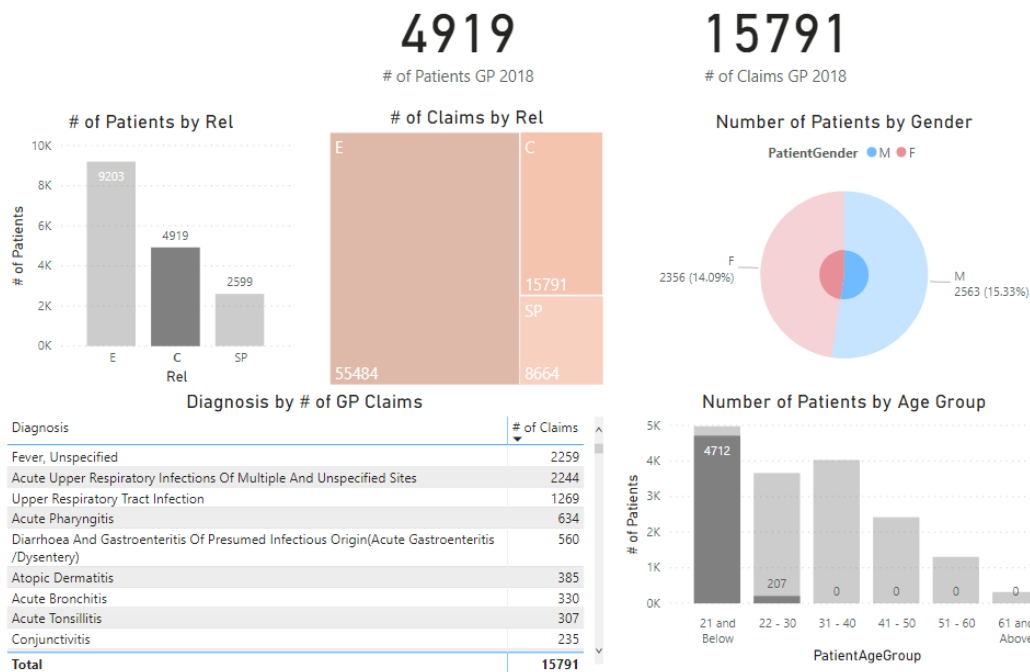


Figure 14: Diagnoses of GP Claims based on Patient Relationship (C) 2018

Children claims are the 2nd largest segment among the 3 segments of E, SP and C. As shown, there were 4,919 patients who made claims under GP in year 2018. Among these 4,919 patients, they made an accumulative 15,791 claims. Number of gender split among M and F were almost 50-50 with M at 2,563 and F at 2,356 patients, respectively. As expected, the patient age group category has only 2 categories, 21 and Below as well as 22 - 30; most companies do not cover for children who are above certain age groups. There were a total of 4,712 patients under 21 and Below while only 207 under 22 - 30. Based on the diagnoses by the number of claims, the diagnoses differ from E and SP slightly, with the highest number of claims being “Fever” (2,259), followed by “Acute Upper Respiratory Infections” (2,244) and “Upper Respiratory Tract Infection” (1,269).

c) GP Overview - by Claim Frequency Trends

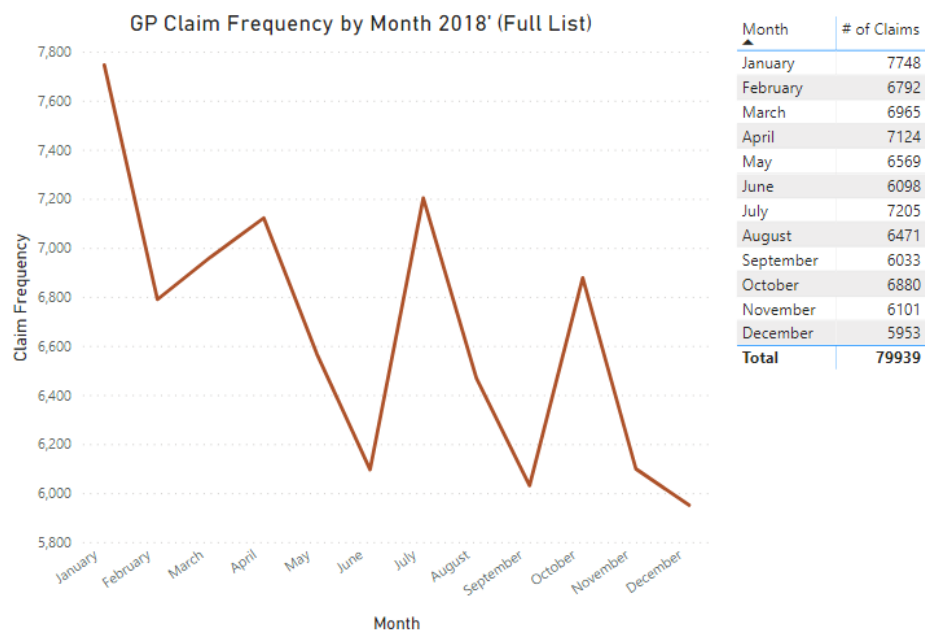


Figure 15: GP Claim Frequency by Month 2018'

The next section looks at the claim frequency trends based on the number of patients in a given month based on the date of claims. A total of 79,939 claims were made under GP in year 2018. As shown based on the line graph, 3 months (January, July, and April) had more than 7,000 claims in those months. January recorded the highest number of claims at 7,748 while 2nd was July at 7,205 claims and April at 7,124. These were the peak months. The lows were in the months of November, June, September, and December. These 4 months recorded sub 6,000 claims. In descending order, November had 6,101 claims, June had 6,098 claims, September had 6,033 claims and December recorded the lowest at 5,953 claims.

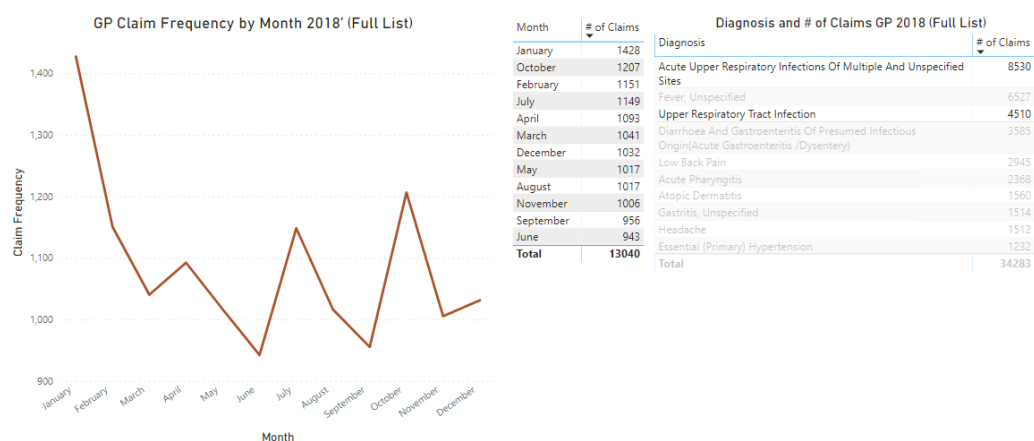


Figure 16: Highest Diagnoses based on GP Claims 2018

In Figure 16, the claim frequency trend was based on the diagnoses under GP claims in year 2018. Here, the most common diagnoses was shown to better understand the claim patterns. A total of 13,040 claims were made with relation to the diagnosis “Upper Respiratory Infections”. Out of the 13,040, 8,530 were “Acute Upper Respiratory Infections” and 4,510 were “Upper Respiratory Tract Infection”. The highest recorded months were January and October which exceed 1,200 claims: January (1,428) and October (1,207). The lows were below 1,000 claims in the months of September (956) and June (943).

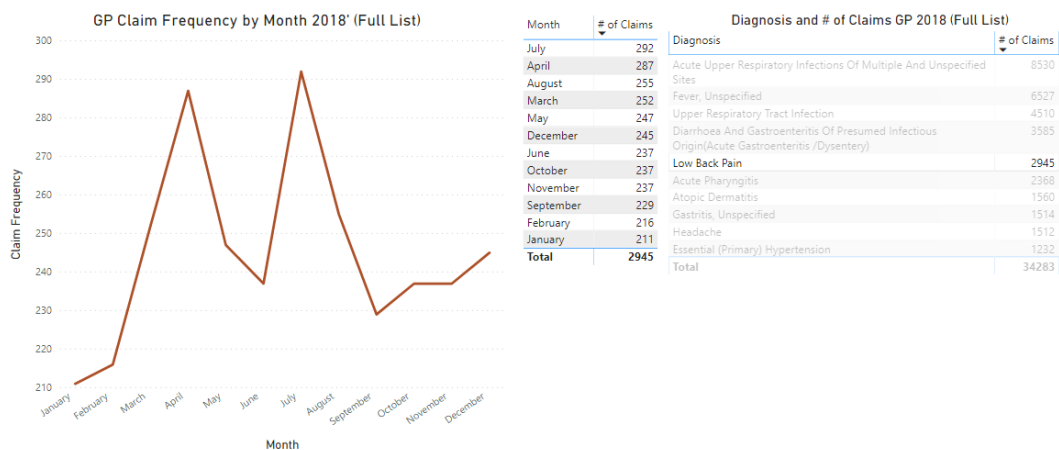


Figure 17: Chronic Condition Diagnosis based on GP Claims 2018

In Figure 17, the attention was shifted towards the chronic condition of “Low Back Pain”. A total of 2,945 claims were made with the diagnosis of “Low Back Pain”. There were 2 months which peaked considerably higher than the rest, which are July (292) and April (287). Lows

were recorded in the first two months of the year 2018, February (216) and January (211) – lows were sub 200 claims while high months were almost 300 claims.

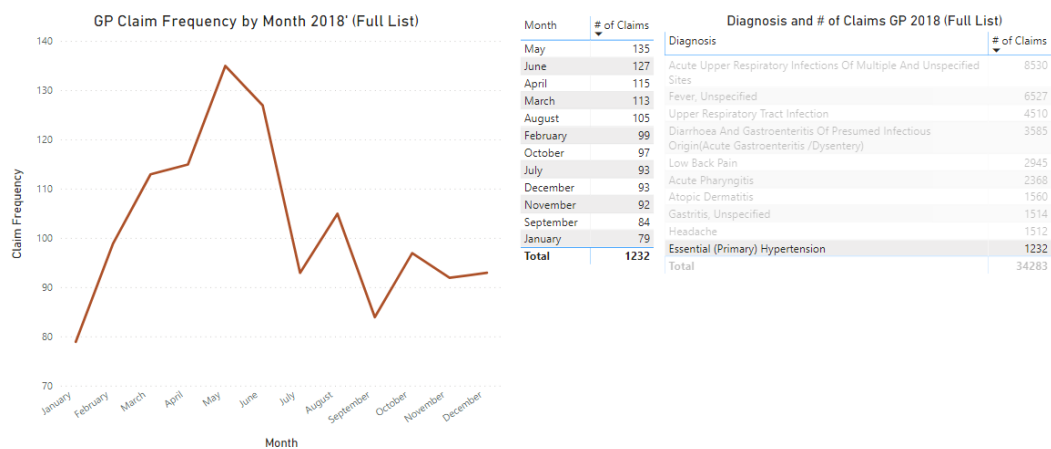


Figure 18: Chronic Condition Diagnosis 2 based on GP Claims 2018

In Figure 18 shows the 2nd chronic condition, “Hypertension”. There were a total of 1,232 claims with the diagnosis of “Hypertension”. High months were May and June which recorded 135 and 127 claims, respectively. Low months were September (84) and January (79) which recorded only claims of below 90.

d) GP Overview - Employee

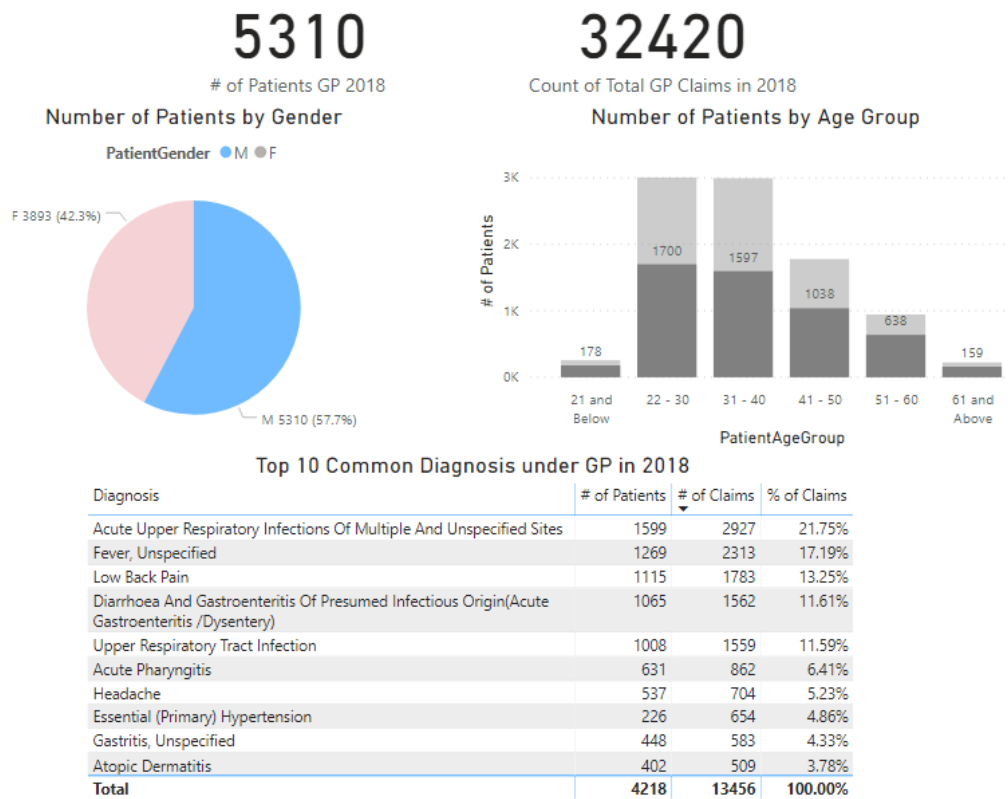


Figure 19: Demographics of Patients (Employees) GP Claims (M) 2018

Here, patients who were employees are split based on their gender. In Figure 19, there were 5,310 males who made a total of 32,420 claims in year 2018. Similarly, the distribution of age group shows that 22 - 30 (1,700) has the highest number of patients, followed by 31 - 40 (1,597) then 41 - 50 (1,038), 51 - 60 (638), 21 and Below (178) and 61 and Above (159). Within the top 10 common diagnosis, a total of 13,456 claims were made by 4,218 male employees. The common diagnoses based on the number of claims however has altered slightly, where among top 3: “Acute Upper Respiratory Infections” (5,470) and “Fever” (3,625) remained as top 2, the 3rd highest claimed diagnosis would be “Low Back Pain” (1,783). “Upper Respiratory Tract Infection” (1,559) dropped to 5th most common diagnosis while “Hypertension” recorded a total of 654 claims.

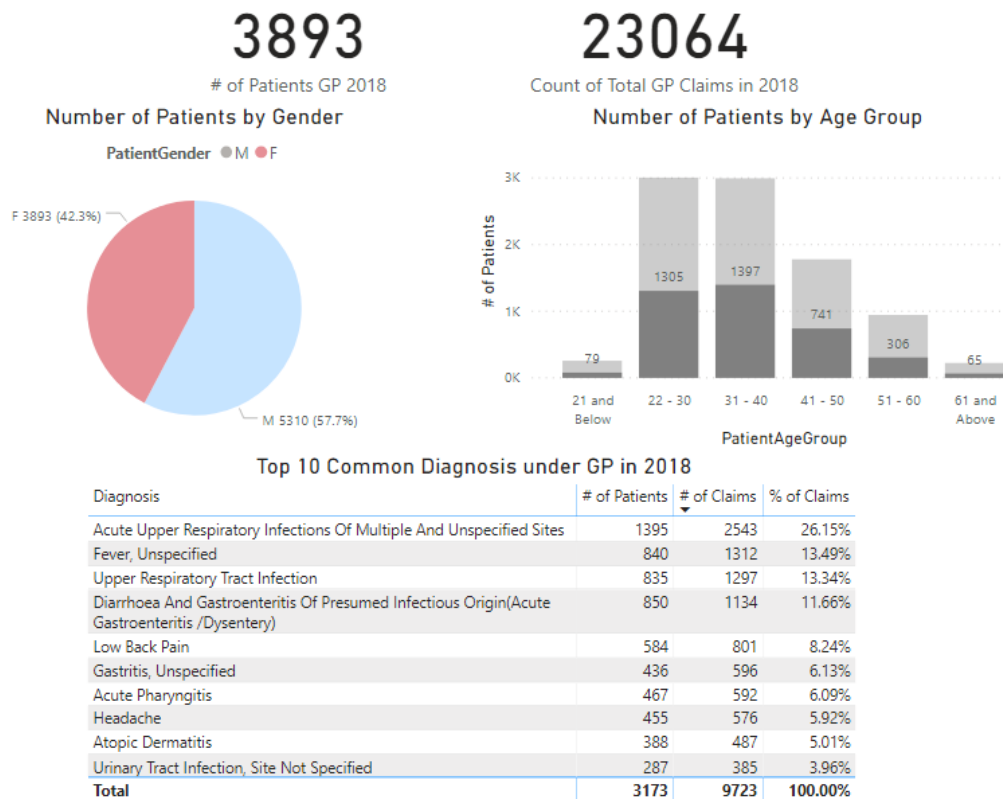


Figure 20: Demographics of Patients (Employees) GP Claims (F) 2018

In Figure 20, there were 3,893 patients who were female employees. They made a total of 23,064 claims under GP in year 2018. For the age group distribution, female employees differed slightly where the highest age group segment would be 31 - 40 (1,397) while 22 - 30 had (1,305) patients. The remaining age group of 41 - 50 (741), 51 - 60 (306), 21 and Below (79) and 61 and Above (65) would be categorized in a descending order. The top 10 diagnosis under GP for female employees showed that, a total of 9,723 claims were made by 3,173 patients. As mentioned, the top 3 common diagnoses based on the number of claims were “Acute Upper Respiratory Infections” (2,543), “Fever” (1,312) and “Upper Respiratory Tract Infection” (1,297). Females only recorded one chronic condition under the top 10 common diagnoses which is “Low Back Pain” (801).

Table 8: Top 10 Business Industry with Highest Patient (Employee) GP Claims 2018

Business Industry	Industry Headcount	# of Patients	% of Patients against Industry Headcount
-------------------	--------------------	---------------	--

Sunway Education Group	1563	1404	89.8%
Sunway Construction Group	1697	1302	76.7%
Sunway Resort Hotel & Spa	878	862	98.2%
Monash University	878	734	83.6%
Shopping Malls	725	576	79.4%
Group Security	717	575	80.2%
Sunway Property - PDD	506	410	81.0%
Sunway Lagoon Theme Park	565	392	69.4%
Sunway Lost World Theme Park	502	290	57.8%
Building Material	300	253	84.3%

The next section looks at the patients who are employees based on their business industry. As shown in Table 8, it has been filtered to show the top 10 business industry with the highest number of patients. Sunway Education Group (1,404) has the highest recorded number of patients (employees) who made claims under GP in year 2018, followed by Sunway Construction Group at (1,302) then Sunway Resort & Spa (862). Monash University (734), Shopping Malls (576), Group Security (575), Sunway Property – PDD (410), Sunway Lagoon Theme Park (392), Sunway Lost World Theme Park (290) and Building Material (253) made the top 10 business industry with the highest number of patients in a descending order. An additional analysis was to compare the number of patients who made claims against the industry headcount for each business industry to see the percentage. For Sunway Education Group, 89.8% of the employees made claims, while Sunway Construction Group 76.7% of the employees, and in 3rd Sunway Resort Hotel & Spa 98,2% of the employees made claims which is also the highest percentage among the top 10 business industry.

Table 9: Top 10 Business Industry with Highest GP Claims 2018

Business Industry	# of Claims	% of Claims (out of entire population)
Sunway Construction Group	7886	18.71%
Sunway Education Group	7557	17.93%
Sunway Resort Hotel & Spa	6190	14.68%
Group Security	4027	9.55%

Monash University	3996	9.48%
Shopping Malls	3970	9.42%
Sunway Lagoon Theme Park	2641	6.26%
Sunway Property - PDD	2116	5.02%
Building Material	1990	4.72%
Sunway Lost World Water Park	1784	4.23%

Table 9 shows the top 10 business industry with the highest number of claims. A collective total of 42,157 claims were made by the top 10 industries. Sunway Construction Group (7,886) had the highest number of claims made under GP in year 2018, followed by Sunway Education Group (7,557) claims while the 3rd largest industry was Sunway Resort Hotel & Spa (6,190) claims. The remaining 7 industries made up the top 10 business industry with the highest number of claims: Group Security (4,027), Monash University (3,996), Shopping Malls (3,970), Sunway Lagoon Theme Park (2,641), Sunway Property – PDD (2,116), Building Material (1,990) and Sunway Lost World Water Park (1,784).

Table 10: Top 10 Diagnoses based on # of GP Claims within Top 10 Business Industry

Diagnosis	# of Patients	# of Claims	% of Claims 17,595
Acute Upper Respiratory Infections	2276	4253	24.17%
Fever	1542	2695	15.32%
Upper Respiratory Tract Infection	1361	2129	12.10%
Diarrhoea and Gastroenteritis	1448	2067	11.75%
Low Back Pain	1252	1916	10.89%
Acute Pharyngitis	835	1110	6.31%
Headache	745	972	5.52%
Gastritis	680	902	5.13%
Essential (Primary) Hypertension	271	790	4.49%
Atopic Dermatitis	598	761	4.33%

Above in Table 10 shows the top 10 diagnoses made by the employees under GP in year 2018 among the top 10 business industry. Within this category, they made a total of 17,595 claims among the top 10 business industry. The top 10 diagnoses made by employees are similar to

the top 10 diagnoses as mentioned previously. “Acute Upper Respiratory Infections” (4,253) claims by 2,276 patients and is the most common diagnosis recorded by clinicians, followed by “Fever” (2,695) claims by 1,542 patients and the 3rd most common diagnosis would be “Upper Respiratory Tract Infection” (2,129) by 1,361 patients. While the 2 chronic conditions were recorded as well, with “Low Back Pain” recording (1,916) claims by 1,252 patients and “Hypertension” (790) claims by over 270 patients. The top 10 diagnoses among the top 10 business industry recorded a sum of 17,595 claims over the total of 55,484 claims by every patient (employee).

e) GP - Usage of Medical Coverage (Employee)

To determine the usage of medical coverage as provided by the employers, the approach taken was to perform a mathematical calculation to derive the total remaining amount available for an employee. To derive the total remaining amount value; each employee is given an employee annual limit signifying the yearly medical insurance coverage amount given by the employer. By taking the subtraction of annual limit minus total amount insured (AnnualLimit – TotalAmtInsured), the total remaining amount would be calculated. Total amount insured is a value derived by taking every claim performed by an employee in a given year and performing the mathematical solution of addition based on the employee’s insured amount per visit to the clinic or hospital.

Table 11: Patients (Employees) based on Total Remaining Amount Range

Total Remaining Amount Range	# of Patients (Employees)	% of Patients (Employees)
0 and Below	75	0.8%
1 to 1000	1337	14.5%
1001 - 2000	3782	41.1%
2001 - 3000	3142	34.1%
3001 - 4000	754	8.2%
4001 and Above	119	1.3%

Based on the findings present in Table 11, the focus will be drawn towards the highlighted rows of Total Remaining Amount Range of ≤ 1000 - reason being that the patients (employees) have either fully utilized or are on the verge of fully utilizing the yearly medical insurance coverage limit provided by the insurance provider. As shown in Table 8, there were a total of 1,412 patients within the 2 highlighted categories - where 0 and Below there were 75 patients (0.8%) and 1 - 1000 there were 1,337 patients (14.5%).

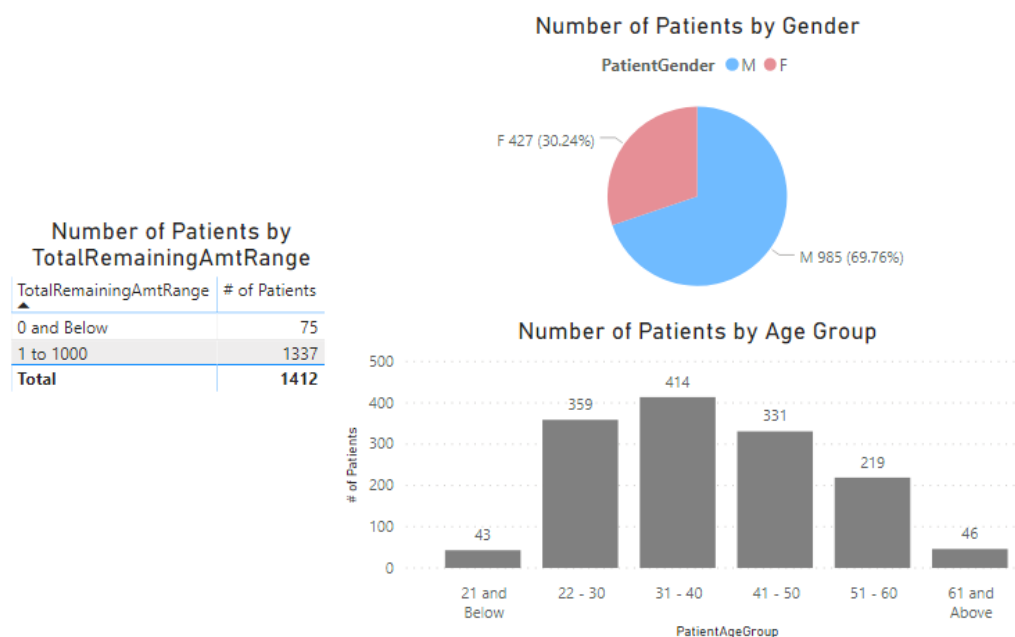


Figure 21: Number of Patients (Employees) based on Total Remaining Amount < 1000

This group of individuals who had a total remaining amount range of < 1,000 are comprised of 985 (69.8%) males and 427 (30.4%) females. 75 of them had a total remaining amount range of less than 0, while the remaining 1,337 patients had a remaining of 1 - 1000. Most patients come from the age group of 31 - 40 (414) while the 2nd highest age group would be 22 - 30 (359) followed by 41 - 50 (331). Age group 51 - 60 would come in at 4th place with 219 patients. 61 and Above and 21 and Below recorded 46 and 43 patients within these age groups, respectively.

Top 10 Business Industry with TotalRemainingAmtRange < 1000			Top 20 Diagnosis based on # of GP Claims 2018		
Business Industry	# of Patients	% of Patients	Diagnosis	# of Patients	# of Claims
Sunway Resort Hotel & Spa	225	19.46%	Acute Upper Respiratory Infections Of Multiple And Unspecified Sites	591	1320
Group Security	202	17.47%	Fever, Unspecified	522	1200
Sunway Construction Group	164	14.19%	Low Back Pain	538	983
Monash University	128	11.07%	Diarrhoea And Gastroenteritis Of Presumed Infectious Origin(Acute Gastroenteritis /Dysentery)	458	766
Shopping Malls	96	8.30%	Upper Respiratory Tract Infection	380	683
Sunway Lagoon Theme Park	94	8.13%	Essential (Primary) Hypertension	145	439
Sunway Education Group	71	6.14%	Gastritis, Unspecified	283	436
Building Material	65	5.62%	Acute Pharyngitis	277	413
Trading & Manufacturing	59	5.10%	Headache	244	345
Quarry	53	4.58%	Atopic Dermatitis	171	235
Total	1156	100.00%	Migraine	135	190
			Myalgia	126	177
			Conjunctivitis	126	158
			Asthma	76	157
			Acute Tonsillitis	112	156
			Gout	54	154
			Urinary Tract Infection, Site Not Specified	108	152
			Cough,Fever, Unspecified	64	151
			Acute Bronchitis	102	145
			Hyperlipidaemia, Unspecified,Essential (Primary) Hypertension	46	120
			Total	1346	8380

Figure 22: Business Industry and Top 20 Diagnosis based on Total Remaining Amount < 1000

Looking at the top 10 business industry with the total remaining amount range of less than 1000, 1,156 patients occupy this segment out of the total 1,412 patients. There were 2 business industry which recorded over 200 patients: Sunway Resort Hotel & Spa (225) and Group Security (202) while there were another 2 which had between 100 to 199 patients: Sunway Construction Group (164) and Monash University (128). The remaining 6 business industry recorded less than 100 patients each, where Shopping Malls had 96 patients, Sunway Lagoon Theme Park had 94 patients, Sunway Education Group recorded 71 patients, Building Material recorded 65 patients, Trading & Manufacturing recorded 59 patients and finally, Quarry had 53 patients. Next, the idea was to identify what are the top 20 diagnosis among this category (total remaining amount range of < 1000) of patients. There were 1,340 patients within this category who contributed to 8,234 claims in total. The 3 most commonly recorded diagnoses would be “Acute Upper Respiratory Infections” (1,320) claims made by 591 patients, “Fever” (1,200) claims made by 522 patients and “Low Back Pain” (983) claims made by 538 patients. Again, “Low Back Pain” (983) and “Hypertension” (439) were the 2 chronic conditions which were within the top 10 diagnosis based on the number of claims.

Total Amount Insured Range based on Top 20 Diagnosis

TotalAmtInsuredRange	# of Patients	# of Claims
601 - 900	485	2780
301 - 600	282	1412
1201 - 1500	252	1909
901 - 1200	199	1373
1501 - 1800	48	344
2101 - 3000	38	277
1801 - 2100	28	241
3001 and Above	14	44
Total	1346	8380

Figure 23: Total Amount Insured Range based on Top 20 Diagnosis under GP 2018

The total amount insured range were analyzed as well based on the top 20 diagnoses to better understand the spending patterns among the remaining amount range of < 1000. As shown in Figure 23, the total insured range starts from > 300 to above 3000. A total of 8,380 claims were made by 1,346 patients within this group category. That is the range of insured amount based on the top 20 diagnosis. The range of 601 - 900 has the highest number of claims (2,780) and patients (485). The 2nd highest range category would be 301 - 600 where there are 1,412 claims and 282 patients followed by the 3rd highest category which is 1,201 - 1,500, this category had higher total claims at 1,909 but less patients at 252 as compared to the 2nd category.

f) SP Overview

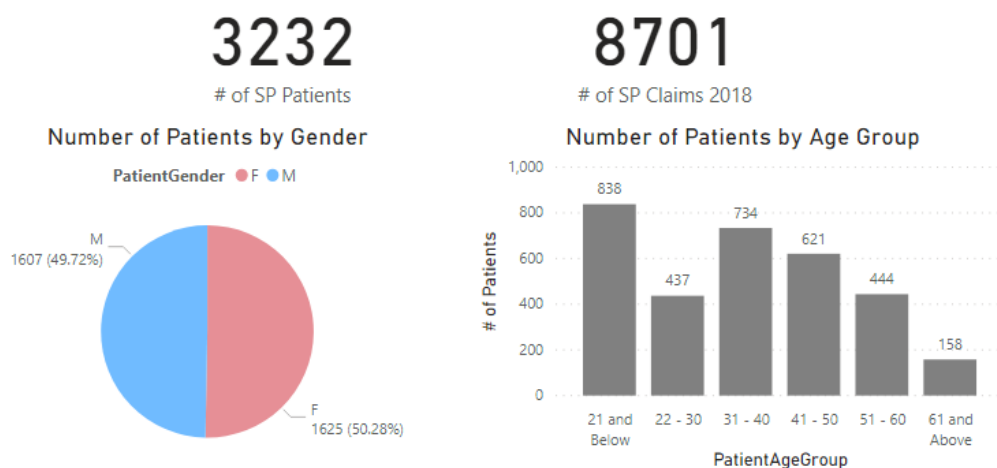


Figure 24: Basic Demographics of SP Patients 2018

Earlier sections were focusing on General Practitioner claims, in Figure 24 onwards, it focuses on Specialists claims. There were 3,232 patients who made 8,701 claims in year 2018 under SP. Out of these 3,232 patients, 1,607 (49.7%) were males while 1,625 (50.3%) were females. The highest patient age group comes from 21 and Below at 838 patients while the 2nd highest is 734 patients, age group of 31 - 40 and the 3rd is 444 patients, age group of 51 - 60. Age groups of 51 - 60 (444), 22 - 30 (437) and 61 and Above (158) are the remaining 3 age groups.

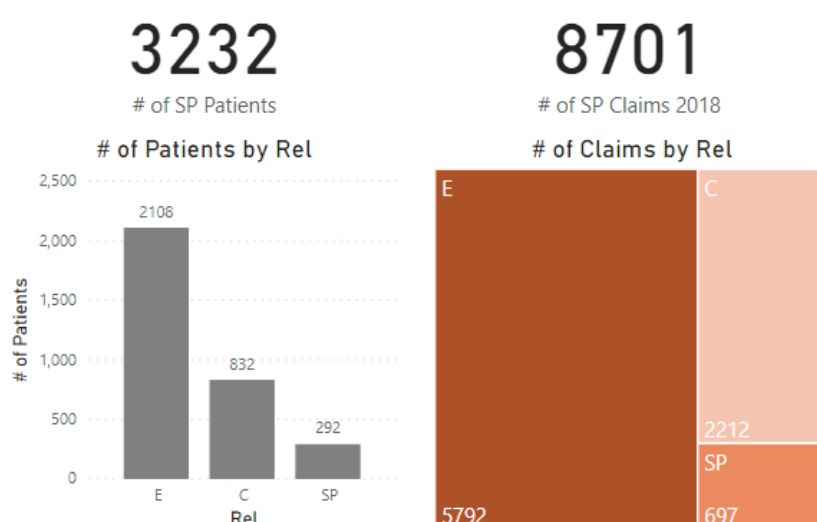


Figure 25: Basic Demographics of SP Patients by Relationship 2018

Out of the 3,232 patients, E representing employees recorded the highest number of patients at 2,108 while C representing child recorded 832 patients and SP representing spouse recorded the least number of patients at only 292. As mentioned, there were 8,701 claims made under SP in 2018. A large percentage at 66.6% (5,792) came under employee claims while child occupied 25.4% (2,212) of the total claims are spouse only occupied 8% (697).

Top 10 Diagnosis under SP in 2018

Diagnosis	# of Patients	# of Claims	% of Claims
Upper Respiratory Tract Infection	190	341	27.59%
Fever, Unspecified	189	281	22.73%
Cough	87	127	10.28%
Acute Nasopharyngitis (Common Cold)	91	119	9.63%
Hypertension	21	65	5.26%
Low Back Pain	29	65	5.26%
Need For Immunization Against Other Single Viral Diseases	48	64	5.18%
Other Disorders Of Eye And Adnexa	42	64	5.18%
Coronary Artery Disease	16	56	4.53%
Need For Immunization Against Other Combinations Of Infectious Diseases(Dpt/Hib/Ipv)	39	54	4.37%
Total	509	1236	100.00%

Figure 26: Top 10 Diagnoses under SP Claims 2018

Figure 26 shows the top 10 common diagnosis which were recorded by each respective specialist during a patient's visit to the specialist centre or hospital. Total claims within the top 10 diagnoses under SP were 1,236 made by 509 patients. "Upper Respiratory Infection" was the most common diagnosis recorded with 341 claims while "Fever" was the 2nd most common diagnosis made at 281 claims and 3rd is "Cough" at 127 claims recorded. The 2 chronic conditions which were present in GP claims were also present in SP claims: "Low Back Pain" and "Hypertension". The number of claims made were 65 for both conditions. Another cause of concern would be the diagnosis of "Coronary Artery Disease", which is a high-risk diagnosis, in 2018, 56 claims were made by 16 patients under SP.

g) SP Overview - Employee

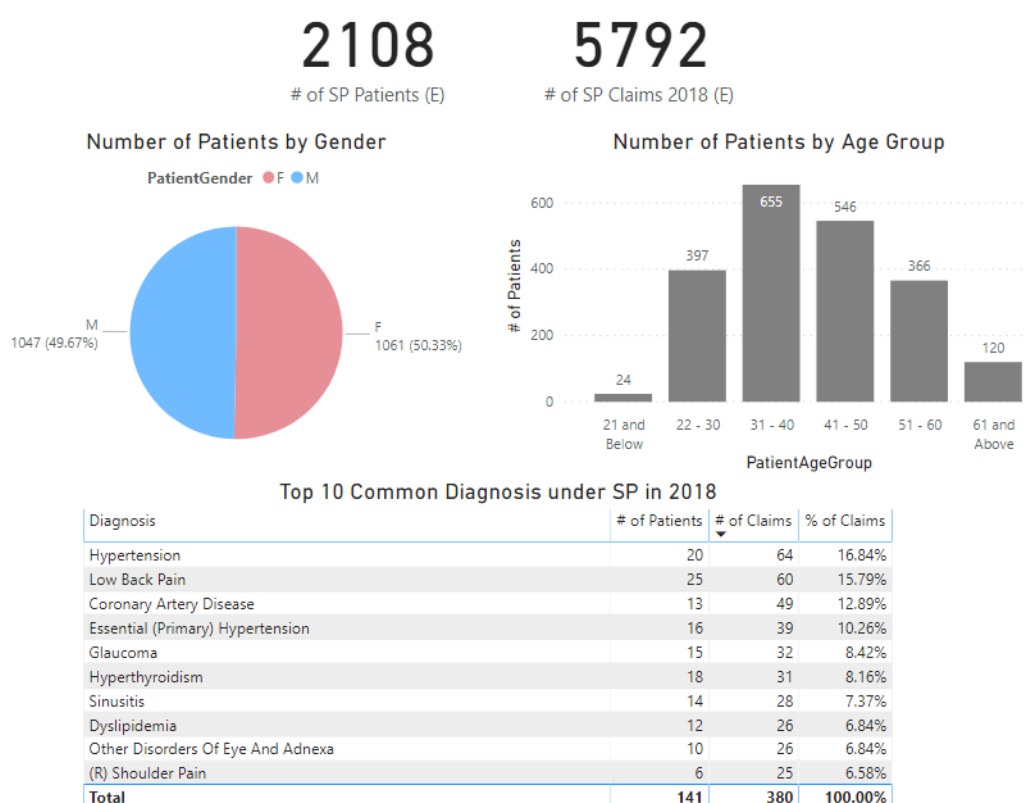


Figure 27: Demographics of Patients (Employees) SP Claims 2018

In Figure 27, 2,108 patients (employees) made a total of 5,792 claims under SP in year 2018. The following section focuses on patients who were employees only. 49.7% or 1,047 were

male employees while 50.3% or 1,061 were female employees. Most employees who made claims under SP come from the age group of 31 - 40 (655) while the 2nd highest comes from 41 - 50 (546) then followed by 22 - 30 (397) and 51 - 60 (366). Less patients were from the age groups of 61 and Above and 21 and Below at 120 and 24 patients, respectively. 141 patients made a collective total of 380 claims as shown in the top 10 common diagnoses. Employees went to specialists more commonly for the following diagnosis such as “Hypertension” (64 + 39 = 103), “Low Back Pain” (60) and “Coronary Artery Disease” (49).

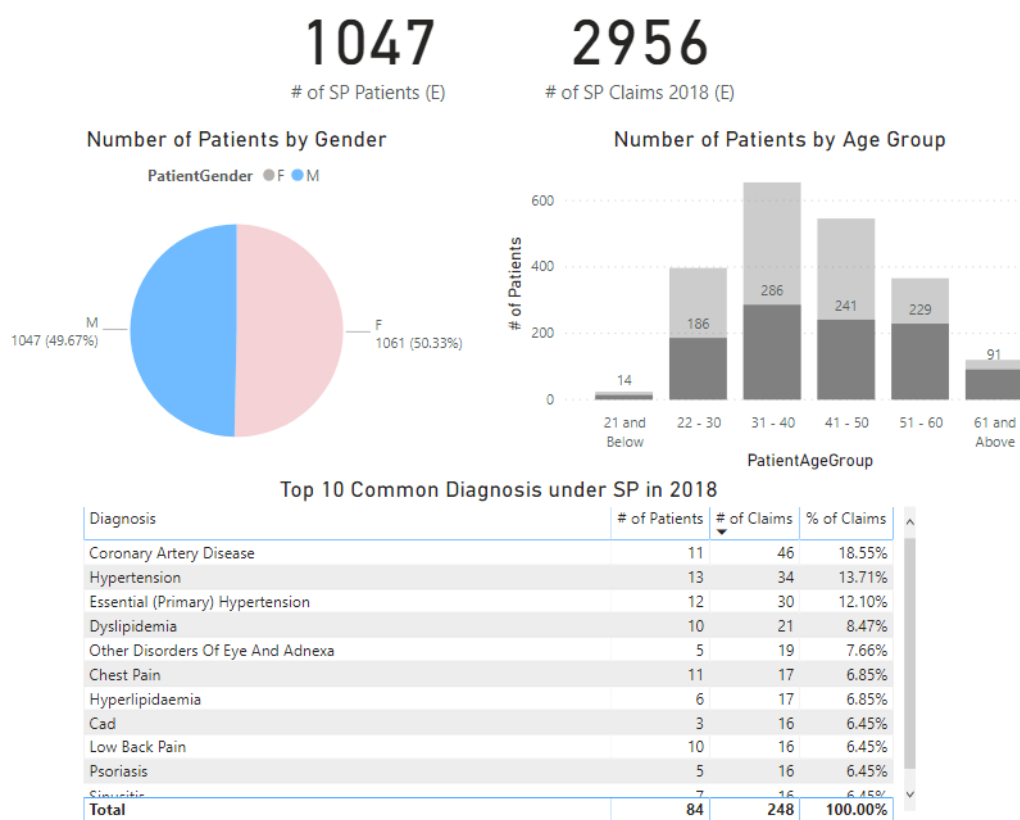


Figure 28: Demographics of Patients (Employees) SP Claims (M) 2018

Figure 28 splits the patients who are employees into gender based (M). There were 1,047 male patients who made 2,956 claims in total. Age group distribution among male patients differ slightly where the highest number of patients come from 31 - 40 (286), 41 - 50 (241) and 51 - 60 (229) in a descending order. The remaining age groups of 22 - 30, 61 and Above and 21 and Below had 186, 91 and 14 patients respectively for each age group. By filtering the top 10 diagnoses only, 248 claims were made by 84 male patients. “Hypertension” is the most

common diagnosis recorded for male patients at 64 claims by 25 patients while the 2nd most common diagnosis would be “Coronary Artery Disease” at 46 claims by 11 patients.

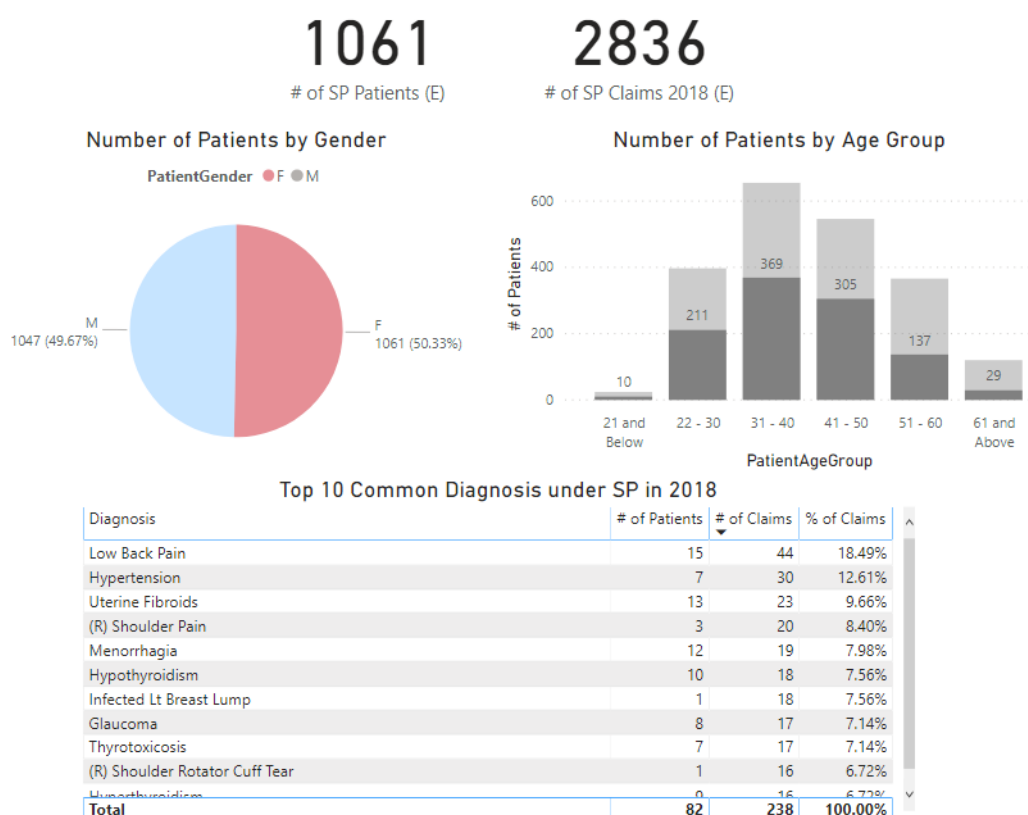


Figure 29: Demographics of Patients (Employees) SP Claims (F) 2018

Figure 29 splits the patients who are employees into gender based (F). There were 1,061 females out of the total 2,108 patients. They made 2,836 claims in total. Female age group distribution was slightly different as compared to male with the highest number of patients from the age group of 31 - 40 (369), 41 - 50 (305) and 22 - 30 (211). The remaining 176 patients comes from the age groups of 51 - 60 (137), 61 and Above (29) and 21 and Below (10). Based on the top 10 common diagnoses as recorded, 238 claims were made by 82 female patients (employees). Female employees were diagnosed with “Low Back Pain” issues more commonly at 44 claims by 15 patients while “Hypertension” was 2nd at 20 claims by 7 patients. “Uterine Fibroids” were also commonly found among female employees where 23 claims were made by 13 patients.

h) IP Overview

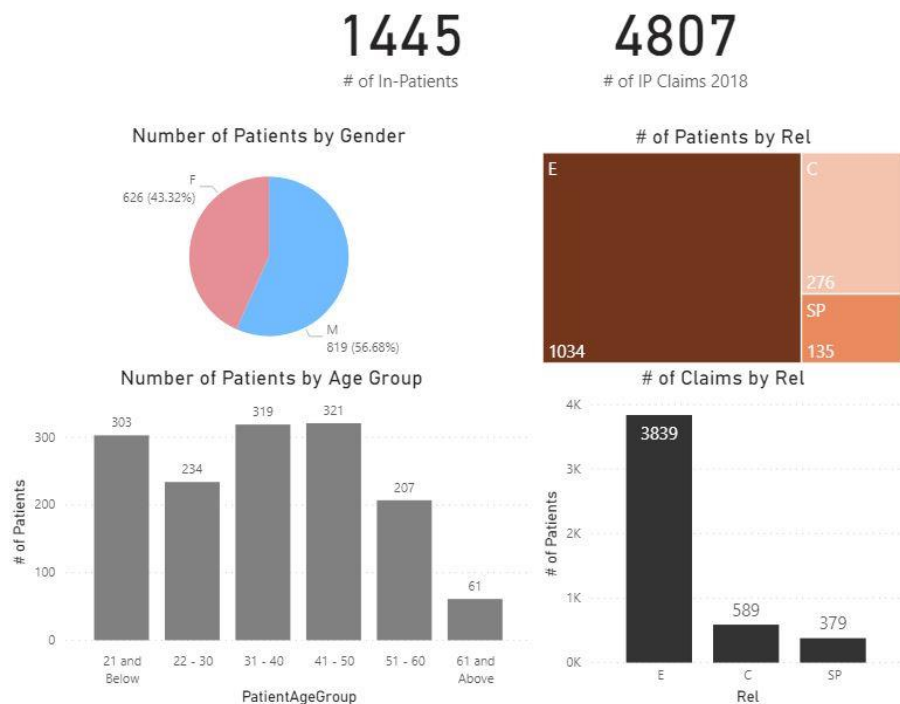


Figure 30: Basic Demographics of IP Patients 2018

Figure 30 onwards focuses on in-patient records. A total of 1,445 patients were admitted into a hospital in year 2018 while during this period and made 4,807 claims. Out of these 1,445 patients, 819 (56.7%) were males while 626 (43.3%) were females. The highest patient age group who were admitted is 41 - 50 with 321 patients, while 31 - 40 is the 2nd highest at 319 patients, then followed by 21 and Below with 303 patients. In 4th would be 22 - 30 (234) then 51 - 60 (207) and finally, 61 and Above (61). Out of the total 1,445 patients, 1,034 came from employees (E), while 135 were from spouse (SP) and 276 children (C). A large percentage of in-patient patients were employees. Looking at claims, E made a total of 3,839 claims while C made 589, SP made the least number of claims at only 379.

Diagnosis	ICDCode	# of Patients	# of Claims Frequency
Gastritis, Unspecified	K29.7	92	197
Dengue Fever (Classical Dengue)	A90	62	117
Diarrhoea And Gastroenteritis Of Presumed Infectious Origin(Acute Gastroenteritis /Dysentery)	A09	57	108
Intervertebral Disc Disorder, Unspecified	M51.9	25	107
Acute Sinusitis	J01	24	91
Other Cataract	H26	19	85
Tear Of Meniscus, Current	S83.2	19	82
Sebaceous Cyst	L72.1	10	72
Anal Fistula	K60.3	7	66
Gastro-Oesophageal Reflux Disease	K21	28	66
Unspecified Lump In Breast	N63	21	66
Acute Bronchitis	J20	30	63
Haemorrhoids	I84	25	63
Unspecified Arthropod-Borne Viral Fever	A94	31	63
Pneumonia, Unspecified	J18.9	29	60
Cellulitis	L03	19	58
Bronchopneumonia, Unspecified	J18.0	25	55
Acute Tonsillitis	J03	25	52
Essential (Primary) Hypertension	I10	17	51
Spondylosis	M47	4	50
Cutaneous Abscess, Furuncle And Carbuncle	L02	6	49
Acute Appendicitis	K35	21	48
Influenza Due To Identified Avian Influenza Virus(Influenza A/H1N1, Influenza A/H5N1)	J09	24	47
Total		1445	4807

Figure 31: Diagnoses under IP Claims 2018

Figure 31 shows the diagnoses of In-Patient patients. It shows all 4,807 claims which were made in the year 2018 under IP claims. The most common admission diagnosis would be “Gastritis” where a total of 197 claims were made with this diagnosis by 92 patients then followed by “Dengue Fever” with 117 claims made by 62 patients, in 3rd would be “Diarrhoea” at 108 claims by 57 patients and in 4th “Intervertebral Disc Disorder” at 107 claims by 25 patients. These 4 diagnoses had recorded a total claim which exceed 100.

i) IP Overview - Employee (Encounters)

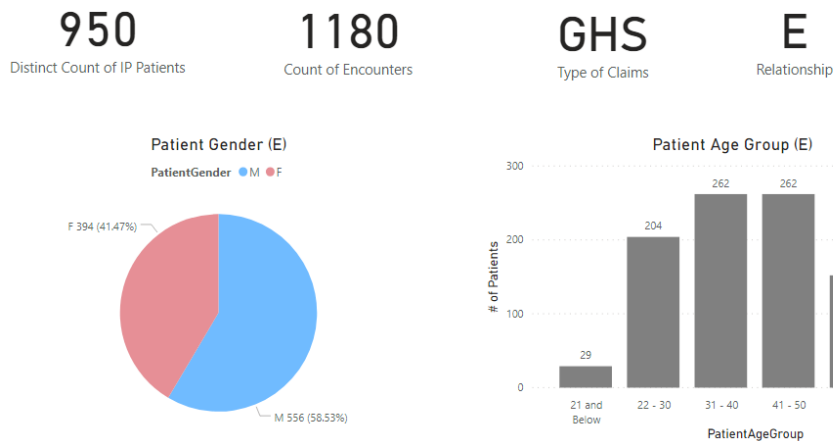


Figure 32: Demographics of Patients (Employees) IP Encounters 2018

Here, the focus turned to in-patient employees based on encounters. The difference here is reflected on the Type of Claims, where each patient visit, it is categorized based on GHS. GHS refers to general hospitalization. As shown in Figure 32, 950 patients who were employees made a total of 1,183 encounters. An encounter would refer to the admission diagnosis and not subsequent follow ups. Out of these 950 patients (employees), 556 (58%) were male employees while the remaining 394 (42%) were female employees. Patient age group distribution shows that 31 - 40 and 41 - 50 had the highest number of patients at 262 each while 22 - 30 had 204 patients. Remaining 3 age groups of 51 - 60 (152), 61 and Above (41) and 21 and Below (29) had less than 200 patients.

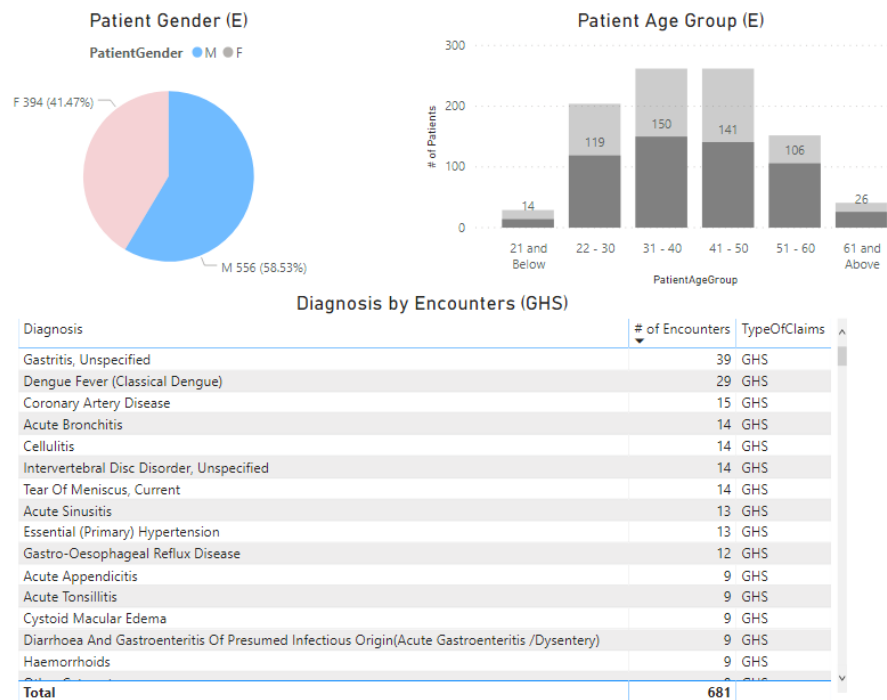


Figure 33: Diagnoses of IP Encounters (M) 2018

Figure 33 shows the recorded diagnoses of male patients based on each encounter. There were 556 male patients who had 681 encounters in total. Out of the 681 encounters, 39 of them represent the diagnosis of “Gastritis” while 29 cases were “Dengue Fever”. In 3rd would be “Coronary Artery Disease” where there were 15 cases.

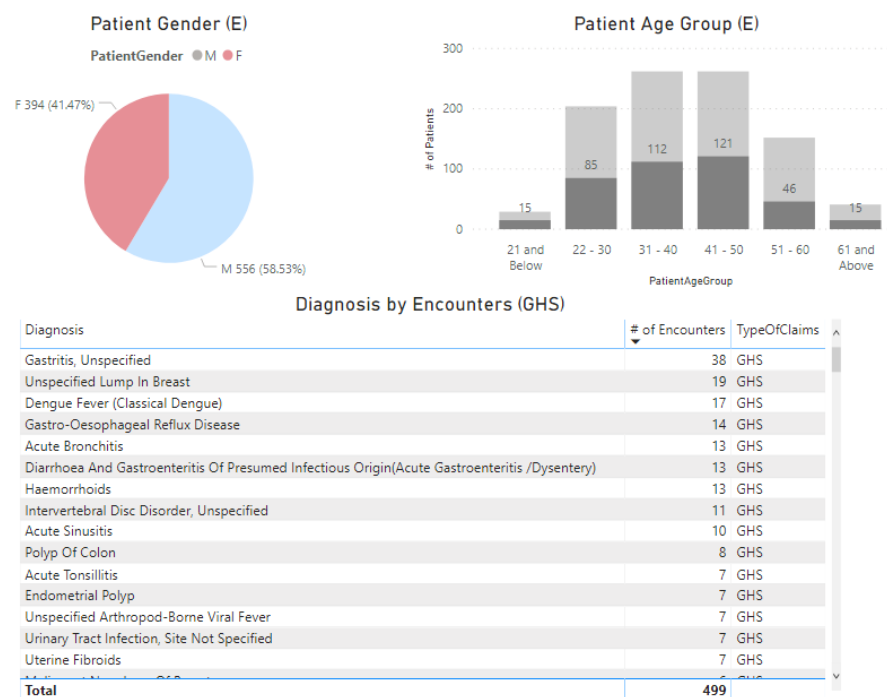


Figure 34: Diagnoses of IP Encounters (F) 2018

Figure 34 shows the recorded diagnoses of female patients based on each encounter. As mentioned previously, there were 394 female patients who had 499 encounters in total. Out of the 499 encounters, “Gastritis” has the highest number of encounters at 38 while 2nd is “Unspecified Lump in Breast” at 19 encounters, and in 3rd it is “Dengue Fever” where there were 17 cases.

j) GP Overview - Drill Down Analysis by Age Group

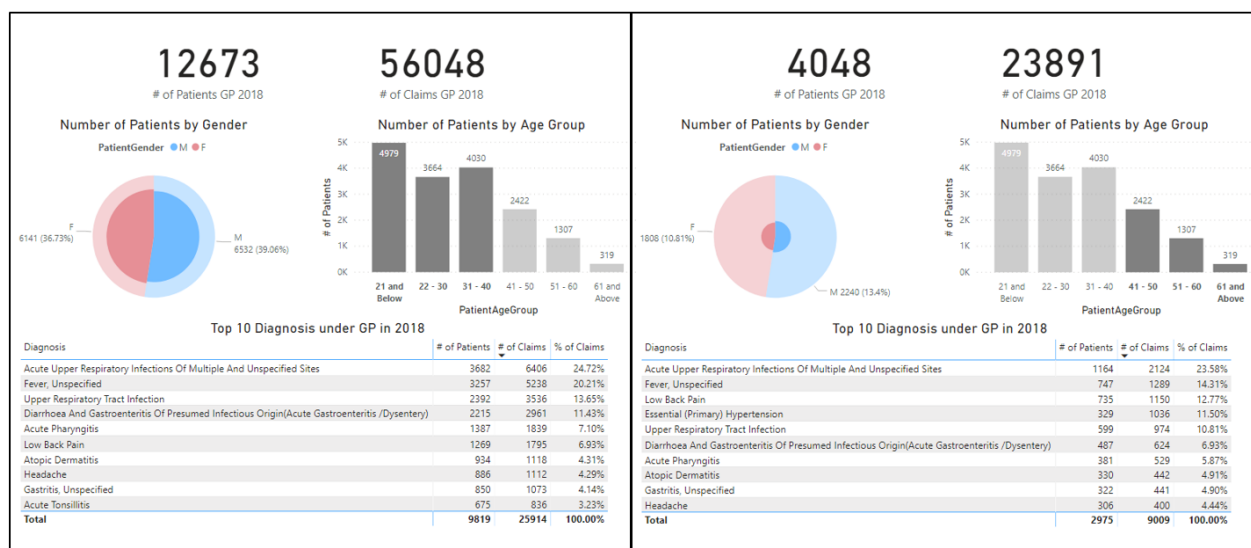


Figure 35: Comparison of Age Group between (Below 40) and (Above 40) - GP

Next, a comparison of GP claims based on age group of Below 40 and Above 40. Below 40 includes 21 and Below, 22 - 30 and 31 - 40 while Above 40 includes 41 - 50, 51 - 60 and 61 and Above. As shown above, there are more patients within the Below 40 segment at 12,673 as compared to the Above 40 segment which only had 4,048 patients. Furthermore, below 40 segments made approximately 50% more claims at 56,048 as compared to Above 40 segment who made 23,891 claims. At the bottom of the comparison shows top 10 diagnoses under GP in 2018, the focus will be on the top 3 diagnosis, the 1st two diagnosis are similar with “Acute Upper Respiratory Infections” and “Fever” for both, however the 3rd common diagnosis showed “Upper Respiratory Tract Infection” for Below 40 and “Low Back Pain” for Above 40.

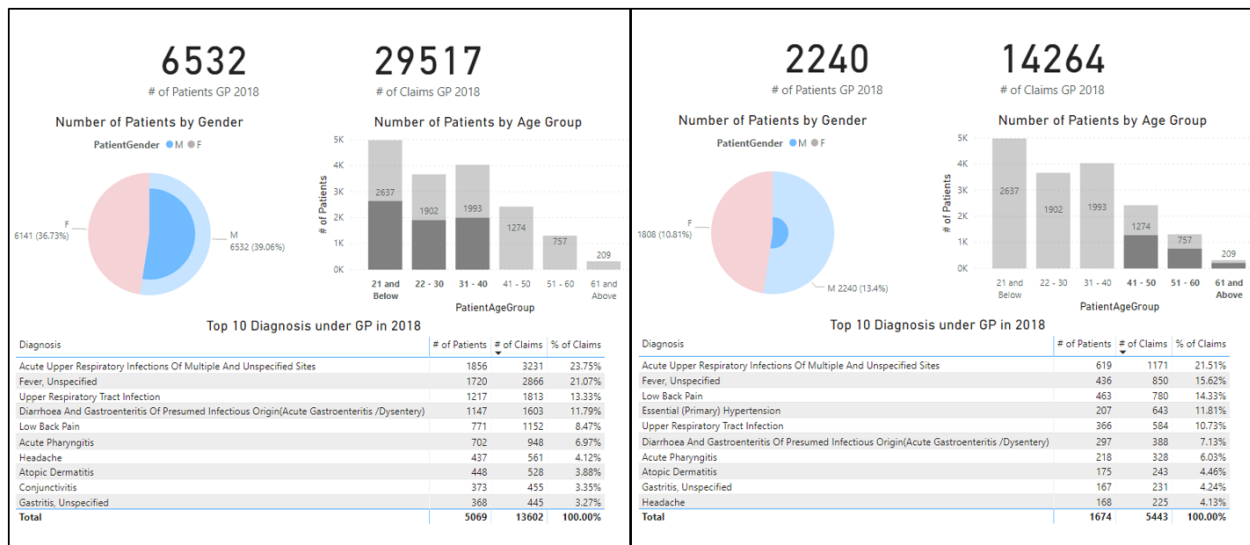


Figure 36: Comparison of (M) Age Group between (Below 40) and (Above 40) - GP

Figure 36 focuses on the comparison of males based on age segment of Below 40 and Above 40 who had made claims under GP in 2018. As shown, below 40 there were 6,532 male patients who made 29,517 claims while Above 40 there were 2,240 male patients who made 14,264 claims. Similarly, the Below 40 segment made almost 50% more claims than Above 40. Moving on, the diagnoses are identical to the description in Figure 36 where “Acute Upper Respiratory Infections” and “Fever” are the 1st two for both segments, however the 3rd common diagnosis showed “Upper Respiratory Tract Infection” for Below 40 and “Low Back Pain” for Above 40.

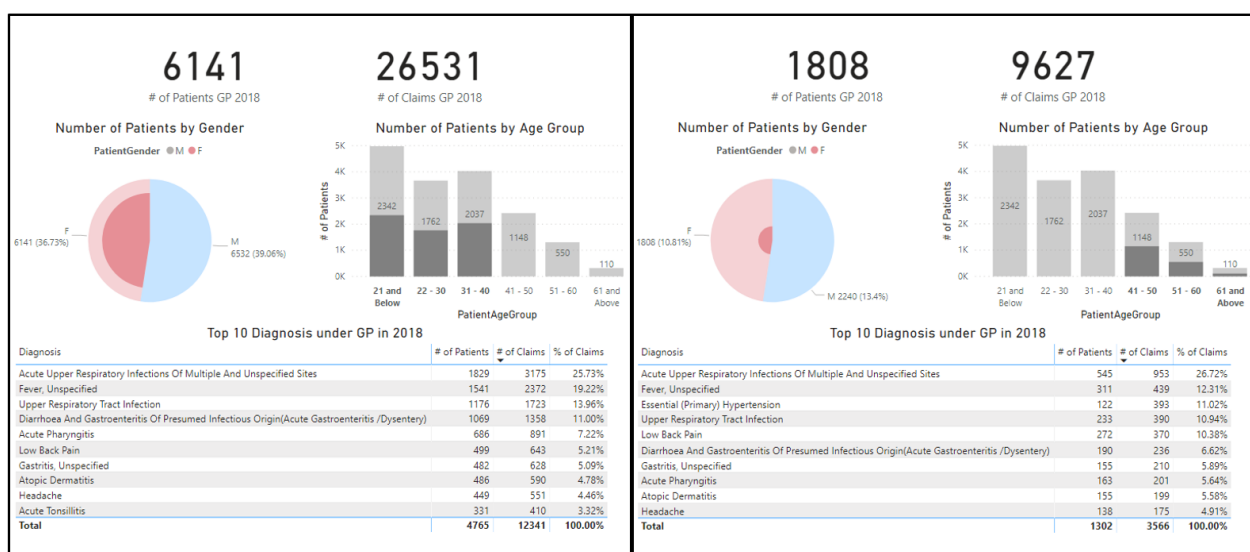


Figure 37: Comparison of (F) Age Group between (Below 40) and (Above 40) - GP

Figure 37 focuses on the comparison of females based on age segment of Below 40 and Above 40 who had made claims under GP in 2018. Below 40 segments had 6,141 female patients while Above 40 only had 1,808 female patients. Out of the total 36, 158 claims made by females, 26,531 came from Below 40 while Above 40 only made 26.6% which is 9,627 claims. Looking at the diagnoses, “Acute Upper Respiratory Infections” and “Fever” are the 1st two for both segments, however the 3rd common diagnosis showed “Upper Respiratory Tract Infection” for Below 40 and “Hypertension” for Above 40.

k) GP Overview - Drill Down Analysis by Age Group (Employee)

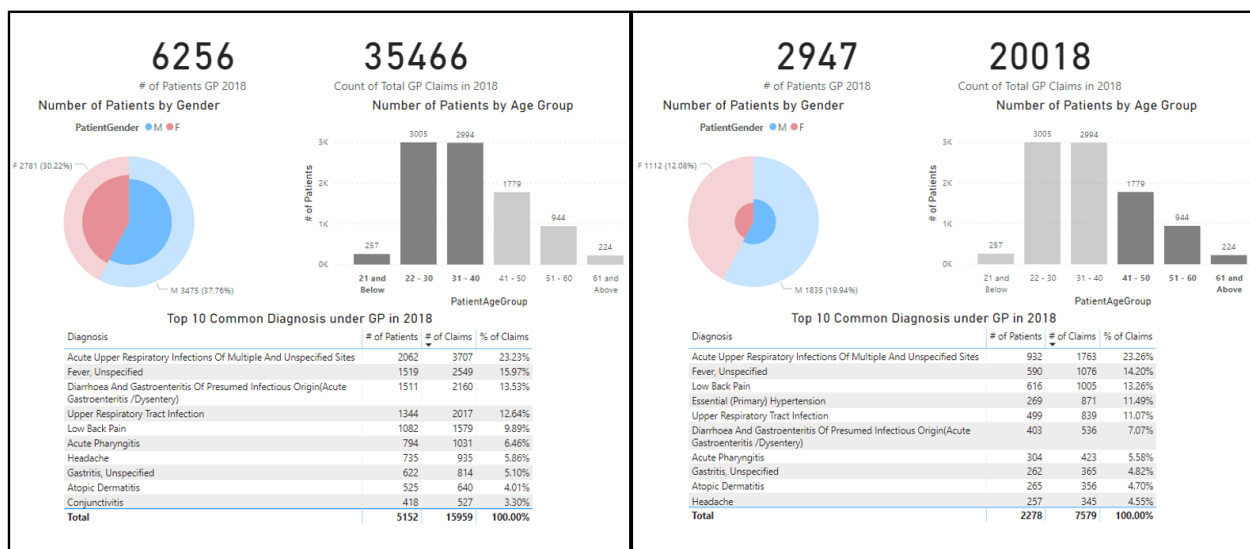


Figure 38: Comparison of Age Group between (Below 40) and (Above 40) - Employee; GP

Moving on, the comparison of GP claims based on age group of Below 40 and Above 40 changed to only include employees (E). As shown in Figure 38, Below 40 segments have 6,256 patients while Above 40 have 2,947 patients. They made a total of 55,484 claims, but 35,466 of those claims were made by Below 40 segments while Above 40 made 20,018 claims. Moving down, the 1st two most common diagnosis were similar for both segments with “Acute Upper Respiratory Infections” and “Fever” for both, however the 3rd most common diagnosis showed “Diarrhoea” for Below 40 and “Low Back Pain” for Above 40.

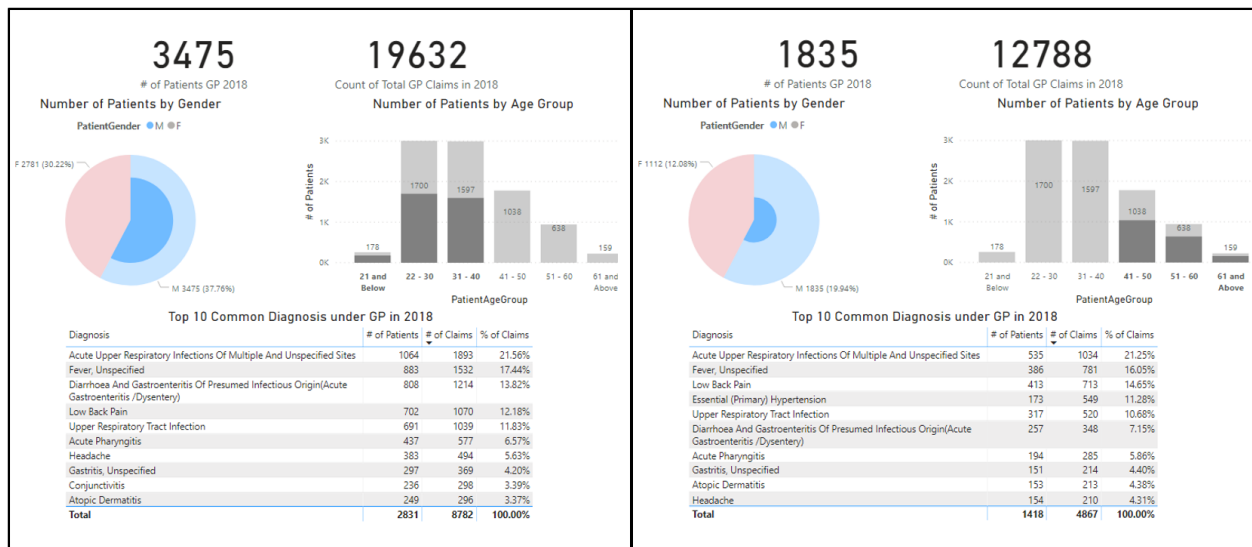


Figure 39: Comparison of (M) Age Group between (Below 40) and (Above 40) - Employee; GP

Figure 39 shows the comparison of male patients (employees) based on the two segments. Below 40 had 3,475 patients who made a total of 19,632 claims while Above 40 had 1,835 patients who made a total of 12,788 claims. The 1st two most common diagnosis were similar for both segments with “Acute Upper Respiratory Infections” and “Fever” for both, however the 3rd most common diagnosis showed “Diarrhoea” for Below 40 and “Low Back Pain” for Above 40 which is identical to Figure 38.

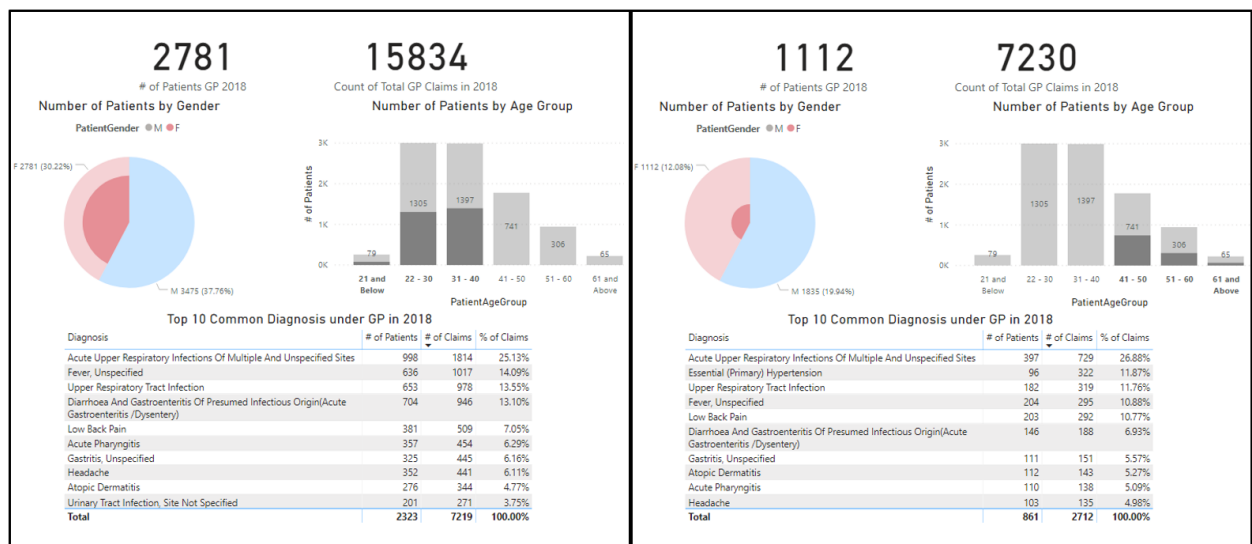


Figure 40: Comparison of (F) Age Group between (Below 40) and (Above 40) - Employee; GP

Figure 40 shows the comparison of female patients (employees) based on the two segments. There were 2,781 patients who were Below 40 while Above 40 had 1,112 patients. Below 40

segment made a total of 15,834 claims, which is 50% more than Above 40 at 7,230 claims. Looking at the diagnoses section, the most common and 3rd most common diagnosis were similar for both segments with “Acute Upper Respiratory Infections” and “Upper Respiratory Tract Infection. The 2nd most common diagnosis however differed whereby Below 40 was “Fever” and Above 40 was “Hypertension”.

l) GP Overview - Clustering

If one is trying to learn something, say music, an approach could be to search for interesting groups or collections (Google Developers, 2020). One may arrange music by genre, while others may arrange music by decade. How one decides to arrange objects may help one understand more about them as individual pieces of music (Google Developers, 2020). In machine learning, to better understand and identify patterns within a dataset can be achieved through an unsupervised learning approach known as Clustering (Google Developers, 2020). Clustering is a profile segmenting approach where similar characteristics and behaviour will be grouped together in clusters.

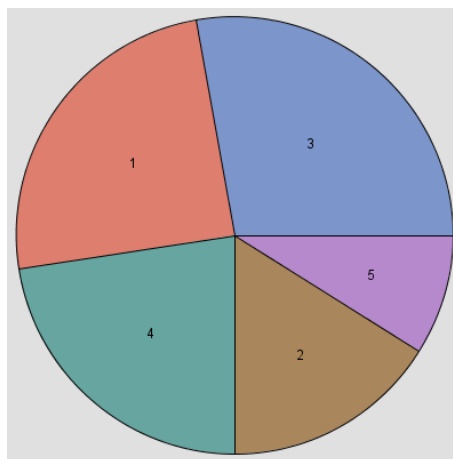


Figure 41: Segment Size of Clustering

The pie chart above in Figure 41 shows the 5 clusters which have been segmented based on the input dataset. Segment 3 was the largest segment with a total number of observations of 44,665 which is around 27.7% of the entire dataset, followed by Segment 1 with 39,982 observations

which occupy 24.8% of the dataset, 4 with 36,312 observations occupying 22.52% of the dataset, then 2 and 5 occupying 16% and 9% respectively.



Figure 42: Profile Segment of Clustering

Figure 42 shows an overview of segments generated through SAS Enterprise Miner based on the input dataset. As mentioned previously, 5 segments were generated. Out of the 5 segments, segment 3 was the largest segment with a total of 44,665 observations. Therefore, only segment 3 would be further analysed as most patient recordings would come from this segment and the purpose is to better understand the relation between the chosen variables.

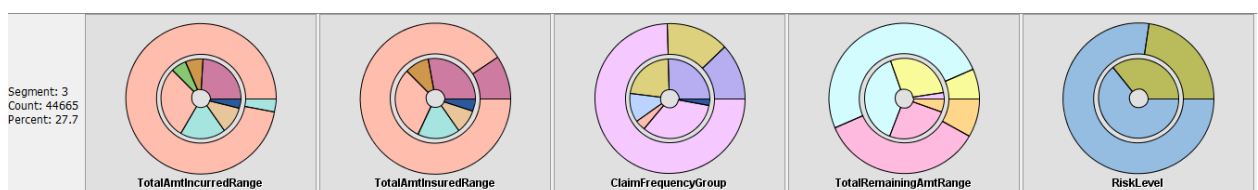


Figure 43: Segment 3 - Largest Segment

Figure 43 shows the profile overview of patients within segment 3. As mentioned previously, there were a total of 44,665 observations which represents approximately 27.7% of the total dataset. 5 variables were considered under this segment including TotalAmtIncurredRange, TotalAmtInsuredRange, ClaimFrequencyGroup, TotalRemainingAmtRange and RiskLevel. Looking closely at the circles in each variable, an inner circle represents all observations while

the outer circle shows the observations within segment 3 only. Segment 3 variable “TotalAmtIncurredRange”, indicates that 97.3% of the observations were incurred a total range of 301 - 600, while a small percentage of 2.7% were within the incurred range of 601 - 900. Notice at the inner circle there are 7 parts but only 2 parts were included within segment 3. This shows that none of the observations incurred less than 300 or more than 900. “TotalAmtInsuredRange”, indicates that 92.3% of the observations were insured a total range of 301 - 600, while a small percentage of 7.7% were within the incurred range of 601 - 900. “ClaimFrequencyGroup”, indicates that 11.8% of the observations claimed within the range of 1 - 5, while 74.2% made claims within the range of 6 - 10, and 13.8% of the claims were within the range of 11 – 15. Looking at the inner circle, there were 6 parts while segment 3 only includes 3 parts - which translates to 44,665 of these observations were within this ranges in terms of claim frequency. This shows that none of the observations had more than 15 claims. “TotalRemainingAmtRange”, indicates that 6.3% of the 44,665 observations had a total remaining amount range of 1 - 1000, while 49.4% had a remaining amount range of 1001 - 2000, 35.1% had 2001 – 3000 remaining amount range, and 8.23% of the observations had 3001 – 4000 remaining amount range. This indicated that the only range that was not included would be 0 and Below, which means segment 3 do not have any observation who has less than 0 remaining amount. “RiskLevel”, is the last variable within this segment. This indicates that 77.2% of the observations had a RiskLevel indicator of L (Low), while a small percentage of 22.8% had a RiskLevel indicator of H (High). An overview of segment 3 would indicate that observations within this segment has similar characteristics as mentioned above.

4.2 Predictive Analysis

4.2.1 Model Overview

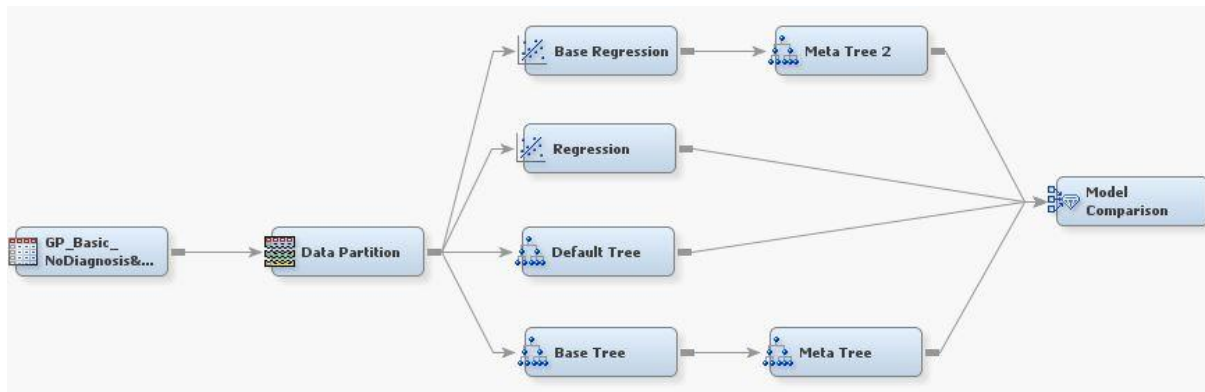


Figure 44: Predictive Model Framework

Figure 44 shows the predictive model framework which has been constructed using SAS Enterprise Miner. Starting from the left, the first node shown was GP_Basic, the dataset “GP” was used to construct the predictive model. The second node was Data Partition, the node was incorporated to split the dataset into two distinct parts: Training (70) and Validation (30) - training uses a portion of the dataset containing the target and input variables to build and train the predictive models constructed while validation performs a validation on the predictive accuracy and suitability of the constructed model. After Data Partition, a total of 4 predictive models were constructed which included: 2 single predictive models (Default Tree and Regression) and 2 stacking ensemble models (Base Regression + Meta Tree and Base Tree + Meta Tree). As mentioned in previous sections, Decision Tree models were preferred over others due to a few reasons such as the Target value being a binary/nominal value and also the interpretability of the predictive results, decision tree is indeed the most easily understandable predictive model as it uses a logical tree and if-else rule approach to describe the prediction. The aim of this research is to mitigate the problem where only experts in this field of interest could understand, this is to provide an alternative approach for professionals to easily understand and interpret a predictive model as well. The last node would be Model Comparison, which compares and chooses the best performing based on accuracy.

4.2.2 Single Model (Decision Tree and Regression)

Decision Tree

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
RiskLevel		NOBS	Sum of Frequencies	112866	48374	
RiskLevel		MISC	Misclassification ...	0.127975	0.128416	
RiskLevel		MAX	Maximum Absolut...	0.999815	1	
RiskLevel		SSE	Sum of Squared E...	23392.93	10073.25	
RiskLevel		ASE	Average Squared ...	0.103631	0.104118	
RiskLevel		RASE	Root Average Squ...	0.321918	0.322674	
RiskLevel		DIV	Divisor for ASE	225732	96748	
RiskLevel		DFT	Total Degrees of ...	112866		

Figure 45: Fit Statistics - Default Tree

Figure 45 shows the Sum of Frequencies (NOBS) where there were 112,866 observations used under Train while 48,374 observations were used for Validation. The Misclassification (MISC) rate recorded were 0.1280 for Train and 0.1284 for Validation. While Averaged Squared Error (ASE) showed that for Train and Validation both recorded values of 0.1036 and 0.1041 respectively which means for both MISC and ASE across the two partitions of Train and Validation were not significantly different.

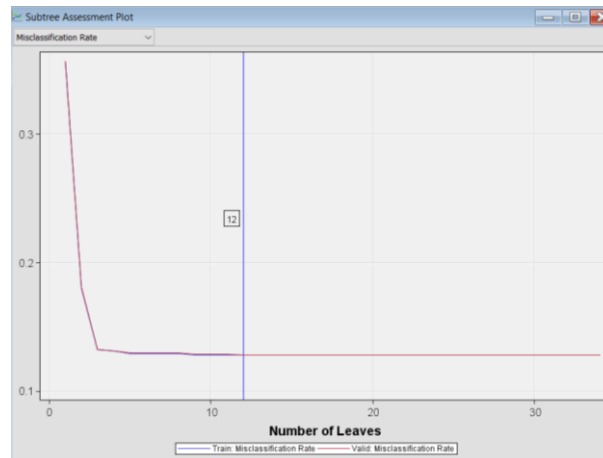
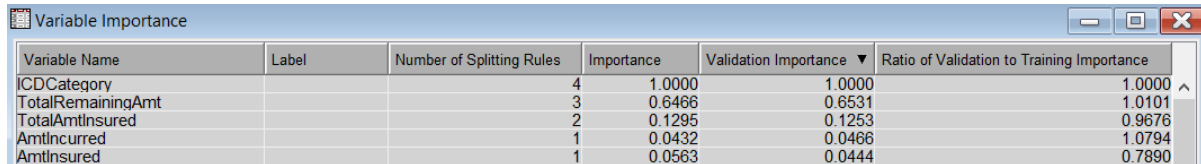


Figure 46: Subtree Assessment Plot - Default Tree

Looking at the Subtree Assessment Plot of the Default Tree, on the left are labels showing the MISC rate against each subtree as the data splits accordingly. There are two lines, blue representing the Train – Misclassification while the Red represents - Validation Misclassification. Both Train and Validation started at Leaf 1 where the MISC was 0.3574 then it dropped steeply until Leaf 3. Based on the Subtree Assessment Plot, the performance of both

Train and Validation had minimal differences. After Leaf 3, Leaf 4 onwards there was a consistent performance without much fluctuation. The tree stopped growing at Leaf 12, which is showed on the plot. This would mean that the total number of leaves for the default tree would be 12 - this shows the concept of under- and over-fitting where any lesser leaves might result in under-fitting while any more leaves might increase complexity leading to over-fitting.



Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance ▼	Ratio of Validation to Training Importance
ICDCategory		4	1.0000	1.0000	1.0000 ^
TotalRemainingAmt		3	0.6466	0.6531	1.0101
TotalAmtInsured		2	0.1295	0.1253	0.9676
AmtIncurred		1	0.0432	0.0466	1.0794
AmtInsured		1	0.0563	0.0444	0.7890

Figure 47: Variable Importance - Default Tree

Figure 47 shows the importance of each predictor towards the Default Tree. As followed the variable importance has been arranged in a descending order from the highest importance to the lowest, the most important being ICDCategory, then followed by TotalRemainingAmt, TotalAmtInsured, AmtIncurred and AmtInsured. These 5 predictors selected have the largest influence towards the construction of the Default Tree model.

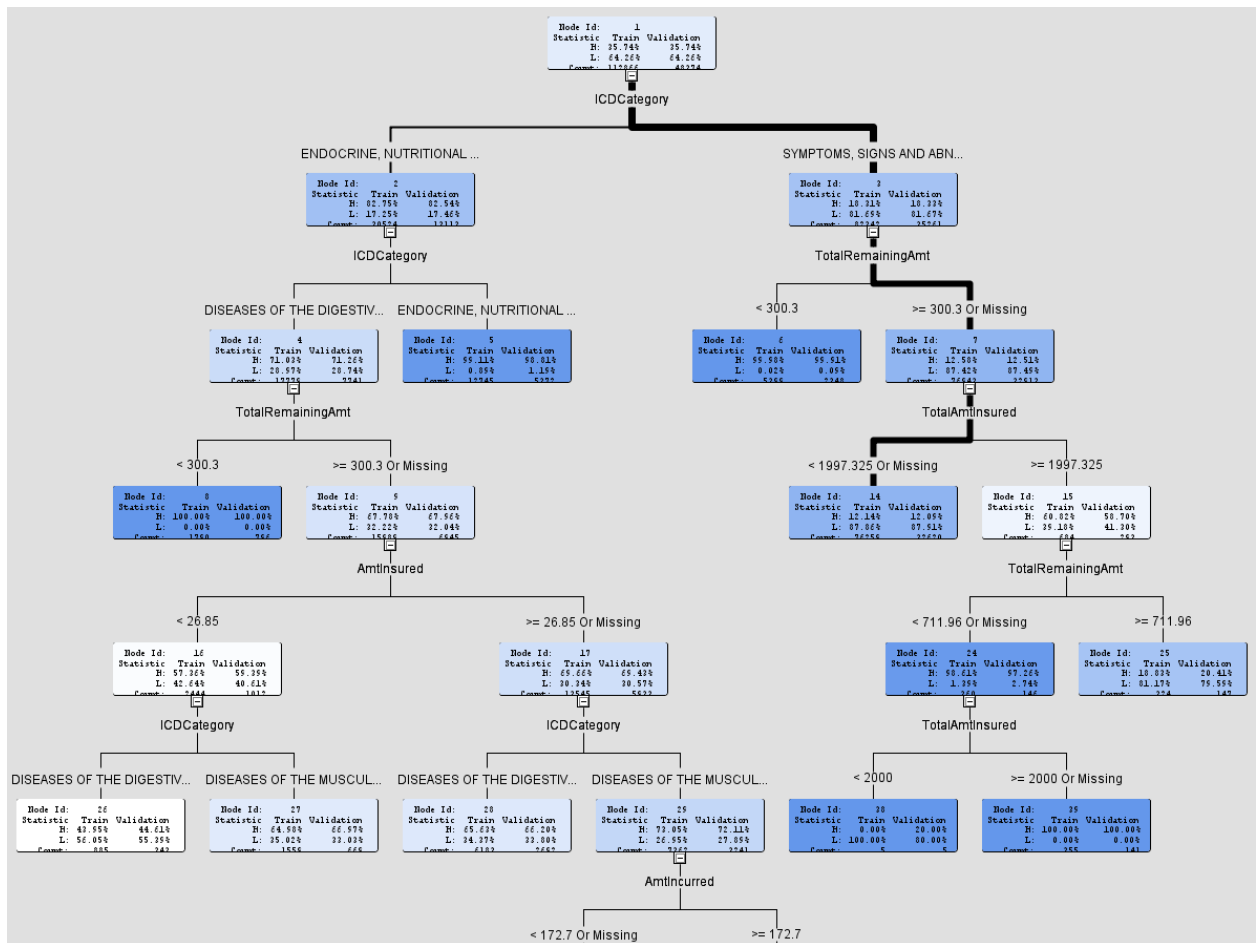


Figure 48: Default Tree - Tree Overview

Moving on, Figure 48 shows the tree overview of the Default Tree constructed by SAS Enterprise Miner. It has a depth of 6 and a total of 12 leaves. There were 2 splits each time as the setting was left on default with 2 branches - generally, in most cases, a decision tree model would have 2 branches only to reduce the complexity of the model. Looking at the tree overview, the chosen path would be represented by the thicker line width where it signifies a higher volume of observations going to those nodes. As shown, the chosen path starts from Node 1 and ends at Node 14 with a depth of 3. The variables which had the largest influence in this path includes the Target, Risk Level, ICD Category, TotalRemainingAmt and TotalAmtInsured. The next section would further explain the chosen path.

```

*-----*
Node = 5
*-----*
if ICDCategory IS ONE OF: ENDOCRINE, NUTRITIONAL AND METAB, DISEASES OF THE
CIRCULATORY SYST, DISEASES OF THE NERVOUS SYSTEM
then
Tree Node Identifier = 5
Number of Observations = 12745
Predicted: RiskLevel=L = 0.01
Predicted: RiskLevel=H = 0.99

```

Figure 49: Node Rules - Node 5

Node 5 shows the highest percentage of high-risk patients. There were a total of 12,745 observations in this node with 99% of them as high risk, these patients would fulfil the criteria as stated above in Figure 49 for the variable of ICDCategory.

```

*-----*
Node = 6
*-----*
if TotalRemainingAmt < 300.3
AND ICDCategory IS ONE OF: SYMPTOMS, SIGNS AND ABNORMAL CLI, INJURY, POISONING AND
OTHER CONS, DISEASES OF THE RESPIRATORY SYST, DISEASES OF THE SKIN AND SUBCUTA,
DISEASES OF THE EAR AND MASTOID, DISE
then
Tree Node Identifier = 6
Number of Observations = 5399
Predicted: RiskLevel=L = 0.00
Predicted: RiskLevel=H = 1.00

```

Figure 50: Node Rules - Node 6

Node 6 consists of 100% high risk patients. However, only 5,399 observations were within this node which is 57.6% less as compared to Node 5. These patients within Node 6 would fulfil the above criteria as shown in Figure 50 for the variables of TotalRemainingAmt and ICDCategory.

Node = 14

if TotalRemainingAmt >= 300.3 or MISSING AND TotalAmtInsured < 1997.33 or MISSING AND ICDCategory IS ONE OF: SYMPTOMS, SIGNS AND ABNORMAL CLI, INJURY, POISONING AND OTHER CONS, DISEASES OF THE RESPIRATORY SYST, DISEASES OF THE SKIN AND SUBCUTA, DISEASES OF THE EAR AND MASTOID, DISE
then
Tree Node Identifier = 14
Number of Observations = 76259
Predicted: RiskLevel=L = 0.88
Predicted: RiskLevel=H = 0.12

Figure 51: Node Rules - Node 14

Node 14 has the highest percentage of low-risk patients. There were 76,259 observations within this node with 88% of them as low risk patients. These patients within Node 14 would fulfil the above criteria as shown in Figure 51 for the variables of TotalRemainingAmt, TotalAmtInsured and ICDCategory.

Regression

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
RiskLevel		AIC	Akaike's Informati...	80050.43		
RiskLevel		ASE	Average Squared ...	0.107096	0.108191	
RiskLevel		AVERR	Average Error Fun...	0.351941	0.356182	
RiskLevel		DFE	Degrees of Freed...	112563		
RiskLevel		DFM	Model Degrees of ...	303		
RiskLevel		DFT	Total Degrees of ...	112866		
RiskLevel		DIV	Divisor for ASE	225732	96748	
RiskLevel		ERR	Error Function	79444.43	34459.92	
RiskLevel		FPE	Final Prediction Er...	0.107672		
RiskLevel		MAX	Maximum Absolut...	0.999434	0.999546	
RiskLevel		MSE	Mean Square Error	0.107384	0.108191	
RiskLevel		NOBS	Sum of Frequencies	112866	48374	
RiskLevel		NW	Number of Estima...	303		
RiskLevel		RASE	Root Average Su...	0.327255	0.328924	
RiskLevel		RFPE	Root Final Predict...	0.328134		
RiskLevel		RMSE	Root Mean Squar...	0.327695	0.328924	
RiskLevel		SBC	Schwarz's Bayesi...	82969.52		
RiskLevel		SSE	Sum of Squared E...	24174.91	10467.28	
RiskLevel		SUMW	Sum of Case Wei...	225732	96748	
RiskLevel		MISC	Misclassification ...	0.137836	0.139682	

Figure 52: Fit Statistics - Regression

Next, the second predictive model which was constructed would be a Regression model. 112,866 observations were used to Train the model and 48,374 were used to Validate the model. The MISC rate recorded were 0.1379 for Train and 0.1397 for Validation. While ASE showed that Train and Validation both recorded values of 0.1071 and 0.1082 respectively which means both MISC and ASE across the two partitions were not significantly different.

Type 3 Analysis of Effects			
Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
AmtIncurred	1	5.6762	0.0172
AmtIncurredRange	11	587.6724	<.0001
AmtInsured	1	6.3837	0.0115
AmtInsuredRange	7	1430.1083	<.0001
BranchName	138	1239.9278	<.0001
BusinessIndustry	39	212.3128	<.0001
ClaimCategory	0	0.0000	.
ClaimFrequency	1	4.9817	0.0256
ClaimFrequencyGroup	6	76.2176	<.0001
Corporate	2	251.9145	<.0001
DeptName	41	33.8556	0.7779
EmpAnnualLimit_RM_	1	145.8007	<.0001
ICDCategory	21	19735.1668	<.0001
MCDays	1	18.2302	<.0001
PatientAge	1	1.8762	0.1708
PatientAgeGroup	5	10.6065	0.0598
PatientGender	1	0.5674	0.4513
TotalAmtIncurred	1	15.8192	<.0001
TotalAmtIncurredRange	8	230.6868	<.0001
TotalAmtInsured	1	37.1110	<.0001
TotalAmtInsuredRange	8	1528.2467	<.0001
TotalRemainingAmt	0	0.0000	.
TotalRemainingAmtRange	5	485.4810	<.0001

Figure 53: Type 3 Analysis of Effects - Regression

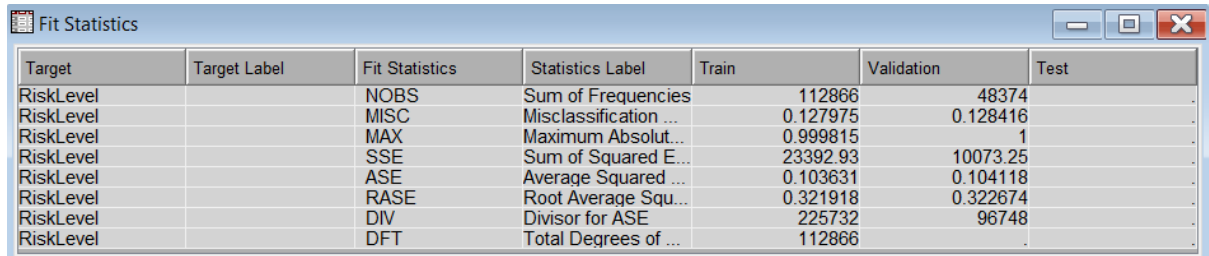
Figure 53 shows the Type 3 Analysis of Effects where it is a test to identify which predictor variables are significant, each will be considered to enter into the model last, to identify the predictive value it carries. Here, as observed there are a total of 23 variables which have differing p-value, the criteria to observe would be that p-value has to be < 0.05 to be considered as significant. Therefore, looking at the list of predictors there are a handful of which are significant including (in no particular order):

AmtIncurred, AmtIncurredRange, AmtInsured, AmtInsuredRange, BranchName, BusinessIndustry, ClaimFrequency, ClaimFrequencyGroup, Corporate, EmpAnnualLimit, ICDCategory, MCDays, PatientAgeGroup, TotalAmtIncurred, TotalAmtIncurredRange, TotalAmtInsured, TotalAmtInsuredRange and TotalRemainingAmtRange. These were the 18 predictors which had p-value of < 0.05 which would suggest that they have an influence in the prediction of “RiskLevel”. As shown in the results above, regression model is not preferable for performing classification predictions as the model is too complex and there are too many variables which were selected, hence, adding to the complex mathematical model. Moreover,

the aim of this research focuses on building a practical ensemble framework for classification predictions - regression models do not increase interpretability but increases complexity.

4.2.3 Ensemble Model (Base Tree + Meta Tree)

Base Tree + Meta Tree



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
RiskLevel		NOBS	Sum of Frequencies	112866	48374	
RiskLevel		MISC	Misclassification ...	0.127975	0.128416	
RiskLevel		MAX	Maximum Absolut...	0.999815	1	
RiskLevel		SSE	Sum of Squared E...	23392.93	10073.25	
RiskLevel		ASE	Average Squared ...	0.103631	0.104118	
RiskLevel		RASE	Root Average Squ...	0.321918	0.322674	
RiskLevel		DIV	Divisor for ASE	225732	96748	
RiskLevel		DFT	Total Degrees of ...	112866		

Figure 54: Fit Statistics - Stacking Ensemble Model (Base Tree + Meta Tree)

Figure 54 shows the predictive results of an Ensemble Model which consists of a base model of Decision Tree and a meta model of Decision Tree. There were 112,866 observations used under Train while 48,374 observations were used for Validation. The MISC rate for Train model was 0.1280 and for Validation model was 0.1284. ASE on the other hand, recorded values of 0.1036 and 0.1041 for Train and Validation respectively - this translates to minimal significance in the variance between MISC and ASE across the two partitions. An observation made was that the other Stacking Ensemble Model of a base model of Regression and a meta model of a Decision Tree yield similar results as the following Ensemble Model.

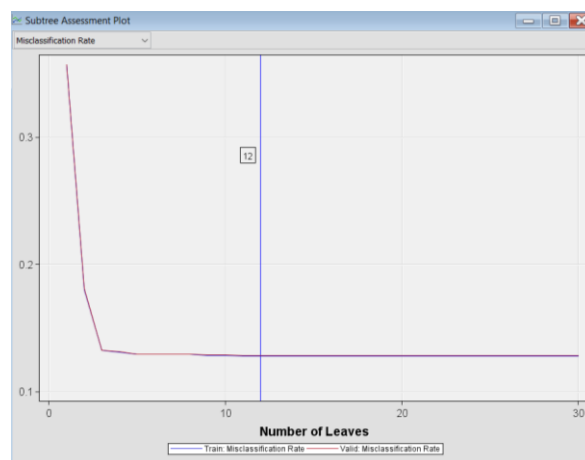
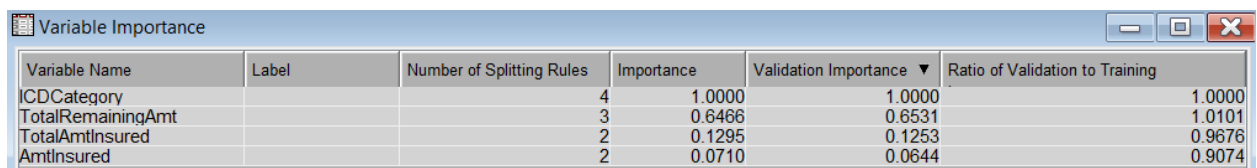


Figure 55: Subtree Assessment Plot - Stacking Ensemble Model (Base Tree + Meta Tree)

Next, the Subtree Assessment plot of the Stacking Ensemble Model of Base Tree + Meta Tree showed similar results with reference to the Default Tree. Similarly, there were 2 lines with reference to the MISC rate, 1 (blue) representing Train - while 1 (red) representing Validation. Both Train and Validation started at Leaf 1 with MISC of 0.3574 then it dropped steeply until Lead 3, subsequently Leaf 4 onwards it was consistent without fluctuations. The Ensemble Tree stopped growing at Leaf 12 as it was at the optimum performance.



Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training
ICDCategory		4	1.0000	1.0000	1.0000
TotalRemainingAmt		3	0.6466	0.6531	1.0101
TotalAmtInsured		2	0.1295	0.1253	0.9676
AmtInsured		2	0.0710	0.0644	0.9074

Figure 56: Variable Importance - Stacking Ensemble Model (Base Tree + Meta Tree)

Figure 56 shows the variable importance, which refers to the influence of each predictor towards the predictive model. In this case, the importance was arranged in a descending order where the highest importance was the first being, ICDCategory, then followed by TotalRemainingAmt, TotalAmtInsured and AmtInsured. These 4 predictors have the largest influence towards the construction of the Ensemble Tree model. This is where the difference between the Default Tree and Ensemble Tree become apparent. In Figure 47 where it showed the variable importance of the Default Tree, there were 5 predictors as compared to the Ensemble Tree where there were only 4 predictors.

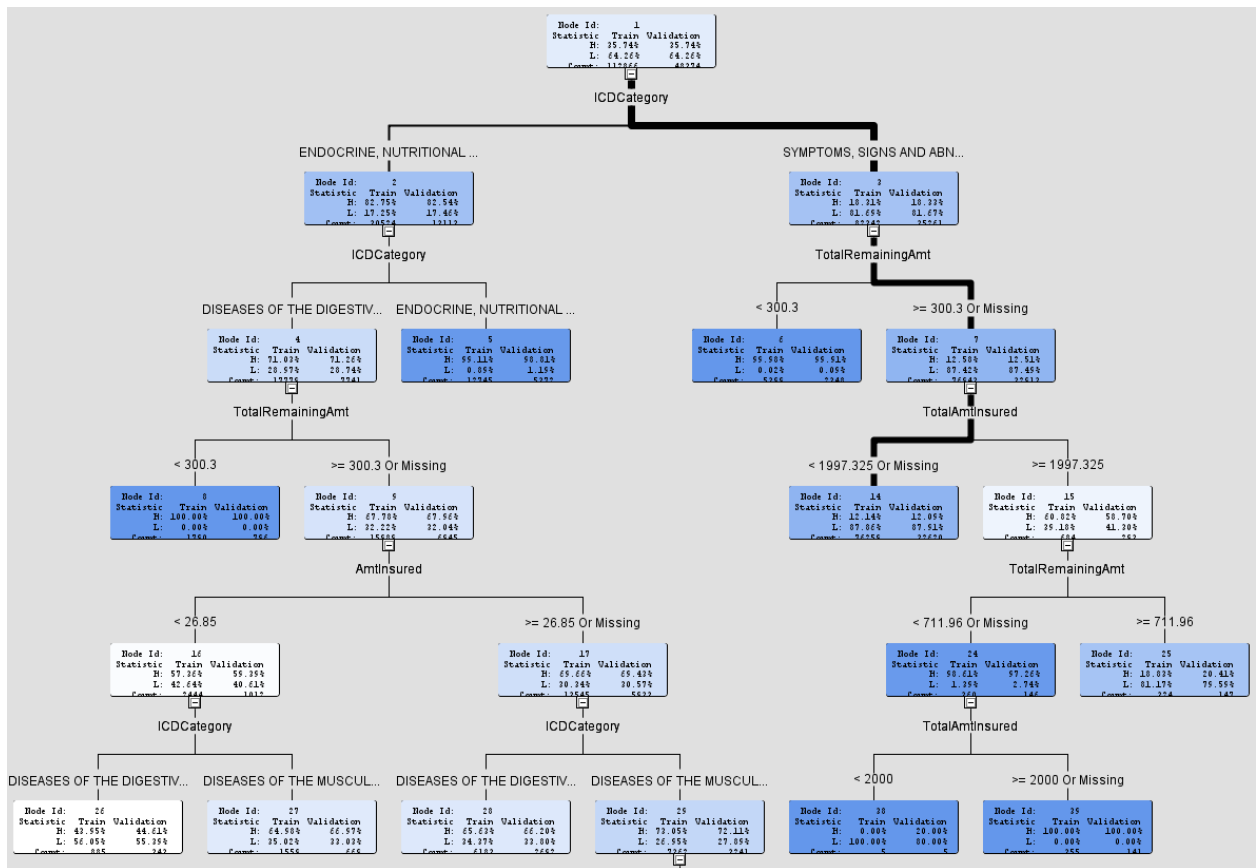


Figure 57: Tree Overview - Stacking Ensemble Model (Base Tree + Meta Tree)

Figure 57 shows the tree overview of the Ensemble Tree constructed by SAS Enterprise Miner. Similarly, the results reflected the outcome of the Default Tree, where it has a depth of 6 and a total of 12 leaves. Looking at the tree overview, the chosen path which is represented by the thicker line width showing a higher volume of observations going to the nodes were as follow: Node 1 > Node 3 > Node 7 and Node 14. The predictors which had an influence on the predictive outcomes are ICD Category, TotalRemainingAmt and TotalAmtInsured. Usually, there would be further exploration of the chosen path to better understand the flow, however, as the results are the reflection of the Default Tree, it is not necessary to explain the same results over.

The level of splits and node rules under the Ensemble Tree model yield the same results as described in detailed under the Default Tree model. Hence, further exploration of the results were not described under this section. In addition, ensemble model of Base Regression and

Meta Tree yielded the same results as the meta learner used is similar to Base Tree and Meta Tree, hence the results did not change. In addition, regression models were not further analyzed as the results did not fulfil the objective of the research.

4.2.4 Model Selection and Evaluation

SAS Enterprise Miner

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree39	Meta Tree	0.12842	0.10363	0.12797	0.10412
	Tree43	Default Tree	0.12842	0.10363	0.12797	0.10412
	Tree46	Meta Tree	0.12842	0.10363	0.12797	0.10412
	Reg5	Regression	0.13968	0.10710	0.13784	0.10819

Figure 58: Fit Statistics - Model Comparison

The last section of the model analysis would revolve around the last node which is Model Comparison. According to the Fit Statistics as shown in Figure 58, the selected model with the label “Y” would be Model Node “Tree39”, of the Model Description of “Meta Tree”. This model would be the Stacking Ensemble Model of (Base Tree + Meta Tree). Based on the results generated by the Model Comparison node, the Stacking Ensemble Model of (Base Tree + Meta Tree) would be the best model for prediction.

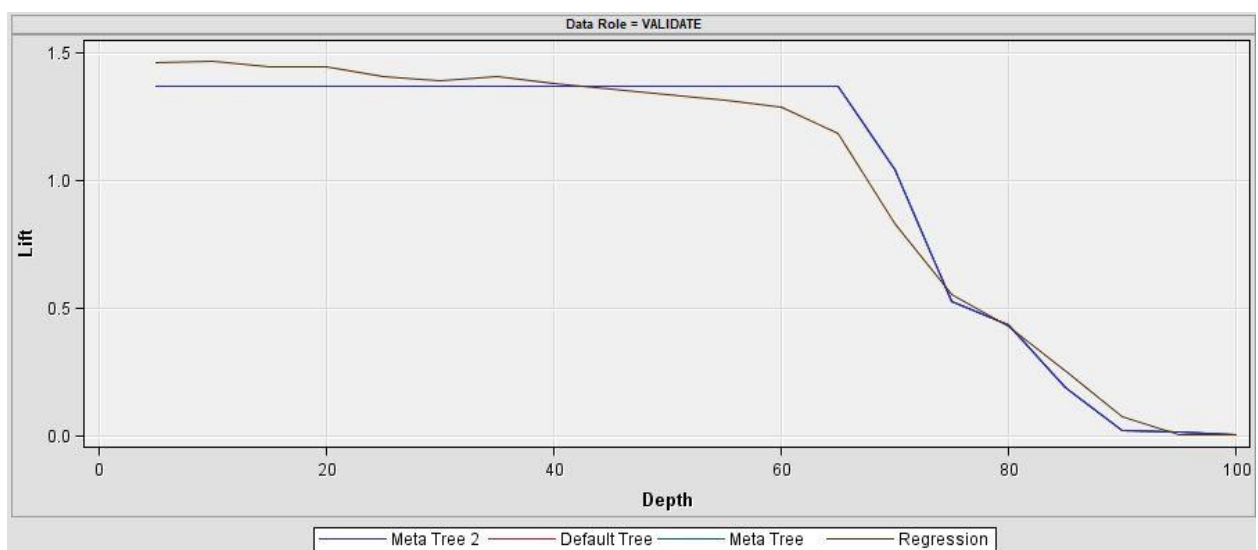


Figure 59: Lift Chart - Model Comparison

Based on the Lift Chart as shown in Figure 59, a basic understanding of Lift would be, it is preferable if Lift is higher. As shown, there were 4 predictive models which were constructed. Default Tree, Meta Tree 2 and Regression were preferable from 0% - 40%, however, between 40% - 75% Meta Tree was preferable then subsequently between 75% - 80% there were minimal differences while between 85% - 95% Default Tree, Meta Tree 2 and Regression were preferred. At approximately 95% all 4 models had minimal differences. In conclusion, with reference to the model selection statistics and lift chart, Meta Tree, was selected as the preferred and best performing model, then followed by Default Tree, Meta Tree 2, and Regression.

As shown in the predictive model construction generated by SAS Enterprise Miner, the two best performing models would be a single Decision Tree model as well as ensemble model of (Base Tree + Meta Tree). And with that, the predictive model selected in this case was the Ensemble Model of Meta Tree. Hence, to test this result and selection, the same models were run on a different platform called Orange. Results are as shown below. Testing was performed to ensure the framework can be applied on various platforms without any restriction or limitation to just a specific platform - SAS Enterprise Miner is a proprietary software while Orange is an open-sourced software.

4.3 Robustness Testing (Healthcare Data and 3 Case Studies)

Robustness testing aims at evaluating the performance of the proposed framework and ensemble predictive model. In this research, robustness testing was performed through two channels, (1) using an open-sourced software, Orange and (2) testing the framework and predictive models on other datasets from varying industries. This provides a better understanding of the performance based on the proposed framework while ensuring the framework and predictive models can be used by different datasets and industries. Firstly, the healthcare dataset was tested using the open-sourced platform, Orange. This will support the findings as shown by SAS Enterprise Miner where the ensemble predictive model will be the best model for prediction with the highest predictive performance and accuracy.

Testing with Orange (an open-sourced platform)

Model	AUC	CA
Stack	0.703	0.730
Tree	0.655	0.666

Figure 60: Orange - Test and Score

As shown above on Figure 60, it is ran using the platform Orange generated by the node of Test and Score. It showed that the AUC (Area under ROC Curve) and CA (Classification Accuracy) both recorded that Stacking Model of Decision Tree was the selected model with AUC of 0.703 and CA of 0.73. Meaning the prediction accuracy of the Stacking Ensemble Model was 73% accurate as compared to a single Decision Tree model where it managed only 66% accuracy.

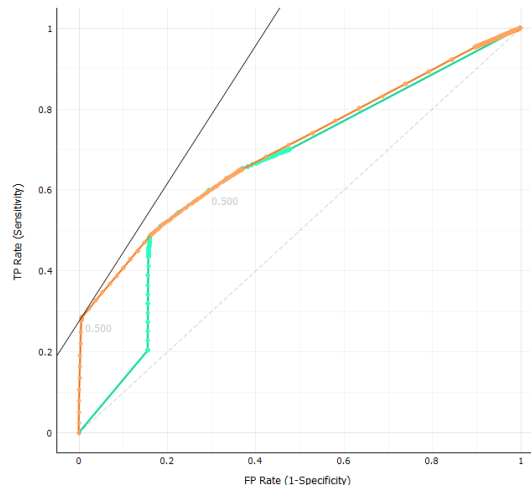


Figure 61: Orange - ROC Chart

Figure 61 shows the ROC Chart for the two models generated using Orange platform - Default Tree and a Stacking Ensemble Tree (Base Tree + Meta Tree). The dotted line would be the baseline, and the further away each lines of each model are away (upwards) from the baseline, the higher the accuracy. From the ROC Chart, orange line represents the Ensemble Tree while the green line represents the Default Tree. As observed, Ensemble Tree has the highest accuracy most of the time. However, even though the accuracy might be the highest, in some cases it may not mean that it is the selected model for prediction due to the specificity and also the complexity. But in this case, as shown in the Test and Score, ensemble model was the selected model due to the higher Classification Accuracy (CA).

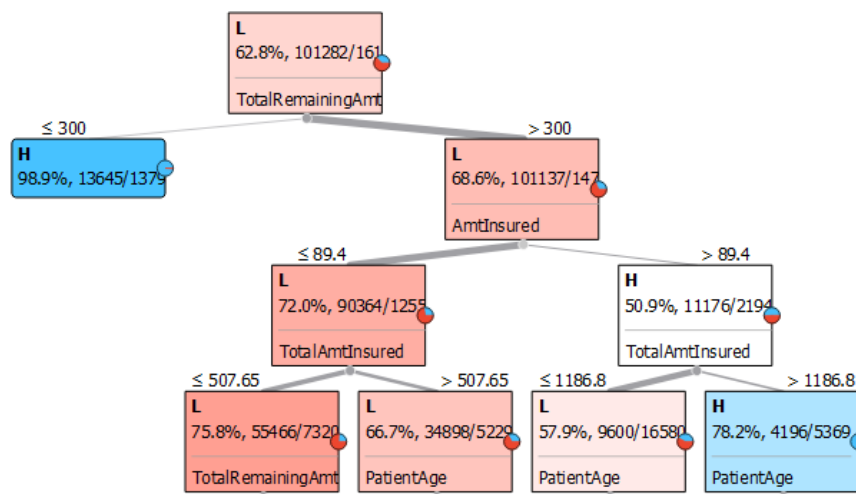


Figure 62: Orange - Tree Overview

As shown above, this is the Tree Overview generated by the platform Orange. Through the tree overview, it is observed that the predictors chosen to predict the target of “RiskLevel” was TotalRemainingAmt, AmtInsured and TotalAmtInsured. Upon closer inspection, the tree overview has a red and blue circle on the right side of each node, this shows the total volume of observations going into each node. For example, looking at the first level split of TotalRemainingAmt of less than or equal to 300 a total of 99% of them are “H” high risk, while above 300 would be 69% “L” low risk, and with the low risk it was observed that the “L” in red occupies around $\frac{3}{4}$ of the circle as compared to the “H” where it occupies almost the whole circle, this shows the number of patients who has the following characteristics would potentially be categorized as a high or low risk patient. Hence, for example, if the TotalRemainingAmt of the patient is less than or equal to 300, there is a 99% chance of high risk. However, this is not entirely true as it depends on the diagnosis as well, which was observed in the tree overview generated by SAS Enterprise Miner. So, the following concludes the comparison between the two platforms of SAS Enterprise Miner and Orange.

To further test the predictive accuracy and outcomes, 3 case studies were conducted to validate the performance and predictive accuracy of the proposed ensemble model. There is a caveat to call out where there are similarities between the data structures of the employee healthcare data and case studies. The target variable must be classification / categorical based meaning, it must be a Yes or No, 1 or 0 or even High, Medium, and Low outcome - there should not be any numerical target values. Additionally, they are structured data where it fits neatly into data tables and includes discrete data types such as text, numbers, and dates.

Case Study 1: Customer Churn - Retail, Loyalty Program

Explanation for each dataset

Table 12: Customer Churn Dataset

Dataset	Description
CubeData	This dataset has 144 variables and 509473 rows of data. It consists of the member's general information such as gender, age, and race. It also contains membership information such as membership type, points balance, spending pattern, redeeming pattern, points expiry, and customer activity. The dataset includes information from June 2016 - June 2017.

A loyalty program is often provided by a company / business to their customers as part of a brand membership. Through a loyalty program, customers may have priority access to new products / services, special promotions, or even exclusive free gifts which will motivate customers to join the loyalty program. Generally, a customer would exhibit loyalty through consistent use of products / services for an extended period. The aim of this case study is to better understand the factors affecting customer churn. By understanding the factors, potential mitigating ways can be proposed to prevent customer churn. Past research have conducted churn analysis in various industries such as banking and telecommunications (Karvana, Yazid, Syalim, & Mursanto, 2019) (Halibas, et al., 2019). In the case study, 2 models were applied which are Decision Tree and a stacking ensemble model using (Base and Meta Tree).

Moving on, the proposed framework was tested and ensemble predictive model on a customer churn environment within the retail loyalty program segment. Similarly, using the open-sourced platform, Orange to generate the following predictive results.

Model	AUC	ČA
Stack	0.730	0.743
Tree	0.664	0.731

Figure 63: Orange - Test and Score (Customer Churn, Loyalty Program)

Orange was used to run the same dataset of churn which yielded a result in favour of the Ensemble Tree of (Base Tree + Meta Tree). The recorded AUC (Area under Curve) was 0.730

as compared to the Default Tree which was 0.664 while the classification accuracy was also in favour of the Ensemble Tree at 0.743 while Default Tree recorded 0.731. As shown in both the proprietary (SAS Enterprise Miner) and open-sourced (Orange) platform, ensemble model tree of (Base Tree + Meta Tree) was the selected predictive model to perform prediction while applying the proposed framework. Testing would justify that the proposed framework is robust while having the capability of performing predictions in various scenarios.

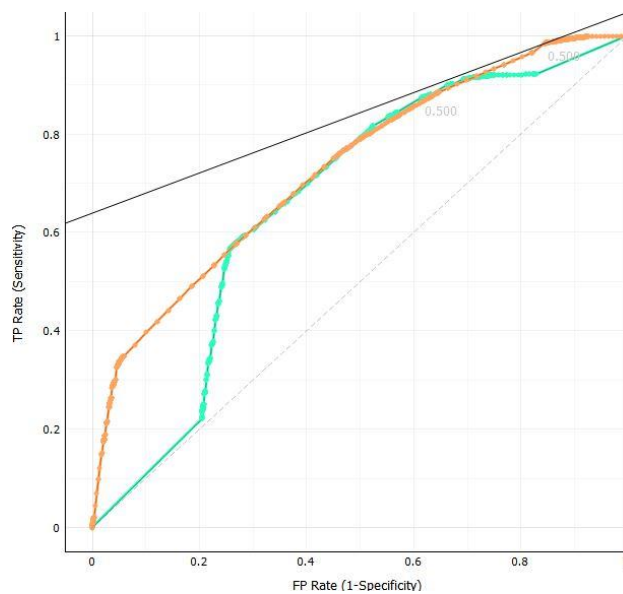


Figure 64: Orange - ROC Chart (Customer Churn, Loyalty Program = “No”)

Figure 64 shows the ROC chart for customer churn = “No”, the two models generated using Orange platform - Default Tree and a Stacking Ensemble Tree (Base Tree + Meta Tree). The dotted line represents the baseline, and the further away each lines of each model are away (upwards) from the baseline, the higher the accuracy. From the ROC chart, orange line represents the Ensemble Tree while the green line represents the Default Tree. As observed, the Ensemble Tree has the highest accuracy most of the time.

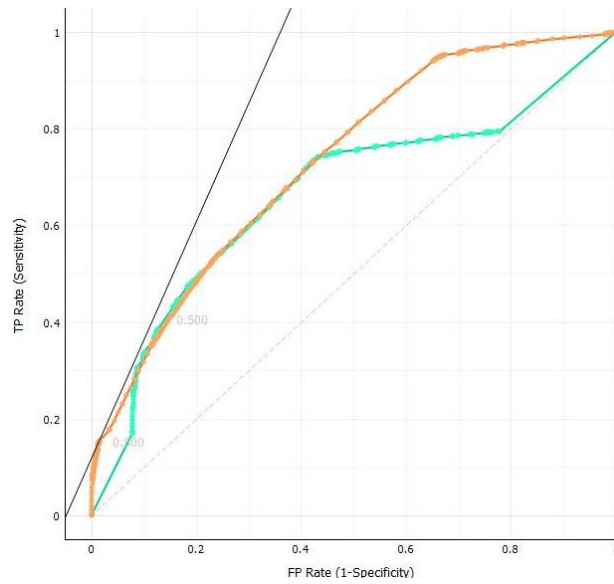


Figure 65: Orange - ROC Chart (Customer Churn, Loyalty Program = “Yes”)

Figure 65 shows the ROC chart for customer churn = “Yes”; similarly, the two models generated using Orange platform - Default Tree and a Stacking Ensemble Tree (Base Tree + Meta Tree). As shown through the ROC chart, the orange line representing the Ensemble Tree performs better than the green line representing the Default Tree. As observed, the Ensemble Tree has the highest accuracy most of the time which is the same as recorded in Figure 64.

Case Study 2: Loan Risk Defaulted

Table 13: Loan Risk Default Dataset

Dataset	Description
Application Data	This dataset has 122 variables and 307512 rows of data. It consists of the applicant’s gender, nature of contract, amount of income, marital status, and income type. The dataset includes information from 2020.

The following case study focuses on applying for a loan in a real business case scenario from a financial institution. The following data is used to minimize the risk of losing money when applicants apply for a loan. This dataset will provide a better understanding of what characteristics should be considered when predicting the potential risk of loan defaulting. There are several research which have investigated the prediction of loan default where the

researcher, Yiheng Li, used a combination of logistic regression and neural network to improve the predictive outcome (Li & Chen, 2021). Another researcher also applied neural network but in combination with a hybrid sampling method that combines clustering with stochastic measure (Chen, Zhang, & Ng, 2018). These research utilized complex data mining techniques which could be an issue for interpretation by practitioners who are not experts in the field of analytics. Hence, the following stacking ensemble model was proposed which would ease the interpretation and enable practitioners to apply the following framework across various classification scenarios.

Lastly, the approach was tested on a totally different industry. The ensemble model approach was applied in a banking industry to better understand the loan risk default behaviour and what may cause loan application to be defaulted. Similar, the dataset was obtained through the online portal, Kaggle. Results were generated through the open-sourced platform, Orange.

Model	AUC	CA
Tree	0.918	0.969
Stack	0.919	0.969

Figure 66: Orange - Test and Score (Loan Risk Defaulted)

As shown through the results generated through Orange, it showed that the predictive outcomes were just marginally favourable towards “Stack” Ensemble Tree where the base and meta model were using decision trees. The recorded AUC shown 0.918 for the “Tree” default tree but 0.919 for the “Stack” tree. This showed almost identical predictive results which translates to both the stacking model, or the individual predictive model achieved almost identical results with the stacking model edging the individual tree marginally. The classification accuracy achieved exact results at 97%. However, it does show that even though the results may be similar, “Stack” tree still achieves a better predictive outcome which has been proven by the previous results as shown. All 3 industries, Retail, HR, and Financial Institution showed favourable results towards the ensemble model approach. With that, the results showed that the

proposed approach of a stacking ensemble model has the robustness to be applied in varying industries.

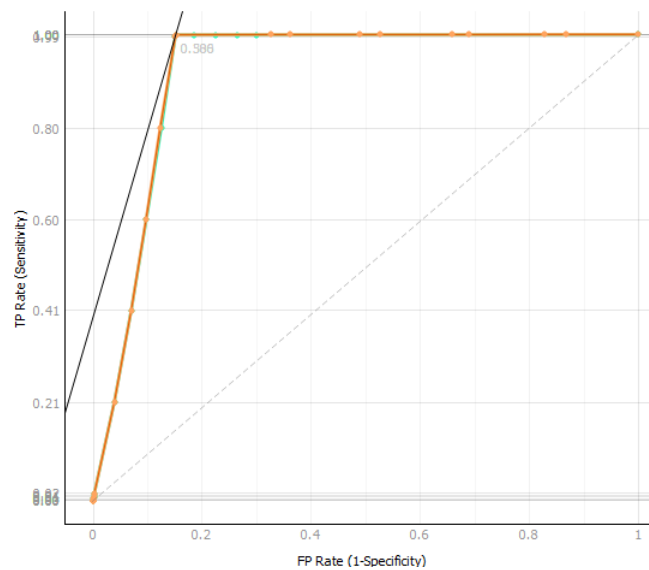


Figure 67: Orange - ROC Chart (Loan Risk Defaulted = "No")

Figure 67 shows the prediction of loan risk defaulted = "No", meaning the risk of loan being defaulted is low. "Stack" tree showed that it performed almost identical to the "Tree" default tree. This has been proven through the results which has been described.

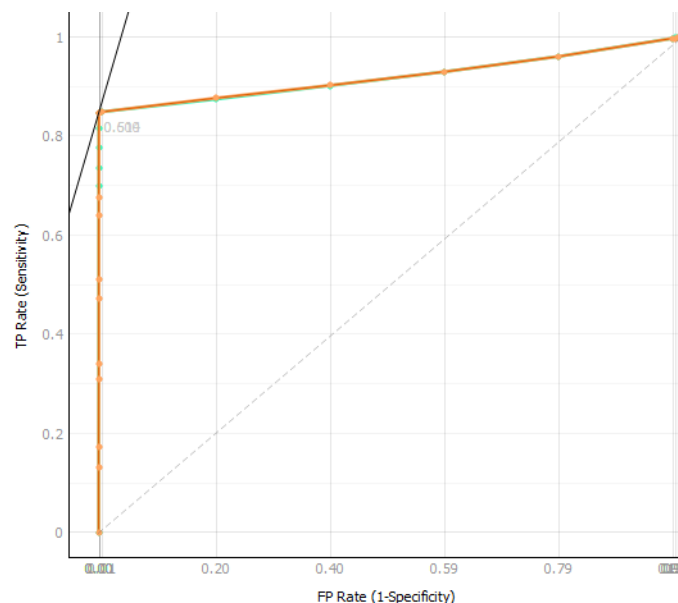


Figure 68: Orange - ROC Chart (Loan Risk Defaulted = "Yes")

Moving on, in Figure 68 the ROC chart shows loan risk defaulted = "Yes", meaning the risk of loan being defaulted is high. Similarly, the results of "Stack" tree showed that it performed

almost identical to the “Tree” default tree. Unfortunately, through the following results, it did not show drastic differences but minimal variance between a stacking ensemble model and an individual model, but it does prove that by applying a stacking ensemble model approach, it can achieve a higher predictive accuracy and it can be applied across various classification environment which has been proven in 3 differing industries (Retail, HR, and Financial Institution).

Case Study 3: Employee Attrition

Table 14: Employee Attrition Dataset

Dataset	Description
IBM HR	This dataset has 35 variables and 54278 rows of data. It consists of the employee's general information such as gender, age, and department. It also includes hourly rate, job role, job satisfaction, marital status, and monthly income information. The dataset includes information from 2017 - 2018.

The following dataset was created by IBM data scientists to identify the employee attrition rate based on the variables provided. Moreover, the focus was to uncover the factors which has direct relation to employee attrition and exploring the important question of what contributes to employee attrition rate. Some of the past research have used various data mining techniques to look at the employee attrition prediction. A paper by Alduayj, used machine learning models such as Support Vector Machine and K-Nearest Neighbour to perform the prediction (Alduayj & Rajpoot, 2018). One of the main concerns that are faced by companies and businesses alike would be the loss of talented employees (Alduayj & Rajpoot, 2018). Another paper used Decision Tree as the main classification technique to predict employee attrition while building a framework for HR to analyse precise behaviours and characteristics contributing to attrition (Yadav, Jain, & Singh, 2018).

Secondly, the proposed approach was tested in an employee attrition environment, this is a dataset used by IBM Human Resource Department where they would like to better understand customer attrition patterns and behaviour. Similarly, the dataset was obtained through the online portal, Kaggle. Results were generated through the open-sourced platform, Orange.

Model	AUC	CA
Tree	0.532	0.771
Stack	0.570	0.837

Figure 69: Orange - Test and Score (Customer Attrition)

As per the results generated through Orange, it showed that favourable predictive outcomes for “Stack” Ensemble Tree where the base and meta model were both decision trees. However, the recorded AUC were not ideal as it was recorded at 0.570 for the “Stack” tree and 0.532 for the “Tree” default tree. The result for AUC reflects the probability of 50-50, which is not ideal in a predictive model. But, for the classification accuracy, it showed that “Stack” tree would obtain 0.837 while the “Tree” default tree would manage 0.771. This would translate to approximately 83% and 77% classification accuracy, respectively. The less-than-ideal results could be due to the small dataset which is similar to the previously telco customer churn. Through the results though, it has proven that an ensemble tree would yield higher predictive performance, which again has successfully shown that the proposed framework of classification ensemble predictive model has the robustness to be applied in different environments.

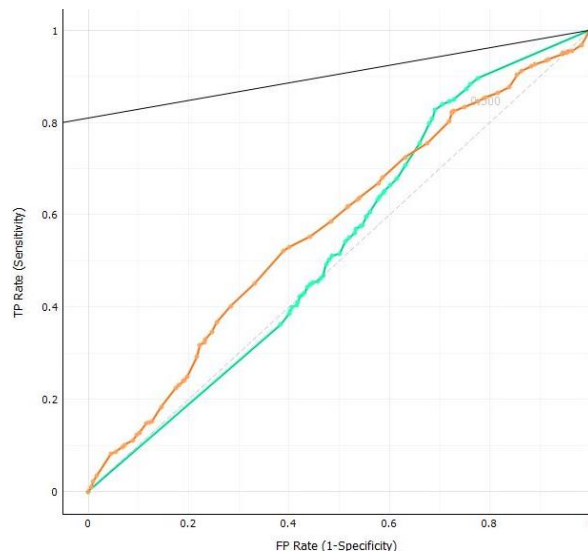


Figure 70: Orange - ROC Chart (Customer Attrition = “No”)

Figure 70 shows the prediction of customer attrition = “No”. “Stack” tree showed that it performed better until specificity of 0.6 then the “Tree” default tree would perform marginally better than the “Stack” tree. However, as shown by the AUC in Figure 69, the predictive outcomes were not ideal as it is reflected on the ROC chart where the dotted line representing the baseline would indicate the predictions as a probability of 50-50.

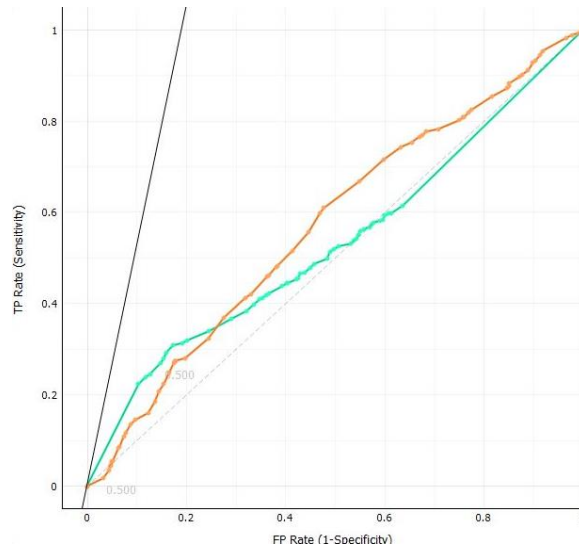


Figure 71: Orange - ROC Chart (Customer Attrition = “Yes”)

Moving on, in Figure 71 the ROC chart shows customer attrition = “Yes”. The difference immediately observed would be “Tree” default tree performed slightly better than the “Stack” tree until specificity of 0.25, subsequently “Stack” tree would perform better than “Tree” default tree. Through these results obtained, it showed that predictive were not satisfactory, but ensemble predictive model would still yield higher predictive performance and accuracy. This would prove that by applying the proposed framework of classification ensemble predictive modelling, it can be applied in various industries, through different datasets - with the only condition that prediction must be classification predictions.

5. Summary and Discussion

Firstly, an overview of GP demographics was performed to identify the number of patients and claims made, while understanding the split between male and female patients as well as the patient age group. This will provide an overview of which gender or age group were driving claims under GP in year 2018. Through further analysis, E which represented employee were driving most of the claims as expected at 9,203 patients who made 55,484 claims while 2nd would be C which represented child at 4,919 patients who made 15,791 claims and SP which represented spouse had only 2,599 patients who made 8,664 claims. This showed that child had almost 50% more patients who made approximately 50% more claims than spouse. The top 10 diagnoses were identified, however, what drawn the attention would be “Acute Upper Respiratory Infections” was the most common diagnosis recorded while “Fever” was the 2nd most common diagnosis and 3rd was “Upper Respiratory Tract Infection”. “Acute Upper Respiratory Infections” and “Upper Respiratory Tract Infection” are both commonly known in layman terms as sore throat, but the written diagnosis when recorded were different, however they both had the same ICD Code of J06.9. Another two diagnoses which were shown raised some concerns as it were “Low Back Pain” and “Hypertension” - this showed that there are 2 chronic conditions which were commonly diagnosed among patients, even though there were only 2,945 and 1,232 claims respectively, the focus will be on these findings. According to an article written by Daniela Koller, “Hypertension” and “Low Back Pain” are chronic conditions within the list of 46 chronic conditions based on ICD-10 codes (Koller, et al., 2014). Chronic conditions would refer to outcomes with “long-term care dependency” (Koller, et al., 2014). More importantly, a research was carried out to identify the relationship between “Low Back Pain” working postures, among individuals who stand and sit most of the working day, the results led to findings which showed that standing at the workplace has associated with “Low

Back Pain” issues in both men and women as shown in the findings here as well (Tissot & Stock, 2009).

Moving on, the major difference between male and female patients’ diagnoses under GP in year 2018 showed that “Hypertension” was recorded for males but not females within the Top 10 Diagnoses under GP. This is quite prevalent among research which have been performed, where “Hypertension” among men consistently have higher levels of “Hypertension” as compared to women of the same age (Everetti & Zajacova, 2015). Similar in another article written by Ellen, and I quote: “distinct gender differences in the incidence and severity of hypertension are well established where males have a higher incidence of hypertension compared to females of the same age” (Gillis & Sullivan, 2016). Even though both males and females recorded “Low Back Pain” issues, males made 1,932 claims while females made 1,013 only - which is almost an increase of 1,000 claims. An assumption which can be made would be males were working in more stressful working environments as compared to females such as Construction - this will be confirmed in the latter stages. This assumption was made based on a research done by AXA where they ranked Building and Construction as the 3rd most stressful job, behind Accounting and Finance and Cleaning Services (Gerrard, 2018). Apart from performing analysis on gender, patients were categorized based on their relationship as well, as mentioned, there were 3 categories, E, SP and C. By categorizing accordingly, it provided a better overview of the claims made within various relationship groups such as the diagnoses and age distribution. This categorization showed that the 2 chronic conditions were only present among E and SP. On the other hand, “Acute Upper Respiratory Infections”, “Fever” and “Upper Respiratory Tract Infection” were common diagnoses among patients regardless of gender and relationship.

A trend analysis looked into the analysis of the claim patterns based on date of claims by months. This is performed to find out the pattern of claims and if it is affected by special

seasons/occasions such as semester breaks, school holidays or festive seasons. These will provide a better understanding of the claim patterns throughout the year. Next, to identify all major term holidays, school holidays and festive seasons then map it against the claims made to identify correlation. School Holidays were from *17th - 25th March; 9th - 24th June; 18th - 26th August and 24th November onwards until 31st December* (One Stop Malaysia, 2018). Festive seasons in Malaysia include Chinese New Year (16th - 17th February), Hari Raya Aidilfitri (15th - 16th June) and Deepavali (6th - 7th November) (One Stop Malaysia, 2018). Besides holidays and festive season, there is a season known as Southeast Asia Haze Season which happens every year between the months of July and August (Kawi, 2018) (EdgeProp MY, 2018) - the reason behind the focusing on haze is due to the high “Upper Respiratory Infections” diagnosis recorded, which triggered an activity to identify if there could be some correlation. An overview of the claim trend were showed based on claims per month, and as observed the number of claims peaked in the month of January at 7,748 claims while December was the lowest at 5,953 claims. Others were July (7205), March (6965) and October (6880) which seen high traffic and number of claims made under GP while the months of June (6098) and September (6101) recorded lower number of claims. Based on the school holidays, there were two correlation which can be made which are 17th - 25th March and 9th - 24th June where in the month of March the number of claims spiked, but even though June was a holiday the number of claims made were one of the least - this led to an assumption where the number of claims made throughout the year has no correlation with term or school holidays. Besides term or school holidays, there were no correlation between the number of claims and festive seasons as well, as during the festive seasons, claims were on the lower end. Hence, similarly, there were no correlation between festive seasons and the trend of claims made. Looking at the trend of claims for the diagnosis of “Upper Respiratory Infections” - (diagnosis of “Acute Upper Respiratory Infections” and “Upper Respiratory Tract Infections” were combined as they

represent the same diagnosis), the highest number of claims were made in January and October, while in the months of July and August, even September, the number of claims were on the lower end except for July, where it was the 4th highest month to record over 1,100 claims for “Upper Respiratory Infections” but even so, there are no concrete evidence which suggest that the “Upper Respiratory Infections” diagnosis spiked during the haze period of July and August. Hence, all hypothesis of the possibility of a spike in claims made during semester breaks, school holidays, festive seasons, and haze season, were not supported by any of the findings made. Looking at the claim trend of “Low Back Pain” and “Hypertension”, “Low Back Pain” peaked considerably in the months of April and July while the lows were sub-200 claims in the months of January and February while “Hypertension” claims peaked in the months of May and June while the lows were in January and September. An interesting finding between the 2 chronic conditions would be the months of highs and lows, they have a similar pattern where the highs were mid-year between April to July while the lows were beginning of the year in January. However, for both of these chronic conditions they appeared more mid-year and the lows were at the beginning and end. An assumption which was made, there are more work to be done towards the mid-year, hence, increased stress as there were more deadlines to meet leading to an increase in these 2 chronic conditions.

The following part of the discussion focuses on 3 specific areas of diagnoses of “Upper Respiratory Infections”, “Low Back Pain” and “Hypertension”. The reason being the focus on these 3 diagnoses was because “Upper Respiratory” issues were the most common among GP claims, while “Low Back Pain” and “Hypertension” were 2 chronic conditions which would be a cause of concern for employers. Looking at the most common diagnosis recorded among employees as well as chronic conditions which are the most prominent would enable employers to reduce such complications, as diagnoses are too widespread and there are too many diagnoses which an employer can focus on, hence, narrowing down the spectrum to specific

types such as most common and common chronic conditions would enable employers to focus on specific problems to solve the current employee population health issues. Looking at “Upper Respiratory” issues, it is commonly recorded among both males and females while within male claims, most patients were from Sunway Construction Group, Sunway Resort Hotel & Spa and Group Security and within female claims, most patients were from Sunway Education Group, Sunway Resort Hotel & Spa and Monash University. Some assumptions which were drawn for males would be related to firstly, long hours in the outdoor working environments for construction and security employees and due to the lack of hydration and for hotel employees and female employees, could be due to the long hours within air-conditioned working environments and constant communication with customers, students, or travellers. Moving on, “Hypertension” issues were recorded more within the older age groups of 41 and Above but Low Back Pain was found more commonly among age groups of 22 - 40. This would be concerning because younger individuals are diagnosed with Low Back Pain issues more commonly and it should be further investigated by the HR. Both “Hypertension” and “Low Back Pain” were more commonly recorded among male over female employees. As discussed with the HR, generally age is one of the factors causing “Hypertension” issues which is considerably true based on the analysis. As for “Low Back Pain” issues, it is more prominent among employees which require constant walking and standing which based on the analysis, it is proven as among male employees, they were from Sunway Construction Group, Sunway Resort Hotel & Spa and Group Security while female employees were from Sunway Education Group, Sunway Resort Hotel & Spa and Monash University - there are considerably longer hours of standing and walking within these industries. This is proven by a research performed by WebMD, where they suggest that working on one’s feet for long hours could spell trouble and contribute to long-term back pain and musculoskeletal disorders (Mozes, 2015).

To determine the usage of medical coverage as provided by the employers, the approach taken was to perform a mathematical calculation to derive the total remaining amount available for an employee. To derive the total remaining amount value; each employee is given an employee annual limit signifying the yearly medical insurance coverage amount given by the employer. By taking the subtraction of annual limit minus total amount insured ($\text{AnnualLimit} - \text{TotalAmtInsured}$), the total remaining amount would be calculated. Total amount insured is a value derived by taking every claim performed by an employee in a given year and performing the mathematical solution of addition based on the employee's insured amount per visit to the clinic or hospital. In this section, a better understanding of the current usage of the medical coverage provided by the employer, to potentially either reduce expenditure or identify patients who are spending excessive amount or have minimal remaining amount. The analysis performed here solved objective #3 to discover characteristic and usage pattern of healthcare benefit provided by the employer. Out of the 9,203 patients who were employees, 75 of them fully utilized the coverage while 1,337 of the patients have almost fully utilized, that is a total of 1,412 (15.3%) employees. Out of the 1,412 employees, 985 were males while 427 were females. This analysis showed that only a minimal number of patients (15.3% to be exact) had < 1000 remaining amount which would translate to the medical coverage being more than sufficient for 85% of the employees, there is a minimal group who had overspend. Thus, the focus of potentially reducing medical expenditure would be focused on identifying the patients (employees) who have chronic conditions to prepare for pro-active measures.

Subsequent sections discussed were on Specialist claims made in year 2018. There were 3,232 SP patients who made 8,701 claims including E (employees), SP (spouse) and C (child). Out of the 3,232 patients, 2,108 (5,792 claims) were employees, 832 (2,212 claims) were child and 292 (697 claims) were spouse. The top 10 diagnoses recorded under SP claims were shown. Based on the list, most were occupied by claims made by relationship = C such as "Upper

Respiratory”, “Fever”, etc. As further analysis showed that, most employee claims recorded diagnosis such as “Hypertension”, “Low Back Pain”, “Coronary Artery Disease” and more, while most spouse diagnosis were “Unspecified Abdominal Pain”, “Hypertension” and more. The following section focused on In-Patient demographic analysis. It showed the difference between the IP analysis and the encounter analysis. Every patient is given a label under Type of Claims where there is GHS and PostGHS, GHS refers to general hospitalization while PostGHS refers to post general hospitalization. For every patient who get admitted, they would have a GHS tagged to the admission, however subsequent follow ups refer to the PostGHS. Hence, PostGHS is not a claim, as it is considered as follow up. So, to look at each encounter, the decision was to filter out and re-create the encounter variable. So instead of 1,034 patients and 3,839 claims, the number of patients and encounters are much lesser, 950 patients, 1,134 encounters. Both male and female employees were commonly admitted for “Gastritis” and “Dengue Fever” encounters.

Finally, the model analysis which was done. Decision Trees would be the suggested predictive model because the target “RiskLevel” is nominal. However, 5 different predictive models (2 single models and 3 ensemble models) were built to identify the best performing model. As shown and described, the selected model by the Model Comparison node was Stacking Ensemble Model of (Base Tree + Meta Tree). This model achieved a prediction accuracy of 87% while the selected predictors to achieve these prediction would include: ICD Category, TotalRemainingAmt and TotalAmtInsured. With these 3 predictors, it would be able to segregate between the low and high-risk individuals. These showed that the ICD Category a patient would be diagnosed with, together with the spending behaviour through the TotalRemainingAmt and TotalAmtInsured would potentially predict the RiskLevel of a patient. Looking at the TotalRemainingAmt and TotalAmtInsured only may lead to bias results, as a patient might have used up the annual medical benefit due to “Accidents”, hence this

cannot be categorized as “H” high risk, which is why by including the ICD Category, it would add value to the prediction. However, the Default Tree also yielded the same results as mentioned - prediction was then performed on another platform called Orange where the 2 best performing models which were the Default Tree and Ensemble Tree to test the results. Similarly, the results showed that Ensemble Tree would be a preferred predictor over the Default Tree.

To address the concerns and achieve the objective of the research, the focus was on proposing a practical classification stacking ensemble model which can be explored and utilized by practitioners who are experts and non-experts in the field of analytics. In general, analytics are still within the experimental phases while some have applied analytics in complex scenarios, real-world applications are still dampened by the complexity and increasing focus on model accuracy (Alharthi, 2018). Focus is too clinical with too many clinical identifications and predictions while little focus has been put into improving interpretability and understandability (Alharthi, 2018). In other areas, predictive analysis and machine learning techniques applied were mathematical formulations and statistical calculations which again increases complexity and would be a challenged to be understood by other who are not experts in the field (Alharthi, 2018). The propose stacking technique will be able to bridge the gap between the complexity and focus on driving higher accuracy and allow for easier interpretation of a predictive model. The proposed stacking ensemble model focuses heavily on the combination of feature selection, feature engineering and stacking model using the base and meta model. There are many variants to an ensemble model by using Boosting / Bagging techniques. However, there are several characteristics which are unique to the Stacking technique as compared to the other 2 techniques such as it is a heterogeneous learner (whereby it allows for the flexibility of combining various learning algorithms) while bagging and boosting are homogenous learners and do not have such flexibility. More importantly, stacking ensemble model provides

advantages such as simplicity, ease of interpretation, improved performance, and the flexibility to combine models induced by various learning algorithms (Menahem, Rokach, & Elovici, 2009) (Rocca, 2019).

By having the ability to reduce / correct the errors of the previous models, it improves the predictive accuracy without altering the complexity of the predictive model. The weaker learners / base models acts as the building blocks to design more interpretable and enhanced predictive outcomes (Rocca, 2019). Moreover, in a stacking model, it provides a platform to reduce bias and variance by using the combinational approach to produce stronger, and more robust models which can be applied across various industries - this reduces the industry bias / data bias model issue.

Table 15: Platform Testing (SAS Enterprise Miner and Orange)

No	Platform	Predictive Technique	Model	Classification Accuracy	Selected Model
1	SAS Enterprise Miner	Ensemble (Stacking)	Base Regression + Meta Tree	86%	N
2		Single	Regression	87%	N
3		Single	Decision Tree	87%	N
4		Ensemble (Stacking)	Base Tree + Meta Tree	87%	Y
No	Platform	Predictive Technique	Model	Classification Accuracy	Selected Model
1	Orange	Ensemble (Stacking)	Base Tree + Meta Tree	73%	Y
2		Single	Decision Tree	67%	N

Through the following robustness test, it was concluded that the following proposed stacking ensemble model can be applied in various platforms. In the following test case, the datasets were applied on 2 different platforms, 1 is a proprietary software (SAS Enterprise Miner) which mainly uses SAS programming language as the foundation while the other is an open-sourced free software (Orange Data Mining) which mainly uses R programming language as the programming engine. By applying the same approach whereby, using a single predictive technique against the ensemble stacking approach to perform a predictive comparison. Through the predictive results, the stacking ensemble model was the selected model for SAS Enterprise Miner because of the reduced complexity and increased robustness even though the classification accuracy were identical. On the other hand, on the Orange platform, the results

were clearer which shows that the stacking ensemble model achieved 6% higher predictive accuracy at 73% instead of 67% of a single predictive technique. This result justifies the robustness of the proposed model when applied on differing predictive platforms and programming languages.

Table 16: Robustness (3 differing industries - Retail / HR / Financial Institution)

No	Industry	Predictive Technique	Model	Classification Accuracy	Selected Model
1	Retail	Ensemble (Stacking)	Base Tree + Meta Tree	74%	Y
		Single	Decision Tree	73%	N
2	HR (Employee Attrition)	Ensemble (Stacking)	Base Tree + Meta Tree	83%	Y
		Single	Decision Tree	77%	N
3	Financial Institution (Loan Default)	Ensemble (Stacking)	Base Tree + Meta Tree	97%	Y
		Single	Decision Tree	97%	N

Moving on, the same concept was applied by testing the robustness of the proposed stacking ensemble model on 3 different industries: Retail, HR, and Financial Institution. This test is to prove that the proposed stacking model is not bias and dataset / industry specific which again proves the robustness. The same approach was applied by using 2 different predictive techniques which are a single model and a stacking ensemble model. Similarly, as shown through the results, the proposed stacking ensemble model approach would achieve higher predictive accuracy across all 3 predictive outcomes. The following test cases were all classification predictive outcomes which translates to 1 / 0 or binary results, hence the models applied were mainly Decision Tree - this is the true beauty of a stacking ensemble model where the flexible capability comes into play. If a target value is a nominal or numerical value, the stacking ensemble model allows a combination of base model as Decision Tree and meta model as Regression to predict the numerical outcomes. By combining various algorithms, one will be able to apply the following stacking ensemble model approach in combination with a hybrid feature selection and feature engineering.

Table 17: Past Research Comparison

No	Author	Predictive Technique	Model	Classification Accuracy
1	Mohammad Hossein	Single	Decision Tree	81%
2	Ritesh Jain	Single	Decision Tree	76%
3	Sai T Moturu	Ensemble (Boosting)	LogitBoost	76%
4	Nicholas (Researcher)	Ensemble (Stacking)	Base Tree + Meta Tree	87%

Finally, by comparing previous research against the proposed predictive model, it provided the opportunity to prove that the proposed stacking ensemble model technique can achieve higher predictive accuracy without increasing the complexity to interpret / translate a statistical / mathematical predictive outcome. As shown in the following table, 2 research deemed a single predictive technique of Decision Tree as the best performing predictive model while another research applied the Boosting technique through LogitBoost. However, the results could not compare to the stacking ensemble model technique which achieved 87% predictive accuracy. This shows that a stacking ensemble model is the preferred option, and it shows the superiority and predictive performance over other ensemble techniques. Through the robustness test, the proposed ensemble model was tested on various datasets / industry and various platforms to conclude that the following proposed model does indeed increase the robustness as mentioned (Menahem, Rokach, & Elovici, 2009) (Rocca, 2019).

6. Conclusion

In conclusion, this research has showed how patient's medical claim patterns and behaviours would potentially affect risk level. Based on the claim pattern and behaviour analyzed, potential strategies could be applied to minimize or reduce such diagnoses while the potential of minimizing medical expenditure can be further explored based on the claim pattern of the patients. Analysing medical claim patterns and behaviours is possibly useful for employers to make decisions such as increasing medical premiums and healthcare plans for their employees. As mentioned, healthcare and medical expenditures have been increasing exponentially over the years, hence, it has triggered organizations and businesses to make the decision to further explore and understand their current employee population health to better understand the claim patterns and behaviours of their employees. Through this analysis, characteristics which affect risk level could be further explored; for instance, TotalRemainingAmt, TotalAmtInsured and ICD Category are factors which have an influence in the prediction of risk level in a patient. These findings will enable employers to make better decisions to prepare proactively measures instead of reactive measures. Moreover, the objective of creating a practical classification ensemble model framework was achieved which also yield better and more accurate predictive results. The framework can be applied in various datasets across a wide range of platform (proprietary and open-sourced) where the prediction type would be classification. The concept of this approach was driven by the issue which has not been addressed whereby predictive models focuses on enhancing and improving accuracy without addressing the issue of practicality and usability by practitioners who are not experts in this field.

Through descriptive and predictive analysis, it successfully showed the characteristics of patients who have the potential to be high risk or low risk. Moreover, through the descriptive analysis and clustering, it presented an opportunity to better understand claim patterns and behaviors while recommendations can be rolled out to possibly prepare proactive measures.

The recommendations mentioned below are very much actionable strategies - it could be implemented upon approval by the Group HR. These are short-term recommendations to curb the current top diagnoses issues which drawn the attention of the Group HR. Furthermore, to ensure the success of such implementations, they could go through a trial period to test the effectiveness of such recommendations. Of course, there are many other recommendations which could be adopted besides the suggested ones below, however, based on the short-term goal in mind, these could be more prominent and effective.

6.1. Research Contribution

a. Theoretical

The proposed stacking ensemble model, in conjunction with its associated framework, tackles crucial design challenges within the data mining lifecycle which has been prevalent in previous research too. Despite the prevalent emphasis on predictive accuracy and performance, Data Preparation often receives insufficient attention. However, within this framework, Data Preparation is explicitly underscored as a pivotal phase in ensemble model construction. Phase 2 encompasses essential tasks like Data Cleaning, Data Processing, Exploratory Data Analysis, Feature Engineering, Feature Selection, and Data Partition. Furthermore, the framework prioritizes the infusion of Model Diversity and Flexibility, recognizing it as a cornerstone in ensemble model design. This entails imbuing both feature and algorithmic levels with diversity and flexibility to adapt to varied predictive scenarios, thus optimizing ensemble model potential. Model Selection within this framework entails two key facets: the identification of base models for training and the selection of diverse learning algorithms based on predictive outcomes. Beyond addressing fundamental design issues, empirical research validates the efficacy of the model diversity approach. Case studies across different industries, including healthcare (predicting high Health-Risk

employees), loyalty programs (forecasting Customer Churn), HR (anticipating Employee Attrition), and finance (predicting Loan Default Risks), underscore the significance of injecting Model Diversity and Flexibility. This research is focused on the stacking ensemble technique, fills a critical gap by addressing flexibility concerns often absent in other ensemble methods like boosting and bagging.

b. Practical

The following framework and stacking ensemble model proposed can contribute to the following practical implications. Firstly, as mentioned there was growing concern with the consistent increase in medical expenditures, hence, the following research bridges the gap to help employers understand the overall employee health population and predict a potential high-risk/cost individuals. Through the following research, the issue of the consistent increase in medical expenditure has been address by proposing the flexible medical coverage selection. These insights were extracted from the past medical usage. As mentioned, to achieve the following, a proposed ensemble stacking model approach will be applied to provide a simplified framework which can be applied by practitioners who are non-experts in the field of analytics. The model and framework allows for better understanding and interpretation which bridges the gap Moreover, there has been a growing concern regarding the interpretability of predictive models which will be taken into consideration in this research, which is rarely addressed in data mining prediction studies. It can be deemed as a problem of general interest within the field of analytics as well. By using an ensemble stacking model approach, it provides advantages such as simplicity; increases robustness; improved performance; and capability of a combined model induced by various models. This shows that the interpretability in previous research and literatures in general requires a subject matter expert to be involved in the analysis

process and an individual without any knowledge in predictive analysis may not fully grasps the concept.

6.2. Research Outcome

Table 18: Research Outcome

Research Questions	Research Objectives	Research Outcomes
What is the usage pattern of employee healthcare claims and what are the factors contributing to a high risk/cost employee (patient)?	To discover and understand the usage pattern of healthcare claims to better understand high-risk/cost employees (patients).	Through the descriptive analysis performed, the usage pattern of healthcare claims and the potential to reduce medical expenditure were explored by providing a flexible medical coverage selection for each employee. This also allows employers to understand the overall employee health population and the potential strategies which can be performed to improve overall health populations. Moreover, the predictive analysis provides an opportunity to identify the factors which contributes to high-risk/cost employees and allows employers to identify these signs and perform early intervention. The 3 predictors include ICD Category, Total Remaining Amt and Total Amt Insured. With these 3 predictors, it would be able to segregate between the low and high-risk individuals. By applying the stacking ensemble model approach and through the literature, which was extracted, the question on how stacking ensemble is a better option as compared to bagging and boosting techniques which is commonly

		used and the advantages of a stacking model approach was answered.
How does an ensemble stacking model approach compare against the existing bagging / boosting techniques applied in existing literatures?	To propose an ensemble stacking model approach as it provides advantages such as simplicity; improved performance; and capability of a combined model induced by various models over bagging and boosting techniques.	As shown through the predictive analysis, the ensemble stacking model does outperform the other single predictive models in terms of predictive performance, accuracy, and simplicity. The predictive accuracy was increased without compromising on the complexity of the model. Furthermore, this was tested on 2 criteria to address the robustness issue, where it was tested using proprietary / open-sourced
Will the proposed ensemble stacking model approach increase predictive accuracy as compared to a single predictive model?	Ensemble stacking model approach would potentially increase the predictive accuracy and can be used by practitioners who are non-experts in the field of analytics while being a more robust model which can be applied across a wide range of classification applications.	platforms and on 3 separate case studies. Overall, the proposed ensemble stacking model was tested for the effectiveness in 4 case studies including healthcare. It also showed the robustness of an ensemble predictive model as it can be applied across various classification scenarios without any concern on predictive performance.

6.3. Limitations and Future Work

With the increase and advancement in the development of health management recommendations, the research presented can be extended accordingly, some extensions are as suggested below:

- The dataset provided in this research were not ideal to be used for prediction as further explorations are required to gather medical health conditions as will be further mentioned in point 2. Hence, the results of this research were obtained but more accurate and enhanced results can be achieved in future works.
- Research did not involve patient medical health conditions such as BMI, Blood Pressure, and other medical information. This was not involved because of the paperwork required and PDPA (Privacy Data Protection Act) which would involve the medical department. Furthermore, the research ethics committee were advising against using such data as it is tedious and could lead to unwanted disputes. However, looking at other research which have been performed in the past in collaboration with the medical department, with medical health conditions data included, it could potentially have more influence and impact the prediction. With that, it may improve the predictive model if the potential of including health conditions into the analysis can be explored.
- Further exploration of implementing an enhanced new algorithm can be explored using the ensemble method which can be used across various domain, to produce more innovative outcomes.
- Further exploration to collaborate with insurance agencies can be considered to explore areas of research and how to fully optimize medical expenditure for organizations. The identification performed in this research was based on the utilization of AmtInsured by patients. Other factors influencing utilization and usage of medical coverage can be

explored to enhance medical coverages for organizations which may be tailored to their needs and employee population health.

- Further exploration to compare the results between before and after COVID19 may lead to more insights to be discovered.
- To explore on the possibility of model fusion and to develop a new classification algorithms based on the ensemble model approach. The concept itself has tremendous exploration which can be done, and it can be applied to even more applications such as fraud detection, real-time threat detection and even cyber-attacks.

6.4. Recommendations

1. Upper Respiratory Infection

- Group HR could install air purification systems in every department to ensure the air within the department would be purified as most employees spend most of their time in the office spaces. So, there is a need to have clean and fresh air.
- Group HR could potentially provide the necessary vaccination to the specific group of target segment which has highest volume of medical claims within the business unit.
- With the specified recommendation, the Group HR could monitor the changes within the next 3 months to observe if there are any changes within the claim pattern.

2. Low Back Pain

- Group HR could start by targeting the business units with the highest medical claims and try to observe the day-to-day operations within the business units to better understand why the business units are experience such an issue.
- Group HR could provide “Back Pain Relief Lumbar Support Cushion Pillow” to help employees with their posture and comfort levels. As consistently sitting on a chair without proper back support could affect the lower back.

- Group HR could encourage employees to do simple exercises such as simple stretching to help loosen the muscles as long sitting hours without any movement could affect the lower back as well.

3. Hypertension

- Group HR could start by targeting the business units with the highest medical claims and try to observe the day-to-day operations within the business units to better understand why the business units are experience such an issue.

Group HR could provide simple exercises and stress relief techniques to help employees relax during their day-to-day operations.

References

- (MAMPU), T. M. (2020). *Big Data Analytics Digital Government Lab (BDA-DGL)*. Retrieved August 24, 2021, from <https://www.malaysia.gov.my/portal/content/30616>
- A, A., NFA, H., H, J., D. M., S, O., & Kamarul, T. (2020). Prediction of Disease Burden and Healthcare Resource Utilization through Simple Predictive Analytics using Mathematical Approaches, an Experience from University of Malaya Medical Centre. *Journal of Health and Translational Medicine*, 1(1), 10-15.
- Abdunabi, T. (2016). A Framework for Ensemble Predictive Modeling. Ontario: University of Waterloo in Electrical and Computer Engineering.
- Abuassba, A. O., Zhang, D., Luo, X., Shaheryar, A., & Ali, H. (2017). Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines. *Hindawi (Computational Intelligence and Neuroscience)*, 1-12.
- Agarwal, R. (2019, July 28). *The 5 Feature Selection Algorithms Every Data Scientist Should Know*. (Towards Data Science) Retrieved April 30, 2020, from <https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2>
- Alduayj, S. S., & Rajpoot, K. (2018). Predicting Employee Attrition using Machine Learning. 2018 *International Conference on Innovations in Information Technology (IIT)*, 93-98.
- Alharthi, H. (2018). Healthcare predictive analytics: An overview with a focus on Saudi Arabia. *Journal of Infection and Public Health*, 11(1), 749-756.
- Alharti, H. (2018). Healthcare Predictive Analytics: An Overview with a Focus on Saudi Arabia. *Journal of Infection and Public Health*, 11(1), 749-756.
- Alonso, S. G., Díez, I. d., Rodrigues, J. J., Hamrioui, S., & López-Coronado, M. (2017). A Systematic Review of Techniques and Sources of Big Data in the Healthcare Sector. *Journal of Medical Systems*, 41(1), 1-9.
- Annamalai, N., Azid, I. A., Kamaruddin, S., & Yeoh, T. (2013). Importance of Problem Statement in Solving Industry Problems. *Applied Mechanics and Materials*, 421(1), 857-863.
- Asif, S., Zhao, M., Tang, F., & Zhu, Y. (2024). LWSE: a lightweight stacked ensemble model for accurate detection of multiple chest infectious diseases including COVID-19. *Multimedia Tools and Applications*, 23967–24003.
- Azmi, N. A., Noor, N. M., Shukri, M. I., & Aidalina Mahmud, R. A. (2022). The Role of Big Data Analytics in Digital Health for COVID-19 Prevention and Control in Asia. *Malaysian Journal of Medicine and Health Sciences*, 173-181.
- Bates, D. W., Sari, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data In Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients. *Health Affairs*, 33(7), 1123-1131.
- Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs*, 33(7), 1123-1131.

- Beh, B. (2019, February 22). *Avoid Overdependence on Employee Health Benefits*. (Focus Malaysia) Retrieved February 26, 2019, from <http://webcache.googleusercontent.com/search?q=cache:bJtd5GUnvJIJ:www.focusmalaysia.my/Income/avoid-overdependence-on-employee-health-benefits+&cd=1&hl=en&ct=clnk&gl=my>
- Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., J. Christian Kryder, R. P., Vempala, S., & Wang, G. (2008). Algorithmic Prediction of Health-Care Costs. *Operations Research*, 56(6), 1382-1392.
- Better Explained. (n.d.). *Understanding the Pareto Principle (The 80/20 Rule)*. (Better Explained) Retrieved March 30, 2020, from <https://betterexplained.com/articles/understanding-the-pareto-principle-the-8020-rule/>
- Birruntha, S. (2024, January 19). "Private healthcare sector to continue to grow in 2024". Retrieved from New Straits Times - Business Times: <https://www.nst.com.my/business/corporate/2024/01/1003070/private-healthcare-sector-continue-grow-2024>
- Brar, K. (2018, January 31). *Aiding Healthcare through Data Analytics*. (The Star Malaysia) Retrieved October 30, 2019, from <https://www.star2.com/health/2018/01/31/aiding-healthcare-data-analytics/>
- Brownlee, J. (2019, November 27). *How to Choose a Feature Selection Method for Machine Learning*. (Machine Learning Mastery) Retrieved April 30, 2020, from <https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>
- Bruno, G., Cerquitelli, T., Chiusano, S., & Xiao, X. (2014). A Clustering-Based Approach to Analyse Examinations for Diabetic Patients. *IEEE International Conference on Healthcare Informatics*. Verona.
- Chandrashekar, G., & Sahin, F. (2014). A Survey on Feature Selection Methods. *Computers and Electrical Engineering*, 40(1), 16-28.
- Chapple, M. (2018, November 8). *Defining the Regression Statistical Model*. (Lifewire) Retrieved March 21, 2019, from <https://www.lifewire.com/regression-1019655>
- Charan, G. (2017, October 19). *Stacking - A Super Learning Technique*. (Medium) Retrieved November 10, 2019, from https://medium.com/@gurucharan_33981/stacking-a-super-learning-technique-dbed06b1156d
- Charfaoui, Y. (2020, January 7). *Hands-On with Feature Selection Techniques: An Introduction*. (Heartbeat) Retrieved April 30, 2020, from <https://heartbeat.fritz.ai/hands-on-with-feature-selection-techniques-an-introduction-1d8dc6d86c16>
- Chen, Y., Zhang, J., & Ng, W. W. (2018). Loan Default Prediction Using Diversified Sensitivity Undersampling. *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, 240-245.
- Choudhury, A. (2019, April 1). *What are Feature Selection Techniques in Machine Learning?* (Analytics India Magazine) Retrieved April 30, 2020, from <https://analyticsindiamag.com/what-are-feature-selection-techniques-in-machine-learning/>
- Chua, A. P. (n.d.). *Clustering Analysis Concepts*. Subang Jaya: Sunway University.

- Chua, A. P. (n.d.). *Decision Tree Concepts*. Subang Jaya: Sunway University.
- Chye, K. H., & Tan, G. (2011). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, 19(2), 64-72.
- Cohn, C. (2015, February 6). *Steps To Identify Your Target Market*. (Forbes) Retrieved August 26, 2021, from <https://www.forbes.com/sites/chuckcohn/2015/02/06/steps-to-identify-your-target-market/?sh=12983b3c229d>
- DeFilippi, R. R. (2018, August 5). *Boosting, Bagging, and Stacking - Ensemble Methods with sklearn and mlens*. (Medium) Retrieved November 11, 2019, from <https://medium.com/@rrfd/boosting-bagging-and-stacking-ensemble-methods-with-sklearn-and-mlens-a455c0c982de>
- Deshmukh, P. M., & Gulhane, P. R. (2016). Importance of Clustering in Data Mining. *International Journal of Scientific & Engineering Research*, 7(2), 247-251.
- Dzeroski, S., & Zenko, B. (2004). Is Combining Classifiers with Stacking Better Than Selecting the Best One? *Kluwer Academic Publishers, Machine Learning*, 54(1), 255-273.
- Eapen, A. G. (2004). *Application of Data Mining in Medical Applications*. Ontario: University of Waterloo.
- EdgeProp MY. (2018, August 8). *The Return of the Haze*. (Edge Prop) Retrieved March 20, 2020, from <https://www.edgeprop.my/content/1412305/return-haze>
- Emmanuel, I., & Stanier, D. C. (2016). Defining Big Data. *Big Data and Advanced Wireless Technologies*, 1-6.
- Everetti, B., & Zajacova, A. (2015). Gender Differences in Hypertension and Hypertension Awareness Among Young Adults. *Biodemography Social Biology*, 61(1), 1-17.
- Fatt, Q. K., & Ramadas, A. (2018). The Usefulness and Challenges of Big Data in Healthcare. *Journal of Healthcare Communications*, 3(2), 1-4.
- Feng, G., & Fan, M. (2024). Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization. *Expert Systems with Applications*.
- Gerrard, N. (2018, November 19). *Construction is Third Most Stressful Industry*. (CIOB (The Chartered Institute of Building)) Retrieved March 20, 2020, from <https://www.constructionmanagemagazine.com/news/construction-third-most-stressful-industry/>
- Ghaleb, E. A., Dominic, P. D., Singh, N. S., & Naji, G. M. (2023). Assessing the Big Data Adoption Readiness Role in Healthcare between Technology Impact Factors and Intention to Adopt Big Data. *Sustainability (Multidisciplinary Digital Publishing Institute)*, 1-25.
- Gillis, E. E., & Sullivan, J. C. (2016). Sex Differences in Hypertension: Recent Advances. *Hypertension*, 68(6), 1322-1327.
- Google Developers. (2020, February 10). *What is Clustering?* Retrieved March 26, 2020, from <https://developers.google.com/machine-learning/clustering/overview>

- Gore, A. (2012). The Digital Earth: Understanding Our Planet in the 21st Century. *The Australian Surveyor*, 43(2), 89-91.
- Gunasekar, T., & Kayalvizhi, S. (2019). Big Data Analytics for Improved Care Delivery in the Healthcare Industry. *International Journal of Online and Biomedical Engineering (iJOE)*, 15(10), 40-51.
- Guo, C., & Chen, J. (2023). Big Data Analytics in Healthcare. In Y. Nakamori, *Knowledge Technology and Systems: Toward Establishing Knowledge Systems Science* (pp. 27-70). Singapore: Springer Nature Singapore.
- Gupta, P. (2017, May 17). *Decision Trees in Machine Learning*. (Towards Data Science) Retrieved March 13, 2019, from <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
- Halibas, A. S., Matthew, A. C., Pillai, I. G., Reazol, J. H., Delvo, E. G., & Reazol, L. B. (2019). Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling. *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, 1-7.
- Hoque, N., Singh, M., & Bhattacharyya, D. K. (2018). EFS-MI: An Ensemble Feature Selection Method for Classification. *Complex & Intelligent Systems*, 4(2), 105-118.
- Hu, H., Li, J., Wang, H., & Daggard, G. (2008). Robustness Analysis of Diversified Ensemble Decision Tree Algorithms for Microarray Data Classification. *2008 International Conference on Machine Learning and Cybernetics (IEEE)*, 1(1), 115-120.
- Ibeh, C. V., Elufioye, O. A., OlorunsogO, T., Asuzu, O. F., Nduubuisi, N. L., & Daraojimba, A. I. (2024). Data analytics in healthcare: A review of patient-centric approaches and healthcare delivery. *World Journal of Advanced Research and Reviews*, 1750-1760.
- Institute of Medicine (US) Committee. (2002). Employer Interest in Promoting the Health of Employees: A Rationale for Corporate Investment in Health. In *The Future of the Public's Health in the 21st Century*. Washington (DC): National Academies Press (US).
- Jain, R. (2015). *Predictive Modeling for Chronic Conditions*. Florida: Florida Atlantic University.
- Jain, R. (2015). *Predictive Modeling for Chronic Conditions*. Florida: Florida Atlantic University.
- Jeremiah Olawumi Arowoogun 1, *. O., Chidi, R., Adeniyi, A. O., & Okolo, C. A. (2024). A comprehensive review of data analytics in healthcare management: Leveraging big data for decision-making. *World Journal of Advanced Research and Reviews*, 1810-1821.
- Jovic, A., Brkic, K., & Bogunovic, N. (2015). A Review of Feature Selection Methods with Applications. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 38(1), 1200-1205.
- Juhi. (2018, February 26). *Simple Guide for Ensemble Learning Methods*. (Towards Data Science) Retrieved July 19, 2019, from <https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2>
- Kai, L. Y., Hsinchun, C., Brown, R. A., Hsing, L. S., & Jen, H. (2014). Healthcare Analytics and Clinical Intelligence: A Risk Predictive Framework for Chronic Care. *24th Annual Workshop on*

- Information Technologies and Systems: Value Creation from Innovative Technologies*, 24, pp. 1-54. Auckland.
- Kar, E. (2015, May 15). *The Evolution of Big Data and its implications*. (Happiest Minds Blogs) Retrieved October 9, 2017, from <http://www.happiestminds.com/blogs/the-evolution-of-big-data-and-its-implications/>
- Karvana, K. G., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. *2019 International Workshop on Big Data and Information Security (IWBIS)*, 33-38.
- Kaushik, S. (2016, November 3). *An Introduction to Clustering and different methods of Clustering*. (Analytics Vidhya) Retrieved March 13, 2019, from <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
- Kaushik, S. (2016, December 1). *Introduction to Feature Selection Methods with an Example (or How to Select the Right Variables?)*. (Analytics Vidhya) Retrieved May 4, 2020, from <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>
- Kawi, M. R. (2018, July 20). *Is the Haze Making a Return this Year*. (New Straits Times) Retrieved March 20, 2020, from <https://www.nst.com.my/news/nation/2018/07/392565/haze-making-return-year>
- Khoo, D. (2024, April 27). *Medical insurance premiums on the rise*. Retrieved from The Star: <https://www.thestar.com.my/business/business-news/2024/04/27/medical-insurance-premiums-on-the-rise>
- Kincade, K. (1998). Data Mining: Digging for Healthcare Gold. *Insurance & Technology (Business Premium Collection)*, 23(2), 1-7.
- Koller, D., Schön, G., Schäfer, I., Glaeske, G., van den Bussche, H., & Hansen, H. (2014). Multimorbidity and Long-Term Care Dependency - A Five-Year Follow-Up. *BMC Geriatrics*, 14(70), 1-9.
- Krishnan, S., Magalingam, P., & Ibrahim, R. b. (2018). Review on Data Analytics Framework in Heart Disease. *Open International Journal of Informatics (OIJI)*, 6(4), 42-53.
- Kruse, K. (2016, March 7). *The 80/20 Rule And How It Can Change Your Life*. (Forbes) Retrieved March 30, 2020, from <https://www.forbes.com/sites/kevinkruse/2016/03/07/80-20-rule/#250ae48b3814>
- Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms, 2nd Edition*. Bangor: John Wiley & Sons.
- Lai, P. K., Mai, C. W., Sulaiman, L. H., & Lim, P. K. (2019). Healthcare Big Data Analytics: Re-Engineering Healthcare Delivery through Innovation. *IRDI Public Health Policy Dialogue Series No.3*, 13(3), 10-13.
- Li, Y., & Chen, W. (2021). Entropy method of constructing a combined model for improving loan default prediction: A case study in China. *Journal of the Operational Research Society*, 72(5), 1099-1109.

- Li, Y., Bai, C., & Reddy, C. K. (2016). A Distributed Ensemble Approach for Mining Healthcare Data under Privacy Constraints. *Information Sciences (Elsevier)*, 330, 245-259.
- Liu, B., Li, Y., Ghosh, S., Sun, Z., Ng, K., & Hu, J. (2020). Complication Risk Profiling in Diabetes Care: A Bayesian Multi-Task and Feature Relationship Learning Approach. *IEEE Transactions on Knowledge and Data Engineering*, 32(7), 1276-1289.
- Malaysian Reserve. (2017, April 3). *Reining in Medical Benefits Cost*. (Malaysian Reserve) Retrieved February 26, 2019, from <https://themalaysianreserve.com/2017/04/03/reining-in-medical-benefits-cost/>
- Mardhiah, A. (2023, November 6). *Rising medical costs in Malaysia: A complex challenge*. Retrieved from The Malaysian Reserve: <https://themalaysianreserve.com/2023/11/06/rising-medical-costs-in-malaysia-a-complex-challenge/>
- Marjudi, S., Setik, R., Ahmad, R. M., Harun, W. A., & Ismail, S. (2020). Cardiovascular Disease Risk Factors among White-Collar Workers towards Healthy Communities in Malaysia. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(6), 2118-2124.
- Mehta, N., & Pandit, A. (2018). Concurrence of Big Data Analytics and Healthcare: A Systematic Review. *International Journal of Medical Informatics*, 114(1), 57-65.
- Menahem, E., Rokach, L., & Elovici, Y. (2009). Troika - An Improved Stacking Schema for Classification Tasks. *Elsevier, Information Sciences*, 179(24), 4097-4122.
- Merican, T. S. (2018, May 13). *Health Sector Challenges*. (The Sun Daily) Retrieved October 30, 2019, from <https://www.thesundaily.my/archive/health-sector-challenges-EUARCH547270>
- Mohd-Tahir, N.-A. (2015). Quality Use of Medicine in a Developing Economy: Measures to Overcome Challenges in the Malaysian Healthcare System. *SAGE Open Medicine*, 1-8.
- Moreira, L. B., & Namen, A. (2018). A Hybrid Data Mining Model for Diagnosis of Patients with Clinical Suspicion of Dementia. *Computer Methods and Programs in Biomedicine (Elsevier)*, 165, 139-149.
- Moturu, S. T., Johnson, W. G., & Liu, H. (2007). Predicting Future High-Cost Patients: A Real World Risk Modeling Application. *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*, 202-208.
- Moturu, S. T., Johnson, W. G., & Liu, H. (2007). Predicting Future High-Cost Patients: A Real-World Risk Modeling Application. *2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007)*, 202-208.
- Mozes, A. (2015, July 28). *Standing All Day at Work May Take Toll on Health*. (WebMD) Retrieved March 24, 2020, from <https://www.webmd.com/back-pain/news/20150728/standing-all-day-at-work-it-may-take-toll-on-health#1>
- Mwadulo, M. W. (2016). A Review on Feature Selection Methods for Classification Tasks. *International Journal of Computer Applications Technology and Research*, 5(6), 395-402.
- Narayanan, M. (2014, December 19). *Ensemble Methods in Predictive Analytics*. (LinkedIn) Retrieved July 19, 2019, from <https://www.linkedin.com/pulse/ensemble-methods-predictive-madhusudanan-n/>

- Obenshain, M. K. (2004). Application of Data Mining Techniques to Healthcare Data. *Infection Control and Hospital Epidemiology*, 25(8), 690-695.
- Olofsson, N. (2017). *A Machine Learning Ensemble Approach to Churn Prediction - Developing and Comparing Local Explanation Models on Top of a Black-Box Classifier*. Stockholm: KTH Royal Institute of Technology in Stockholm.
- One Stop Malaysia. (2018, January 1). *Malaysia School Holidays & Malaysia Public Holidays 2018*. (One Stop Malaysia) Retrieved March 24, 2020, from <https://www.onestopmalaysia.com/holidays-2018.html>
- Pham, H. N., Chatterjee, A., Narasimhan, B., Lee, C. W., Jha, D. K., Wong, E. Y., . . . Chua, M. C. (2019). Predicting Hospital Readmission Patterns of Diabetic Patients using Ensemble Model and Cluster Analysis. *2019 International Conference on System Science and Engineering (ICSSE)*. Dong Hoi.
- Prabhakaran, S. (2018, June 7). *Feature Selection - Ten Effective Techniques with Examples*. (Machine Learning Plus) Retrieved April 30, 2020, from <https://www.machinelearningplus.com/machine-learning/feature-selection/>
- Raghupathi, W., & Raghupathi, V. (2014). Big Data Analytics in Healthcare: Promise and Potential. *Health Information Science and Systems*, 2(3), 1-10.
- Rahm, E. (2016). Big Data Analytics. *IT - Information Technology*, 4(58), 155-156.
- Ranawana, R., & Palade, V. (2006). Multi-classifier systems: Review and a roadmap. *Int. J. Hybrid Intell. Syst.*, 3(1), 35-61.
- Rathi, M. (2010). Regression Modeling Technique on Data Mining for Prediction of CRM. *International Conference on Advances in Information and Communication Technologies*. Kochi.
- Raul, A., Patil, A., Raheja, P., & Sawant, R. (2016). Knowledge Discovery, Analysis And Prediction in Healthcare using Data Mining and Analytics. *International Conference on Next Generation Computing Technologies*, 475-478.
- Ravanshad, A. (2018, April 28). *Ensemble Methods*. (A Medium Corporation) Retrieved July 19, 2019, from <https://medium.com/@aravanshad/ensemble-methods-95533944783f>
- Rawale, S. (2018, August 1). *Feature Selection Methods in Machine Learning*. (A Medium Corporation) Retrieved April 30, 2020, from <https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc>
- Ray, S. (2015, August 14). *7 Types of Regression Techniques you should know!* (Analytics Vidhya) Retrieved March 21, 2019, from <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>
- Rocca, J. (2019, April 23). *Ensemble Methods: Bagging, Boosting and Stacking*. (Towards Data Science) Retrieved November 8, 2019, from <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205>
- Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, 53(12), 4046-4072.

- Rose, D. (2021, July 13). *Asking the Right Data Science Questions*. (LinkedIn) Retrieved August 26, 2021, from <https://www.linkedin.com/pulse/asking-right-data-science-questions-doug-rose/>
- Rouse, M. (2015, January). *Ensemble Modeling*. (Search Business Analytics) Retrieved July 19, 2019, from <https://searchbusinessanalytics.techtarget.com/definition/Ensemble-modeling>
- Sahoo, P. K., Mohapatra, S. K., & Wu, S.-L. (2016). Analyzing Healthcare Big Data With Prediction for Future Health Condition. *IEEE Access*, 4, 9786-9799.
- Sanjeevi, M. (2017, October 6). *Chapter 4: Decision Trees Algorithms*. (Medium Corporation) Retrieved March 13, 2019, from <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>
- Sharkey, A. J. (2012). *Combining artificial neural nets: ensemble and modular multi-net systems*. Sheffield: Springer Science & Business Media.
- Singh, A. (2018, June 18). *A Comprehensive Guide to Ensemble Learning (with Python codes)*. (Analytics Vidhya) Retrieved November 7, 2019, from <https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/>
- Singh, R. K., Agrawal, S., Sahu, A., & Kazancoglu, Y. (2023). Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda. *The TQM Journal*, 262-291.
- Sippe, R. (2015). *The Relation between Big Data and Informational Privacy in the Context of the Healthcare*. Delft: Delft University of Technology.
- Spradlin, D. (2012, September 1). *Are You Solving the Right Problem?* (Harvard Business Review) Retrieved August 26, 2021, from <https://hbr.org/2012/09/are-you-solving-the-right-problem>
- Srivathsan, A., Abdou, A., Al-Khatib, T., Apadinuwe, S.-C., Badiane, M. D., Bucumi, V., . . . Kanyi, S. K. (2024). District-Level Forecast of Achieving Trachoma Elimination as a Public Health Problem By 2030: An Ensemble Modelling Approach . *Clinical Infectious Diseases*, 101-107.
- Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. *Journal of the Society of Academic Emergency Medicine*, 18(10), 1099-1104.
- Streefkerk, R. (2018, June 20). *Primary and secondary sources*. (Scribbr) Retrieved August 26, 2021, from <https://www.scribbr.com/citing-sources/primary-and-secondary-sources/>
- Talukder, M. S., & Akter, S. (2024). An improved ensemble model of hyper parameter tuned ML algorithms for fetal health prediction. *International Journal of Information Technology* , 1831-1840.
- TechDifferences. (2018, January 13). *Difference Between Linear and Logistic Regression*. (TechDifferences) Retrieved March 21, 2019, from <https://techdifferences.com/difference-between-linear-and-logistic-regression.html>
- Tekieh, M. H. (2012). *Analysis of Healthcare Coverage using Data Mining Techniques*. Ontario: University of Ottawa.
- Tekieh, M. H. (2012). *Analysis of Healthcare Coverage using Data Mining Techniques*. Ontario: University of Ottawa.

- Tissot, F., & Stock, K. M. (2009). Studying the Relationship Between Low Back Pain and Working Postures Among Those Who Stand and Those Who Sit Most of the Working Day. *Ergonomics (Taylor & Francis Online)*, 52(11), 1402-1418.
- Tuysuzoglu, G., Birant, D., & Pala, A. (2017). Ensemble Methods in Environmental Data Mining. *Intechopen*, 2-17.
- Valiance Solutions. (2016, August 11). *Improving Predictions with Ensemble Model*. (Data Science Central) Retrieved July 19, 2019, from <https://www.datasciencecentral.com/profiles/blogs/improving-predictions-with-ensemble-model>
- Wales, U. o. (2021, April 23). *Primary and Secondary Sources*. Retrieved August 26, 2021, from <https://www.library.unsw.edu.au/study/information-resources/primary-and-secondary-sources>
- Wang, D. D., Quan, T. X., Khoo, A., Sridharan, S., Ramachandran, S., Ng, S. H.-X., & Rahman, S. N. (2017). Cluster Analysis on Utilization Patterns of Patients with Chronic Diseases Based on Flattened Electronic Medical Records. *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. Exeter.
- Wang, Y., Kung, L., & Byrd, T. A. (2018). Big Data Analytics: Understanding its Capabilities and Potential Benefits for Healthcare Organizations. *Technological Forecasting & Social Change*, 126(1), 3-13.
- Wunker, S. (2012, December 15). *Asking the Right Questions*. (Forbes) Retrieved August 26, 2021, from <https://www.forbes.com/sites/stephenwunker/2012/12/15/asking-the-right-question/?sh=77b5bb326930>
- Xiong, Z., Li, H., Liu, Z., Chen, Z., Zhou, H., Rong, W., & Ouyang, Y. (2024). A Review of Data Mining in Personalized Education: Current Trends and Future Prospects. *Cornell University*, 1-25.
- Yadav, S., Jain, A., & Singh, D. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. *2018 IEEE 8th International Advance Computing Conference (IACC)*, 349-354.
- Yuvaraj, N., & SriPreethaa, K. R. (2017). Diabetes Prediction in Healthcare Systems using Machine Learning Algorithms on Hadoop Cluster. *Cluster Computing*, 1(1), 1-9.
- Zhong, H., & Xiao, J. (2017). Enhancing Health Risk Prediction with Deep Learning on Big Data and Revised Fusion Node Paradigm. *Scientific Programming*, 1(1), 1-19.
- Zin, C. S., Rahman, N. S., Nazar, N. I., Kurdi, A., & Godman, B. (2023). Trends in the Cost of Medicines, Consultation Fees and Clinic Visits in Malaysia's Private Primary Healthcare System: Employer Health Insurance Coverage. *Journal of Multidisciplinary Healthcare*, 1683-1697.

Appendix

List of Publications and Papers Presented

- i. Chan, K. W., Lee, A. S. H., & Zainol, Z. (2020). Profiling patterns in healthcare system: A preliminary study. International Journal of Advanced Computer Science and Application, 11(4), 661-668*
- ii. Chan, K. W., Lee, A. S. H., & Zainol, Z. (2021). A Framework for Predicting Employee Health Risks using Ensemble Model. International Journal of Advanced and Applied Sciences, 8(9), 29-38*
- iii. Predicting Employee Health Risks using Classification Ensemble Model - presented at 5th International Conference on Information Retrieval and Knowledge Management - CAMP 21'*

Research Ethics Approval Letter



15th November 2019

Ref No: PGSUREC 2019/055

Nicholas Chan Khin Whai
Department of Computing and Information Systems
School of Science and Technology
Sunway University

Dear Nicholas (Student ID: 13025382),

Research Ethics Approval of "Profiling Patterns using Predictive Analysis in Healthcare Systems"

Sunway University Research Ethics Committee has reviewed the documents you have submitted on 14th October 2019 for your project and has given approval for your proposed study.

The final list of documents reviewed and approved by the Committee is as per below:

1. Cover Sheet

Please remember that you are required to inform the Committee of (1) any changes in your research procedures and (2) the date when the project is completed.

The Committee wishes you the best success for your project.

Yours sincerely,

A handwritten signature in black ink, appearing to read "Kenneth Feinstein".

Assoc. Prof. Kenneth Alan Feinstein
Chair,
Sunway University Research Ethics Committee

Sunway University (SU-025-B)
Sunway University Sdn Bhd (603074)
A member of the Sunway Education Group
No. 5, Jalan Universiti, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia
Tel: +6 03 7491 8622 Fax: +6 03 5635 8630 university.sunway.edu.my

Owned and governed by the
Jeffrey Cheah
Foundation
A small green logo of a plant with two leaves.
Nurturing the Seeds of Wisdom

Description of Datasets

Table 19: Data Understanding

1. GP_2016_2018	
Variable	Description
1. Corporate	Indicating employee corporation
2. StartDate	Medical insurance start date (every employee is the same depending on the year)
3. ExpiryDate	Medical insurance expiry date (every employee is the same depending on the year)
4. Entity Code	Indicating employee entity code identifier
5. Entity Name	Indicating employee entity name identifier
6. Business Industry	Indicating employee business industry
7. Staff Member ID (D)	Patient's unique identification number (Spouse/Child)
8. Staff ID	Employee's unique identification number
9. MC	Number of Medical Certificate (MC) in days
10. DTDISABILITY	Date of disability/date of visit
11. Category	Indicating category of claims
12. Emp Annual Limit (RM)	Medical insurance yearly limit for an employee (different employee levels get different limit)
13. Dep Annual Limit (RM)	Medical insurance yearly limit for a dependent (Spouse/Child) (different dependent get different limit with reference to the employee)
14. AmtIncurred	Amount spent/incurred (in Ringgit Malaysia, RM) during a visit to the clinic or hospital
15. AmtInsured	Amount insured (in Ringgit Malaysia, RM) by the medical insurance company during a visit to the clinic or hospital
16. ExcessPaid	Excess amount due based on the difference between (Amount Incurred – Amt Insured), in Ringgit Malaysia RM
17. TypeOfClaims	Indicating claim type of a patient
18. PatientGender	Indicating patient's gender
19. PatientAge	Indicating patient's age
20. Rel	Indicating patient's relationship (Employee, Spouse or Child)
21. MedicalProviders	Indicating the clinic or hospital which the patient had visited
22. Diagnosis	Patient's medical condition as diagnosed by the medical doctor
23. DischargeDate	Date of discharge of a patient if hospitalized/admitted

24. MCDays	Number of Medical Certificate (MC) in days
25. BranchName	Indicating the branch of an employee
26. DeptName	Indicating the department of an employee
27. LTM	Indicating if an employee is on Long Term Medication
28. DRName	Indicating the name of the doctor who attended to the patient
29. MiCaresClaimID	Patient's unique claim identification number used by the medical insurance company
30. ICDCode	ICD, International Statistical Classification of Diseases and Related Health Problem. ICD codes are alphanumeric indicators used by doctors, health insurance companies and health agencies to represent diagnoses. Every disease, disorder, injury, infection, and symptom has its own ICD code.
31. med_fee	Medical fee surcharge in Ringgit Malaysia, RM
32. xray_fee	X-ray surcharge in Ringgit Malaysia, RM
33. lab_fee	Laboratory surcharge in Ringgit Malaysia, RM
34. inject_fee	Injection surcharge in Ringgit Malaysia, RM
35. surg_fee	Surgeon surcharge in Ringgit Malaysia, RM
36. screen_fee	Screening surcharge in Ringgit Malaysia, RM
37. dressing_fee	Medical dressing surcharge in Ringgit Malaysia, RM
38. others_fee	Other extra miscellaneous surcharge in Ringgit Malaysia, RM
39. referral_fee	Referral surcharge in Ringgit Malaysia, RM

Table 20: SP - Data Understanding

2. SP_2016_2018	
Variable	Description
1. Entity Code	Indicating employee entity code identifier
2. Corporate	Indicating employee corporation
3. Entity	Indicating employee entity name identifier
4. CUR Industry	Indicating employee business industry
5. BranchName	Indicating the branch of an employee
6. DeptName	Indicating the department of an employee
7. Staff Member ID (D)	Patient's unique identification number (Spouse/Child)
8. Employee Identifier	Employee's unique identification number
9. PatientGender	Indicating patient's gender

10. PatientAge	Indicating patient's age
11. Rel	Indicating patient's relationship (Employee, Spouse or Child)
12. Category	Indicating category of claims
13. Emp Annual Limit	Medical insurance yearly limit for an employee (different employee levels get different limit)
14. Dep Annual Limit	Medical insurance yearly limit for a dependent (Spouse/Child) (different dependent get different limit with reference to the employee)
15. AmtIncurred	Amount spent/incurred (in Ringgit Malaysia, RM) during a specialist visit to the clinic or hospital
16. AmtInsured	Amount insured (in Ringgit Malaysia, RM) by the medical insurance company during a specialist visit to the clinic or hospital
17. ExcessPaid	Excess amount due based on the difference between (Amount Incurred – Amt Insured), in Ringgit Malaysia RM
18. StartDate	Medical insurance start date (every employee is the same depending on the year)
19. ExpiryDate	Medical insurance expiry date (every employee is the same depending on the year)
20. DTDISABILITY	Date of disability/date of visit to the specialist
21. TypeOfClaims	Indicating claim type of a patient
22. MC (Days)	Number of Medical Certificate (MC) in days
23. Diagnosis	Patient's medical condition as diagnosed by the medical doctor
24. MCDays	Number of Medical Certificate (MC) in days
25. LTM	Indicating if an employee is on Long Term Medication
26. MedicalProviders	Indicating the specialist clinic or hospital which the patient had visited
27. DischargeDate	Date of discharge of a patient if hospitalized/admitted
28. DRName	Indicating the name of the doctor who attended to the patient
29. MiCaresClaimID	Patient's unique claim identification number used by the medical insurance company
30. ICDCode	ICD, International Statistical Classification of Diseases and Related Health Problem. ICD codes are alphanumeric indicators used by doctors, health insurance companies and health agencies to represent diagnoses. Every disease, disorder, injury, infection, and symptom has its own ICD code.
31. EICCFee	Consultation, amount incurred during specialist visit
32. CPCCFee	Consultation, co-payment incurred during specialist visit
33. EPCCFee	Consultation, payment eligible incurred during specialist visit

34. ENPCCFee	Consultation, payment ineligible incurred during specialist visit
35. EIMCFee	Medication, amount incurred during specialist visit
36. CPMCFee	Medication, co-payment incurred during specialist visit
37. EPMCFee	Medication, payment eligible incurred during specialist visit
38. ENPMCFee	Medication, payment ineligible incurred during specialist visit
39. EIXSUFee	X-ray, amount incurred during specialist visit
40. CPXSUFee	X-ray, co-payment incurred during specialist visit
41. EPXSUFee	X-ray, payment eligible incurred during specialist visit
42. ENPXSUFee	X-ray, payment ineligible incurred during specialist visit
43. EILTFee	Laboratory, amount incurred during specialist visit
44. CPLTFee	Laboratory, co-payment incurred during specialist visit
45. EPLTFee	Laboratory, payment eligible incurred during specialist visit
46. ENPLTFee	Laboratory, payment ineligible incurred during specialist visit
47. EIProcFee	Procedure, amount incurred during specialist visit
48. CPProcFee	Procedure, co-payment incurred during specialist visit
49. EPProcFee	Procedure, payment eligible incurred during specialist visit
50. ENPProcFee	Procedure, payment ineligible incurred during specialist visit
51. EIPTFee	Physiotherapy, amount incurred during specialist visit
52. CPPTFee	Physiotherapy, co-payment incurred during specialist visit
53. EPPTFee	Physiotherapy, payment eligible incurred during specialist visit
54. ENPPTFee	Physiotherapy, payment ineligible incurred during specialist visit
55. EIAFFee	Admin Fee, amount incurred during specialist visit
56. CPAFFee	Admin Fee, co-payment incurred during specialist visit
57. EPAFFee	Admin Fee, payment eligible incurred during specialist visit
58. ENPAFFee	Admin Fee, payment ineligible incurred during specialist visit
59. EIADTFee	Accidental Dental Treatment, amount incurred during specialist visit
60. CPADTFee	Accidental Dental Treatment, co-payment incurred during specialist visit
61. EPADTFee	Accidental Dental Treatment, payment eligible incurred during specialist visit
62. ENPADTFee	Accidental Dental Treatment, payment ineligible incurred during specialist visit
63. EIOCTFee	Out-Patient Cancer Treatment, amount incurred during specialist visit

64. CPOCTFee	Out-Patient Cancer Treatment, co-payment incurred during specialist visit
65. EPOCTFee	Out-Patient Cancer Treatment, payment eligible incurred during specialist visit
66. ENPOCTFee	Out-Patient Cancer Treatment, payment ineligible incurred during specialist visit
67. EIOKDTFee	Out-Patient Kidney Dialysis Treatment, amount incurred during specialist visit
68. CPOKDTFee	Out-Patient Kidney Dialysis Treatment, co-payment incurred during specialist visit
69. EPOKDTFee	Out-Patient Kidney Dialysis Treatment, payment eligible incurred during specialist visit
70. ENPOKDTFee	Out-Patient Kidney Dialysis Treatment, payment ineligible incurred during specialist visit
71. EIOthFee	Others/Rounding Adjustment, amount incurred during specialist visit
72. CPOthFee	Others/Rounding Adjustment, co-payment incurred during specialist visit
73. EPOthFee	Others/Rounding Adjustment, payment eligible incurred during specialist visit
74. ENPOthFee	Others/Rounding Adjustment, payment ineligible incurred during specialist visit

Table 21: IP - Data Understanding

3. IP_2016_2018	
Variable	Description
1. Corporate	Indicating employee corporation
2. StartDate	Medical insurance start date (every employee is the same depending on the year)
3. ExpiryDate	Medical insurance expiry date (every employee is the same depending on the year)
4. Entity Code	Indicating employee entity code identifier
5. Entity	Indicating employee entity name identifier
6. Business Industry	Indicating employee business industry
7. Staff ID (D)	Patient's unique identification number (Employee, Spouse or Child)
8. DTDISABILITY	Date of disability/date of admission/hospitalization
9. AmtIncurred	Amount spent/incurred (in Ringgit Malaysia, RM) during an admission/hospitalization
10. Category	Indicating category of claims

11. AnnualLimitAmtPerDisability	Medical insurance yearly limit for a patient (Employee, Spouse or Child)
12. AmtInsured	Amount insured (in Ringgit Malaysia, RM) by the medical insurance company during an admission/hospitalization
13. ExcessPaid	Excess amount due based on the difference between (Amount Incurred – Amt Insured), in Ringgit Malaysia RM
14. TypeOfClaims	Indicating claim type of a patient
15. PatientGender	Indicating patient's gender
16. PatientAge	Indicating patient's age
17. Rel	Indicating patient's relationship (Employee, Spouse or Child)
18. MedicalProviders	Indicating the hospital which the patient had been admitted/hospitalized
19. Diagnosis	Patient's medical condition as diagnosed by the medical doctor
20. DischargeDate	Date of discharge of a patient if hospitalized/admitted
21. MCDays	Number of Medical Certificate (MC) in days
22. BranchName	Indicating the branch of an employee
23. DeptName	Indicating the department of an employee
24. LTM	Indicating if an employee is on Long Term Medication
25. DRName	Indicating the name of the doctor who attended to the patient
26. MiCaresClaimID	Patient's unique claim identification number used by the medical insurance company
27. ICDCode	ICD, International Statistical Classification of Diseases and Related Health Problem. ICD codes are alphanumeric indicators used by doctors, health insurance companies and health agencies to represent diagnoses. Every disease, disorder, injury, infection, and symptom has its own ICD code.
28. RBInc	RB, amount incurred during admission/hospitalization
29. RBCP	RB, co-payment incurred during admission to hospital
30. RBE	RB, payment eligible incurred during admission to hospital
31. RBI	RB, payment ineligible incurred during admission to hospital
32. ICUInc	ICU, amount incurred during admission to hospital
33. ICUCP	ICU, co-payment incurred during admission to hospital
34. ICUE	ICU, payment eligible incurred during admission to hospital
35. ICUI	ICU, payment ineligible incurred during admission to hospital

36. SFInc	Surgeon fee surcharge, amount incurred during admission to hospital
37. SFCP	Surgeon fee surcharge, co-payment incurred during admission to hospital
38. SFE	Surgeon fee surcharge, payment eligible incurred during admission to hospital
39. SFI	Surgeon fee surcharge, payment ineligible incurred during admission to hospital
40. AFInc	Anesthetist fee surcharge, amount incurred during admission to hospital
41. AFCP	Anesthetist fee surcharge, co-payment incurred during admission to hospital
42. AFE	Anesthetist fee surcharge, payment eligible incurred during admission to hospital
43. AFI	Anesthetist fee surcharge, payment ineligible incurred during admission to hospital
44. OTInc	Operating theatre surcharge, amount incurred during admission to hospital
45. OTCP	Operating theatre surcharge, co-payment incurred during admission to hospital
46. OTE	Operating theatre surcharge, payment eligible incurred during admission to hospital
47. OTI	Operating theatre surcharge, payment ineligible incurred during admission to hospital
48. HSSAIInc	HSS - nursing care / procedure, amount incurred during admission to hospital
49. HSSACP	HSS - nursing care / procedure, co-payment incurred during admission to hospital
50. HSSAE	HSS - nursing care / procedure, payment eligible incurred during admission to hospital
51. HSSAI	HSS - nursing care / procedure, payment ineligible incurred during admission to hospital
52. HSSBInc	HSS - medicine / pharmacy / injection, amount incurred during admission to hospital
53. HSSBCP	HSS - medicine / pharmacy / injection, co-payment incurred during admission to hospital
54. HSSBE	HSS - medicine / pharmacy / injection, payment eligible incurred during admission to hospital

55. HSSBI	HSS - medicine / pharmacy / injection, payment ineligible incurred during admission to hospital
56. HSSCInc	HSS - laboratory / diagnostic / x-ray, amount incurred during admission to hospital
57. HSSCCP	HSS - laboratory / diagnostic / x-ray, co-payment incurred during admission to hospital
58. HSSCE	HSS - laboratory / diagnostic / x-ray, payment eligible incurred during admission to hospital
59. HSSCI	HSS - laboratory / diagnostic / x-ray, payment ineligible incurred during admission to hospital
60. HSSDInc	HSS - therapy / physiotherapy, amount incurred during admission to hospital
61. HSSDCP	HSS - therapy / physiotherapy, co-payment incurred during admission to hospital
62. HSSDE	HSS - therapy / physiotherapy, payment eligible incurred during admission to hospital
63. HSSDI	HSS - therapy / physiotherapy, payment ineligible incurred during admission to hospital
64. HSSEInc	HSS - medical supplies, amount incurred during admission to hospital
65. HSSECP	HSS - medical supplies, co-payment incurred during admission to hospital
66. HSSEE	HSS - medical supplies, payment eligible incurred during admission to hospital
67. HSSEI	HSS - medical supplies, payment ineligible incurred during admission to hospital
68. HSSFInc	HSS - others, amount incurred during admission to hospital
69. HSSFCP	HSS - others, co-payment incurred during admission to hospital
70. HSSFE	HSS - others, payment eligible incurred during admission to hospital
71. HSSFI	HSS - others, payment ineligible incurred during admission to hospital
72. PVIInc	In-hospital physician visit surcharge, amount incurred during admission to hospital
73. PVCP	In-hospital physician visit surcharge, co-payment incurred during admission to hospital
74. PVE	In-hospital physician visit surcharge, payment eligible incurred during admission to hospital

75. PVI	In-hospital physician visit surcharge, payment ineligible incurred during admission to hospital
76. LFIInc	Lodger fee surcharge, amount incurred during admission to hospital
77. LFCP	Lodger fee surcharge, co-payment incurred during admission to hospital
78. LPE	Lodger fee surcharge, payment eligible incurred during admission to hospital
79. LFI	Lodger fee surcharge, payment ineligible incurred during admission to hospital
80. AMBInc	Ambulance fee surcharge, amount incurred during admission to hospital
81. AMBCP	Ambulance fee surcharge, co-payment incurred during admission to hospital
82. AMBE	Ambulance fee surcharge, payment eligible incurred during admission to hospital
83. AMBI	Ambulance fee surcharge, payment ineligible incurred during admission to hospital
84. MRInc	Medical report fee surcharge, amount incurred during admission to hospital
85. MRCP	Medical report fee surcharge, co-payment incurred during admission to hospital
86. MRE	Medical report fee surcharge, payment eligible incurred during admission to hospital
87. MRI	Medical report fee surcharge, payment ineligible incurred during admission to hospital
88. DPIInc	Daycare procedure surcharge, amount incurred during admission to hospital
89. DPCP	Daycare procedure surcharge, co-payment incurred during admission to hospital
90. DPE	Daycare procedure surcharge, payment eligible incurred during admission to hospital
91. DPI	Daycare procedure surcharge, payment ineligible incurred during admission to hospital
92. PHDTInc	Pre-hospital diagnostic surcharge, amount incurred during admission to hospital
93. PHDTCP	Pre-hospital diagnostic surcharge, co-payment incurred during admission to hospital

94. PHDTE	Pre-hospital diagnostic surcharge, payment eligible incurred during admission to hospital
95. PHDTI	Pre-hospital diagnostic surcharge, payment ineligible incurred during admission to hospital
96. PHSCInc	Pre-hospital specialist surcharge, amount incurred during admission to hospital
97. PHSCCP	Pre-hospital specialist surcharge, co-payment incurred during admission to hospital
98. PHSCE	Pre-hospital specialist surcharge, payment eligible incurred during admission to hospital
99. PHSCI	Pre-hospital specialist surcharge, payment ineligible incurred during admission to hospital
100. PHInc	Post hospitalization surcharge, amount incurred during admission to hospital
101. PHCP	Post hospitalization surcharge, co-payment incurred during admission to hospital
102. PHE	Post hospitalization surcharge, payment eligible incurred during admission to hospital
103. PHI	Post hospitalization surcharge, payment ineligible incurred during admission to hospital
104. AOPCInc	Annual out-patient cancer fee, amount incurred during admission to hospital
105. AOPCCP	Annual out-patient cancer fee, co-payment incurred during admission to hospital
106. AOPCE	Annual out-patient cancer fee, payment eligible incurred during admission to hospital
107. AOPCI	Annual out-patient cancer fee, payment ineligible incurred during admission to hospital
108. AOPKInc	Annual out-patient kidney fee, amount incurred during admission to hospital
109. AOPKCP	Annual out-patient kidney fee, co-payment incurred during admission to hospital
110. AOPKE	Annual out-patient kidney fee, payment eligible incurred during admission to hospital
111. AOPKI	Annual out-patient kidney fee, payment ineligible incurred during admission to hospital

112. UCInc	Uncovered charges incurred by patient; amount incurred during admission to hospital
113. UCCP	Uncovered charges incurred by patient; co-payment incurred during admission to hospital
114. UCE	Uncovered charges incurred by patient; payment eligible incurred during admission to hospital
115. UCI	Uncovered charges incurred by patient; payment ineligible incurred during admission to hospital
116. OInc	Other extra chargers, amount incurred during admission to hospital
117. OCP	Other extra chargers, co-payment incurred during admission to hospital
118. OE	Other extra chargers, payment eligible incurred during admission to hospital
119. OI	Other extra chargers, payment ineligible incurred during admission to hospital

Drill Down Analysis by Age Group (SP)

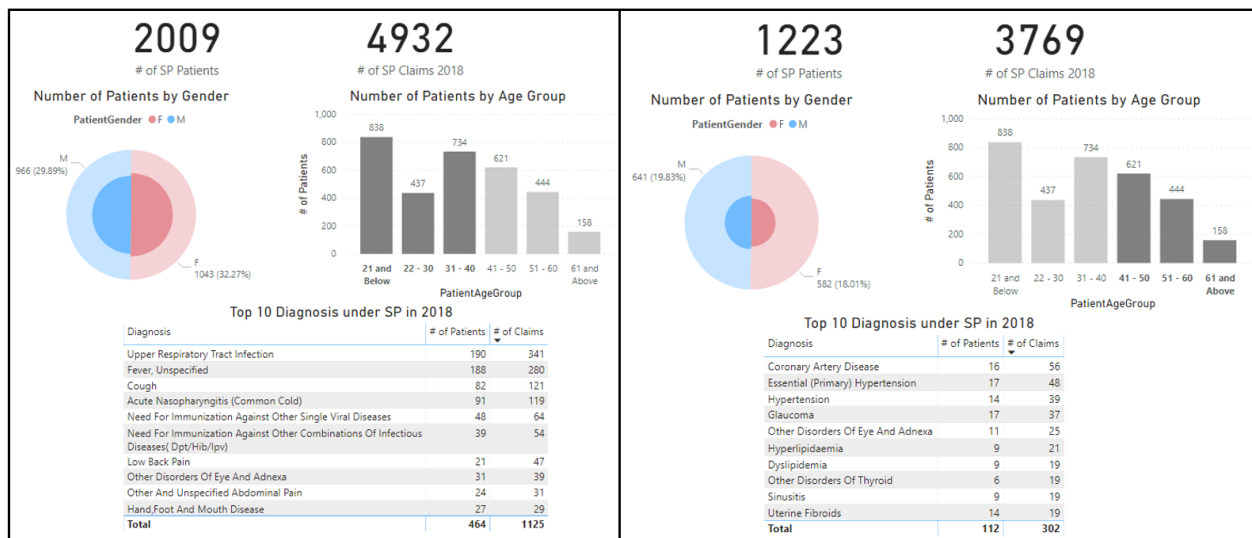


Figure 72: Comparison of Age Group between (Below 40) and (Above 40) - SP

Figure 93 onwards focuses on the comparison of SP (Specialist) claims based on age segment of Below 40 and Above 40. There were more patients among the Below 40 segment at 2,009 patients who made 4,932 claims in total, while Above 40 segment had 1,223 patients who made 3,769 claims in total. Looking at the top 10 diagnoses however, there is a major difference. Below 40 shows “Upper Respiratory Tract Infection”, “Fever” and “Cough” for the 3 most

common diagnosis but Above 40 shows “Coronary Artery Disease”, “Essential (Primary) Hypertension” and “Hypertension” for the 3 most common diagnosis.

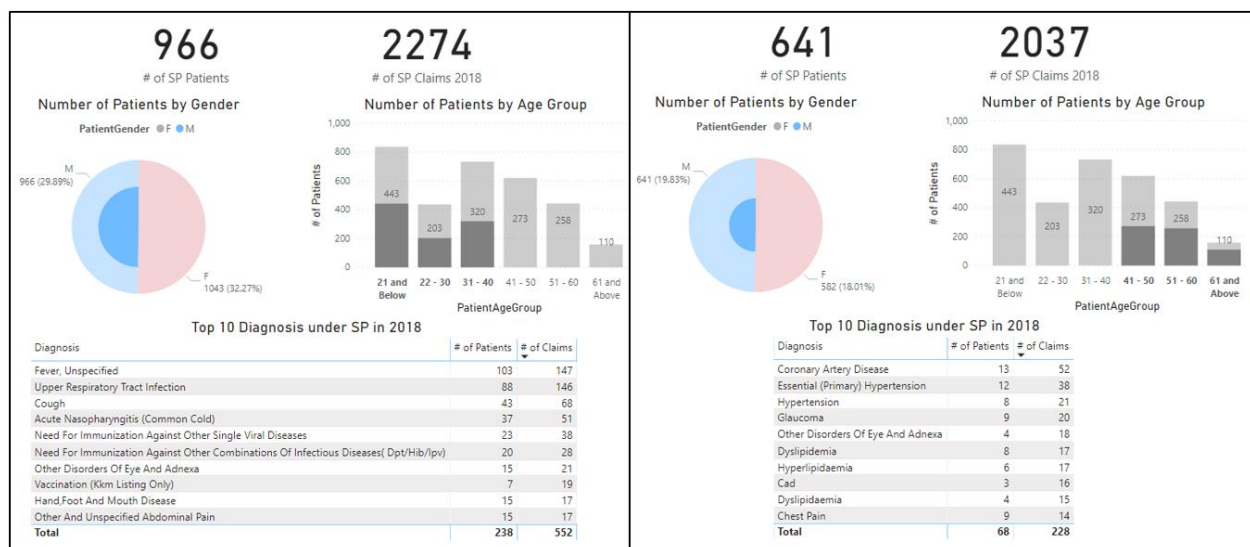


Figure 73: Comparison of (M) Age Group between (Below 40) and (Above 40) - SP

Figure 94 shows male patients who were Below 40 on the left and Above 40 on the right. There were a total of 966 male patients who were Below 40 and 641 male patients who were Above 40. Below 40 segment made a total of 2,274 claims while Above 40 made 2,037 claims. Looking at the diagnoses, it is very different between Below 40 and Above 40 based on the common 3 diagnoses, where Below 40 segment were commonly diagnosed with “Fever”, “Upper Respiratory Tract Infection” and “Cough”, Above 40 segment, however, were diagnosed with “Coronary Artery Disease”, “Essential (Primary) Hypertension” and “Hypertension”.

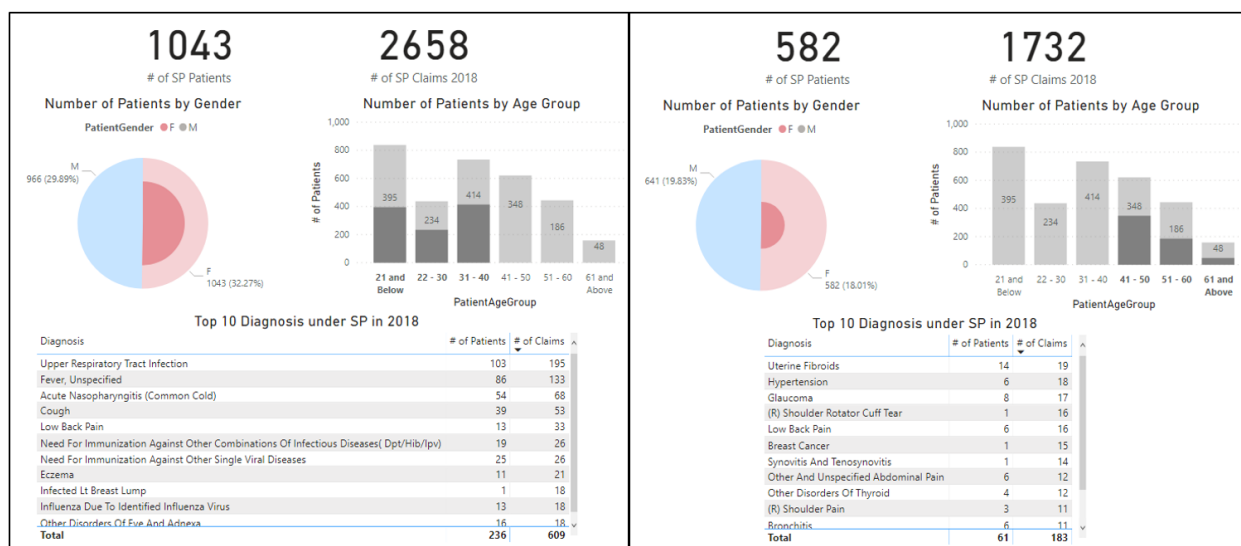


Figure 74: Comparison of (F) Age Group between (Below 40) and (Above 40) - SP

Figure 95 shows female patients who were Below 40 on the left and Above 40 on the right. Below 40 had 1,043 patients who made a total of 2,658 claims while Above 40 had 582 patients who made 1,732 claims. Similar to male patients, the diagnosis under female patients differs based on the common 3, where Below 40 patients were diagnosed with “Upper Respiratory Tract Infection”, “Fever” and “Acute Nasopharyngitis”, on the other hand, Above 40 were diagnosed with “Uterine Fibroids”, “Hypertension” and “Glaucoma”.

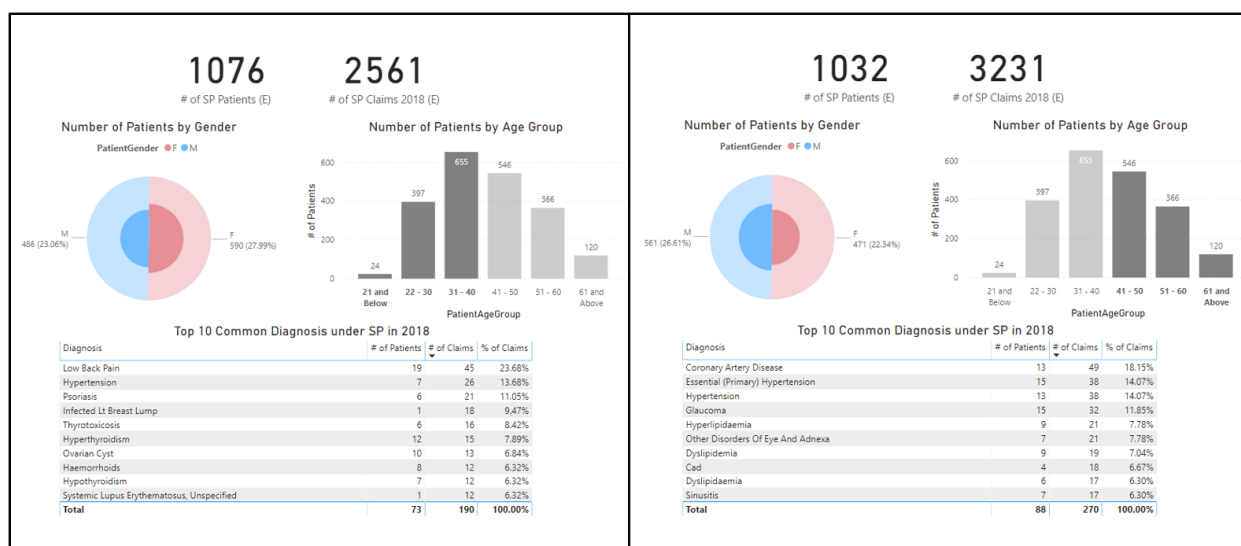


Figure 75: Comparison of Age Group between (Below 40) and (Above 40) - Employee; SP

Figure 96 focuses on SP (Specialist) claims but only those who are employees. Similarly, the segment were split into 2 groups, Below 40 and Above 40. Below 40 segment comprised of

1,076 patients who made 2,561 claims while Above 40 comprised of 1,032 patients who made 3,231 claims. The common diagnoses recorded among employees were “Low Back Pain”, “Hypertension” and “Psoriasis” for Below 40 while Above 40 were “Coronary Artery Disease”, “Essential (Primary) Hypertension” and “Hypertension”.

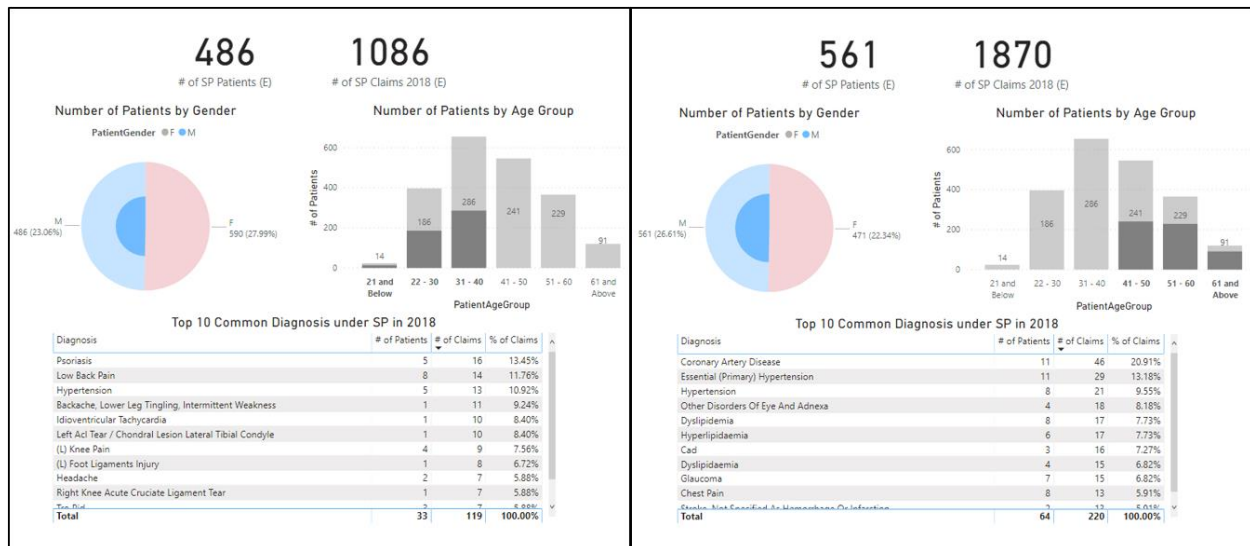


Figure 76: Comparison of (M) Age Group between (Below 40) and (Above 40) - Employee; SP

Figure 97 shows male patients (employees) who made claims under SP. Below 40 segment had 486 patients who made a total of 1,086 claims while Above 40 had 561 patients who made 1,870 claims. Common diagnoses which were recorded among male employees were “Psoriasis”, “Low Back Pain” and “Hypertension” for Below 40 while Above 40 were “Coronary Artery Disease”, “Essential (Primary) Hypertension” and “Hypertension”.

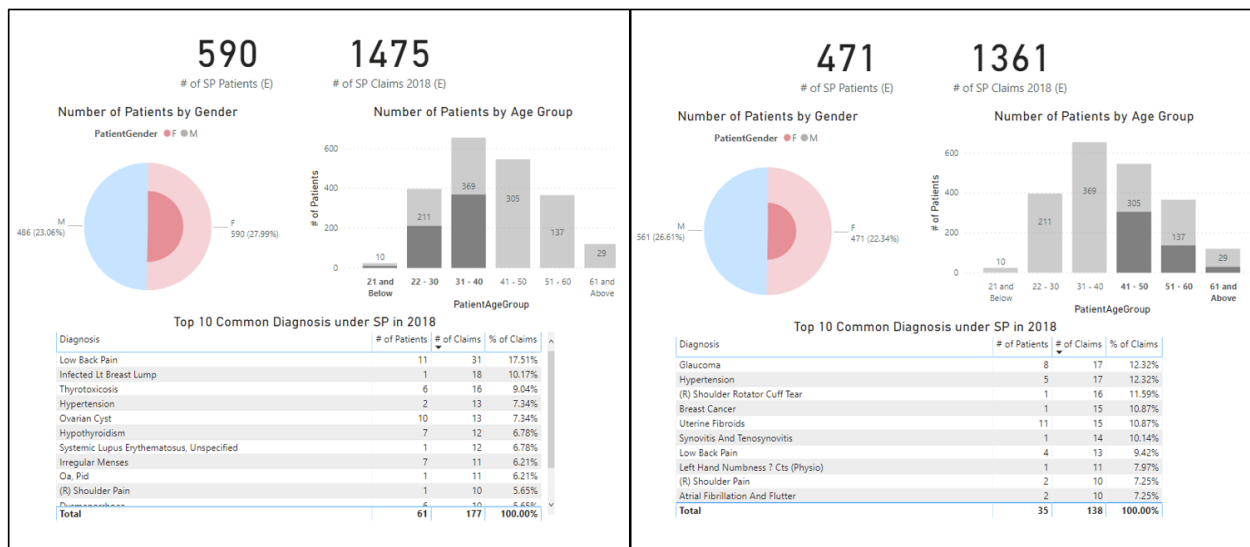


Figure 77: Comparison of (F) Age Group between (Below 40) and (Above 40) - Employee; SP

Figure 98 shows female patients (employees) who made claims under SP. Below 40 segment had 590 patients who made a total of 1,475 claims while Above 40 had 471 patients who made 1,361 claims. As shown at the bottom of the illustration, the common diagnoses recorded among female employees who were Below 40 include “Low Back Pain”, “Thyrototoxicosis” and “Hypertension” - “Infection Lt Breast Lump” was ignored because even though claim was higher, only 1 patient made it. Above 40 on the other hand, recorded “Glaucoma”, Hypertension” and “Uterine Fibroids”. Similarly, there were 2 diagnoses (“Shoulder Rotator Cuff Tear” and “Breast Cancer” by Above 40 segment which were ignored because the claims were made by only 1 patient.