Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

A framework for predicting employee health risks using ensemble model

Nicholas Khin-Whai Chan [1], Angela Siew-Hoong Lee [1,*], Zuraini Zainol [2]

[1]*Department of Computing and Information Systems, Sunway University, Sunway, Malaysia*
[2]*Department of Computer Science, Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, Malaysia*

ARTICLE INFO

Article history:
Received 16 December 2020
Received in revised form
20 April 2021
Accepted 10 June 2021

Keywords:
Data analytics
Predictive analysis
Ensemble modeling
Stacking
Framework

ABSTRACT

Through the phenomenon of data, big data and data analytics have provided an opportunity to collect, store, process, analyze and visualize an immense amount of information. Healthcare is recognized as one of the most information-intensive sectors. An urge to explore analytics has been sparked by the rapid growth of data within the healthcare sector. Most employers in Malaysia provide medical benefits that are included in the medical insurance plan for their employees. Data collected such as the history of medical claims are stored with the HR (Human Resource) which contributes to the potential of analyzing and recognizing trends within medical claims to better understand the use and overall health of the employee population. Patients with higher risk will generally convert into patients with high costs. Hence, early intervention of these patients will allow employers to potentially minimize costs and plan preventative steps. In predictive analysis, Decision Trees and Regression are typical techniques applied. The proposed framework combines an ensemble technique known as Stacking. As opposed to a single predictive model, an ensemble predictive model would yield better performance and accuracy. The objective of this paper is therefore to review current practices and past research within the healthcare sector while suggesting a practical framework for classification ensemble modeling. Preliminary findings indicated that an ensemble model can produce higher predictive accuracy and performance than a single model.

1. Introduction

Through the phenomenon of data, big data, and data analytics, has provided an opportunity to collect, store, process, analyze and visualize an immense amount of information (Gore, 2012). The future of digital information technology has undeniably been influenced by Big Data-behind all the data, the ability to capture, process, and analyze data to uncover meaningful insights and trends that could potentially affect business processes, business operations, generate business opportunities and, more importantly, minimize resources such as time and monetary (Rahm, 2016). The healthcare industry is recognized as one of the most information-intensive sectors (Eapen, 2004). An urge to explore analytics has been sparked by the rapid growth of data within the healthcare sector. Healthcare data is consistently increasing on a regular basis (Eapen, 2004)-which can be converted into information and knowledge. With health insurance data and information held with the HR (Human Resource), employers may theoretically analyze and identify trends in past medical claims that could then help them make specific decisions such as understand the overall health of the employee population and current usage of medical premium coverage. Patients at high risk will generally convert into patients with high costs. Hence, early intervention of these patients will allow employers to potentially minimize costs and plan preventative steps. A better understanding of the overall health of the employee population would potentially allow employers to prepare preventative measures instead of reactive measures. There has been minimal research performed within the field of interest in Malaysia, thus, the insights extracted could provide a breakthrough in assisting employer decision support.

An analytical approach would be to perform predictive analysis using techniques such as Decision Tree, Regression, and Clustering. An example of healthcare predictive analysis would be the use of

* Corresponding Author.
Email Address: angelal@sunway.edu.my (A. S. H. Lee)
https://doi.org/10.21833/ijaas.2021.09.004
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-3388-2372

Clustering and Decision Tree to predict healthcare coverages of an individual based on the significant attributes (Tekieh, 2012), while in another research, the development of a predictive model as a decision support system was proposed to predict the occurrence of Asthma and Diabetes (Jain, 2015). There are various researchers who have applied various techniques to perform prediction such as using a Multi-Level Clustering approach and Decision Tree, k-Medoids Clustering, Naïve Bayes, Bayesian Network, Multilayer Perceptron Model, and Logistic Regression. However, this research has too many clinical identifications and predictions while there has been little focus on potentially proposing a practical framework for classification ensemble modeling which can be applied by practitioners who are not experts in the field (Alharthi, 2018). These techniques which were applied in the models are based on mathematical formulations, statistical calculations, and technically too complex to be understood by others who are not experts in the field (Alharthi, 2018). The focus is consistently driving towards identifying an algorithm that can perform accurate predictions or increase predictive accuracy-but there should be more focus put into solving the underlying issue of practicality and interpretability (Bates et al., 2014). Ensemble learning refers to a combination of learners who are trained to solve the same problem (Tuysuzoglu et al., 2017). 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 et al., 2017). A practical framework to build ensemble models which can be applied by practitioners who are not experts is unclear while a comprehensive ensemble system framework may be too complex to understand (Abdunabi, 2016).

2. Literature review

2.1. Big data analytics and data mining

Big data analytics is inevitably connected to that of data science (Wang et al., 2018). In the early 2000s, an evolution of big data development broke through where it was defined by 3Vs: Volume, Velocity, and Variety (Wang et al., 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 analyze raw data into valuable pieces of information (Raghupathi and Raghupathi, 2014). Data is being generated at an exponential rate in almost every sector, especially in healthcare where it is described as one of the most data extensive industries (Raghupathi and Raghupathi, 2014). 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 et al., 2018). 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 and Raghupathi, 2014).

Data mining can be described as a process that 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). Data mining would usually include tools and techniques such as classification, regression, and clustering to perform analysis (Alonso et al., 2017). Each data mining technique would be used for different purposes depending on the objective. Classification and prediction are the most common modeling objectives-classification refers to the prediction of categorical labels (discrete or binary) while prediction refers to continuous value functions (Alonso et al., 2017). Cross-Industry Standard Process for Data Mining also known as CRISP-DM proposes a 6-step process methodology for data mining: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Koh and Tan, 2011). Business understanding is considered as the most important step where the business objectives, problem statements, and success criterion would be defined (Koh and Tan, 2011). 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 modeling (Koh and Tan, 2011). Modeling is the 4th step where data analysis would be performed classification models, regression models, association, clustering is some analytical techniques that are applied (Koh and Tan, 2011). The evaluation step would allow for a comparison of models based on their predictive accuracy prior to model selection (Koh and Tan, 2011). Once the model has been evaluated and selected, deployment can proceed with the actual implementation of the selected model (Koh and Tan, 2011).

3. Ensemble methods

As stated by Tukey, using two regression models with the first model fitted to the data and the second to the right the errors from the first model would enhance the prediction outcome. Ensemble learning refers to a combination of learners who are trained to solve the same problem (Tuysuzoglu et al., 2017). The fundamental idea of ensemble learning was to combine weak learners into strong learners, with the ability to provide better generalization errors while reducing the over-fitting of outputs (Tuysuzoglu et al., 2017). Ensemble modeling is proposed in this paper because, in predictive modeling, a single model based on one dataset could potentially

contain bias, high variance, or anomalies that will affect the prediction outcome. 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. When combining multiple models, we can reduce the uncertainty, even if they are not good individually, as we will not suffer from random errors from a single source. There are 3 main advantages when applying ensemble modeling which includes: (i) more accurate prediction outcomes, (ii) a more stable and robust model because by aggregating the results into multiple models, it potentially reduces noisy data as compared to individual models, (iii) capturing of linear and non-linear relationships in data through ensemble modeling of 2 different models (Ramzai, 2019).

This paper focuses on the method, Stacking. Stacking is a different approach to combining models with the concept of meta-learning (Solutions, 2016). This approach does not have any empirical formula for the weight function or any similar functionality as bagging or boosting (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. Stacking involves training a learning algorithm and combining the predictions with various other learning algorithms. 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." An example of the stacking ensemble method would be to combine a decision tree and regression model to predict an outcome. Stacking was introduced back in 1992 (Wolpert, 1992). Ensemble methods can be referred to as blending whereby the numbers are blended to produce a prediction. As mentioned, stacking has the basis of two learners which are the base learner and meta-learner. Predictions from a model are used as input for the following sequential layer and combined to form a collection of new predictions. During prediction, the output from the base learner is combined with the meta-learner's outcome to produce a combined final prediction (Menahem et al., 2009). Base learners would fit current data while meta-learner would take on predictions of the base learner.

The decision tree was the chosen modeling technique to be implemented within the framework for classification ensemble model prediction. A 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 and SriPreethaa, 2019). 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 and SriPreethaa, 2019). 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 (Chandrasekar et al., 2017). 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 a leaf node (Chandrasekar et al., 2017). 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 and SriPreethaa, 2019). It is also interpreted as a unique set of rules form which is characterized and denoted by its hierarchical organization rules (Raul et al., 2016). This hierarchy will allow for simple but powerful outcomes to make strategic decisions (Raul et al., 2016). 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 and SriPreethaa, 2019). Moreover, a Decision Tree can be often said to be mimicking human level of thinking hence, it is easily understandable and interpretable.

4. Findings of existing study

There are many predictive techniques that can be applied across different scenarios depending on the nature of the predictive outcomes. For example, in the study by Bruno et al. (2014), they used an approach known as clustering to identify groups of patients with similar characteristics and examination history in a dataset with variable data distribution. While applying a classification technique known as a decision tree to perform prediction (Bruno et al., 2014). Through this study, they identified that age, gender, and HDL (High-Density Lipoproteins) cholesterol were key drivers to determine diabetic patients (Bruno et al., 2014). They used a clustering technique known as multi-level clustering.

Moving on, a study performed by Jain (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) (Jain, 2015). Findings showed that by combining all 5 models, it managed to yield higher predictive accuracy (Jain, 2015). However, combining these techniques would increase the complexity of the predictive model as well. Key factors which were mentioned include, blood pressure, BMI, age, ethnicity, and smoking status (Jain, 2015). Abdunabi (2016) performed a study on proposing a framework for ensemble predictive modeling by applying a technique known as fusion modeling 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 modeling (Abdunabi, 2016). Tekieh (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. Tekieh (2012) managed to

identify 4 factors which were access to care, age, the poverty level of family, and race/ethnicity as the key factors which would show disparity among healthcare coverages. Furthermore, Tekieh (2012) applied the k-means clustering technique to discover groups of people with health coverage problems and inconsistencies. Tekieh (2012) demonstrated that the decision tree models provide higher accuracy than the models based on neural networks. Lin et al. (2014) proposed a Bayesian Multitask learning approach for healthcare predictive analytics for risk profiling within Electronic Health Records (EHR). Lin et al. (2014) concluded that age, body weight, gender, smoking status, ICD-codes, and medications are key drivers when determining risk profiling among patients. Lin et al.'s (2014) analysis showed that the BMTL (Bayesian Multitask Learning) approach can create significant potential impacts on clinical practice in reducing the failures and delays in preventive interventions.

Bates et al. (2014) performed an analytical study on how analytics can be applied to identify high-risk and high-cost patients. Bates et al. (2014) proposed approaches and techniques such as decision trees 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. Alharthi (2018) focused 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. Alharthi (2018) suggested that there is a lack of actionable knowledge towards meaningful progress for better patient outcomes and improves the quality of care.

4.1. Problem statement and objectives

Employers have a lack of understanding of the current employee population health profiles hence, the importance of having a better understanding of the employee health profile would lead to an overview of the health risk of an employee. High-risk employees may indirectly refer to a high-cost employee as well. Current research in healthcare such as medical conditions, healthcare coverages, etc. uses different machine learning algorithms such as Naïve Bayes, Bayesian Network, Decision Tree, Neural Network, and boosting and bagging techniques (Tekieh, 2012; Jain, 2015; Moturu et al., 2007). However, none of these proposed algorithms offer combined predictive models into a combination model, which could potentially lead to better prediction outcomes. Ensemble models have been applied across various fields such as weather forecasting, finance, manufacturing, security, and medicine. Moreover, there has been minimal focus on proposing a practical framework for classification ensemble modeling that can be applied by practitioners. The objective of an ensemble model aims at producing better predictive outcomes and accuracy which is why the paper focuses on developing a practical framework for developing an ensemble model to perform classification predictions. The focus revolves around clinical identifications and predictions which require an expert in the field to perform analysis, while little focus has been put into developing a practical framework that can be applied across a wide range of classification applications.

5. Framework for ensemble predictive modeling

Fig. 1 shows the proposed Framework for Ensemble Predictive Modelling. It consists of 4 main phases: (i) Business Understanding, (ii) Data Preparation, (iii) Model Development, and (iv) Model Evaluation.

5.1. Phase 1: Business understanding

- Market Trend(s): Understanding the current trends would enable an organization to identify alternatives to increase efficiency and effectiveness in strategic planning. Current trends show and act as a guide for improvement. In the 21st century, analytics has become such a booming interest that every organization wants to understand and discover how analytics can be implemented in their processes.
- Objective(s): Once the current market trend(s) has been discovered, the next phase would be to identify business objectives to help strategize the direction and results you hope to achieve. Every organization, business, or company is driven by specific objectives. There are short and long-term objectives and are usually based on tangible, measurable, or physical results. Objective(s) focuses on actions and tasks which help build towards the goal achievement. In analytics, objective(s) will allow for a clear and precise formulation of the problem statement to help solve an analytical problem.
- Problem Statement: Problem statement is a phrase described as the mental process to identify, discover, analyze, and solve problems (Annamalai et al., 2013). It involves a process to solve a problem including identifying and discovering the problem, strategy to tackle the area of concern, understanding of the problem, researching alternative solutions, and actions taken to achieve goals (Annamalai et al., 2013). In this paper, the problem revolves around the increasing medical costs which is an issue employers are looking extensively into, to reduce medical expenditure or potentially identify employees (patients) with higher risk or cost while providing proactive measures and solutions to prevent further deterioration.

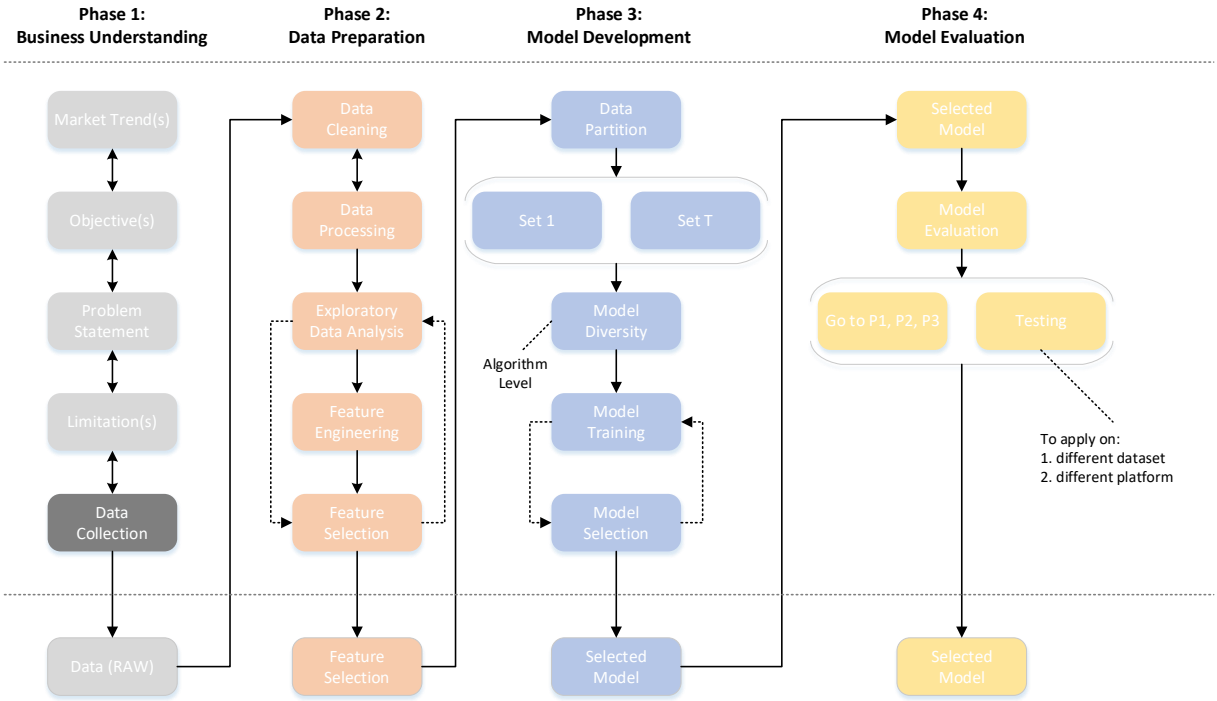

Fig. 1: Proposed practical framework for ensemble model building

5.2. Phase 2: Data preparation

- Data Cleaning and Data Processing: As observed, there are missing values, variations in wordings, inconsistent naming, etc. these findings observed are usually present in raw data format, hence, data preprocessing will be performed. The conversion of capital letters into a standardized format is known as normalization. Next, identifying anomalies such as missing or null values, spelling errors, format errors, and data anomalies (e.g., negative age). As these data will affect the outcome of the analysis, hence, it is best to remove, replace or transform these data. For missing or null values, we perform a process called imputation or replacement to replace the missing values with average/mean/median values, depending on the data structure. Data anomaly are abnormal data which are found in the dataset during data cleaning, for instance, a patient with negative age (-75), we would assume that there was a wrong input and replace the value with a positive age.
- Exploratory Data Analysis: Exploratory Data Analysis includes 2 main 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 behavior 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: Variable creation is the process whereby new variables are created with the requirements of the analysis. Variable creation is done to minimize variation in a variable through binning such as the amount incurred. 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.
- Feature Engineering (Target): No referencing or benchmark was used to categorize the "RiskLevel" of the patient. To identify the "RiskLevel" of a patient, more exploration was needed to better understand the data and how we could formulate the different levels of conditions to fulfill the Target. Firstly, there was a tagging of "LTM" referring to Long Term Medication as shown next to the highlighted column with an indicator of "Y" means yes or "N" means. Patients with a "tagging" of LTM can be referred to as higher risk patients as they have certain chronic conditions. Secondly, based on research by Koller et al. (2014), he labelled 46 different conditions as chronic conditions hence, using the same labeling, we label patients who were diagnosed with similar conditions as high risk, 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.
- Feature Selection (Fig. 2): It is a good practice to choose the features (variables or predictors) which will be useful as it is common to have features that are irrelevant and meaningless. Feature selection is

a process of selecting a subset of relevant features from all features without any transformation while validating it with regard to the analysis objective and building predictive models (Jović et al., 2015). There is a rule that states "garbage-in garbage-out" which is why data is 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).

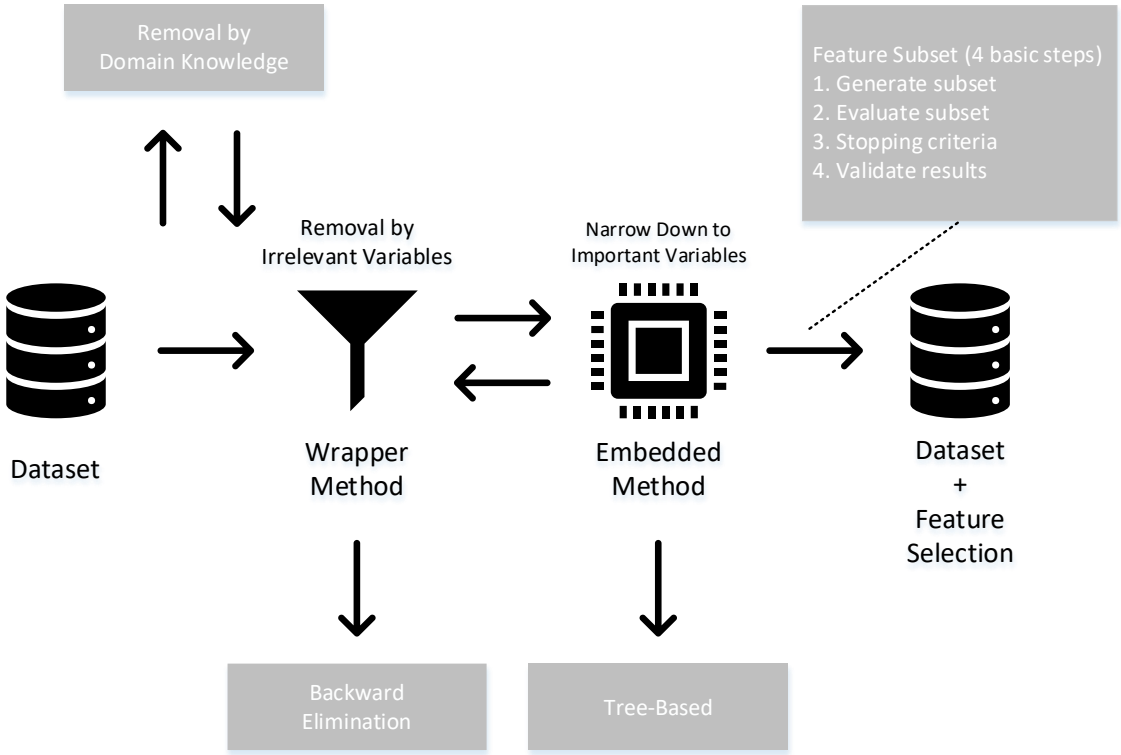

Fig. 2: Feature selection process

Remove features that have no relevance to the objective. Apply the wrapper method using the backward elimination technique, where we would start by using all the features then removing an irrelevant feature at each iteration. This process would be repeated until the satisfactory performance has been achieved. In this paper, we performed wrapper method to removed features with high variance or the set variance threshold (features with values which are the same), such as "ICD Codes" where the number of classes would exceed 512, hence these features should be removed to avoid overfitting where the analysis would yield too specific results. Next, to apply the tree-based embedded method where we would use a Decision Tree to further help to perform feature selection. Even though embedded methods are a combination of both filter and wrapper methods, in this paper, we applied the wrapper method prior to running the embedded method to further enhance the feature selection process as there we would be able to minimize high feature dimensionality prior to applying the embedded method to achieve more efficient and effective feature selection. When performing the feature selection process, a simple feature subset process will be carried out according to the following steps: (i) generate subset, (ii) evaluate subset, (iii) stopping criteria, and (iv) validate results. Once a feature subset has been selected, it will be evaluated in step 2-this process between steps 1 and 2 will be repeated multiple times 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.

- Data Understanding: The dataset used in this paper was obtained from the human resource department in a collaboration project. The dataset provided was in relation to employee claims history which includes basic demographic information, diagnosis, the amount incurred and insured. The initial dataset provided had approximately 40 attributes but after removal through domain knowledge, wrapper, and embedded methods (feature selection process), only 20 relevant attributes were selected to proceed with the analysis. Some of the attributes include Gender, Age, Relationship, Claim Frequency, MCDays, ICD Category, Amount Incurred, Amount Insured, and the Risk Level (which is the target value).

5.3. Phase 3: Model development

- Data Partition: Data partition refers to the allocation of data to perform various tasks such as model training, model evaluation, and model testing. This is one of the most crucial aspects of

predictive modeling. 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, the analysis might not be accurate as there might be bias in the model outcome.

- Model Diversity (Algorithm Level): The concept behind model diversity at the algorithm level focuses on injecting a variety of algorithms and techniques used to train models by applying the same training dataset and features to compare predictive accuracy and performance. As this paper 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.
- Model Training: 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 the decision tree as the main algorithm for classification prediction is one of the easiest algorithms to understand due to the if-else algorithm approach. Fig. 3 shows the ensemble method framework.

The meta learner and base learner would both be Decision Tree models (Fig. 3). The learners have been chosen based on the idea of easy interpretation and usability because of the functionality it would present towards individuals who are not experts in the field which is why we chose to combine Decision Tree and Decision Tree. Decision Tree has a nature that is low bias and high variance. Hence, combining multiple decision trees would reduce the variance while maintaining the low bias nature-this approach of combining decision trees would potentially increase the accuracy of prediction. The decision tree has the flexibility to handle missing values as well as selecting the important variables while ignoring irrelevant variables. This is the flexibility presented by an ensemble stacking method by combining various predictive techniques as required.

- Model Selection: Model selection involves selecting the best model based on the nature of the prediction. Once the ensemble method and model diversity has been decided, then model selection would be achieved by selecting the best-performing models.

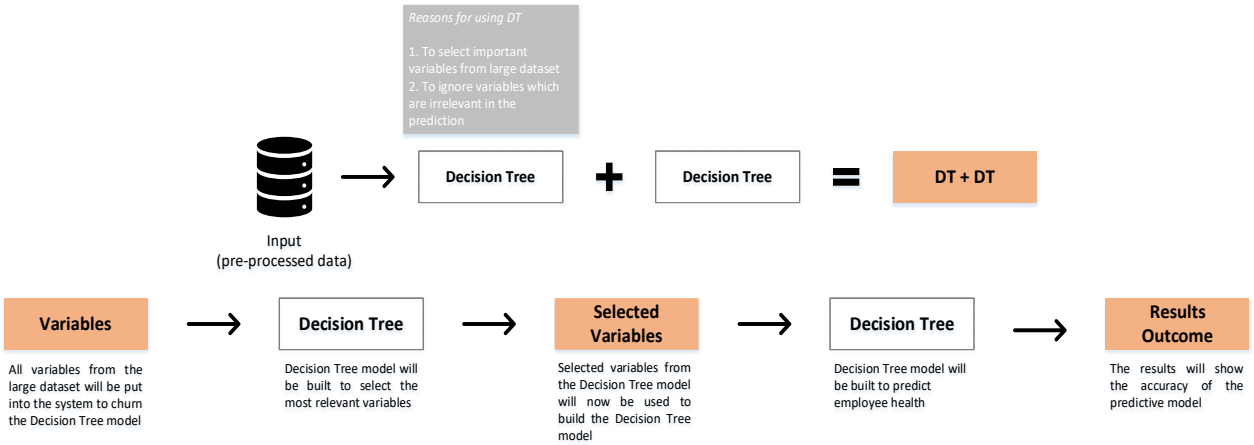

Fig. 3: Ensemble method framework (base tree+meta tree)

5.4. Phase 4: Model evaluation

Model evaluation is an important process in predictive modeling (Abdunabi, 2016). It is more evident when performing ensemble predictive modeling, as the performance and diversity of a model/algorithm must be evaluated completely to assess the effectiveness and predictive accuracy of the ensemble model (Abdunabi, 2016). 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 paper, we aim to test a practical ensemble framework using the ensemble method of stacking. In the analytics environment, the robustness of a model is tested by applying different datasets in various scenarios to discover the robustness and specificity of the model. A robust predictive model should achieve optimal performance by producing accurate and reliable results even with an increasing level of noisy data (Hu et al., 2008). Moreover, we suggested applying the framework on different platforms to ensure that all the frameworks can be applied on the various platform which reduces the limitation of the framework. Platform testing was done on proprietary software and an open-sourced software.

6. Preliminary results (testing with orange)

As shown in Fig. 4, ran using the platform Orange generated the node of Test and Score. It showed that the AUC (Area under ROC Curve) and CA (Classification Accuracy) both recorded that the Stacking Model of Decision Tree was the selected model with an 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. This shows that applying the proposed framework with the ensemble stacking model yield higher predictive accuracy than a single predictive model.

Model	AUC	CA
Stack	0.703	0.730
Tree	0.655	0.666

Fig. 4: Orange–test, and score

Fig. 5 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 line of each model is away (upwards) from the baseline, the higher the accuracy. From the ROC Chart, the 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, the ensemble model was the selected model due to the higher Classification Accuracy (CA).

Fig. 6 shows the Tree Overview generated by the platform Orange. It is observed that the predictors chosen to predict the target of "RiskLevel" were 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 you observe that the "L" in red occupies around ¾ 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.

7. Conclusion

In conclusion, this paper showed how patients' medical claim patterns and behaviors would potentially affect risk levels. Analyzing medical claim patterns and behaviors is potentially useful for employers to make decisions such as increasing medical premiums and healthcare plans for their employees. 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 behaviors of their employees. Through this analysis, characteristics that affect risk level could be further explored; for instance, TotalRemainingAmt, TotalAmtInsured, and ICD Category are factors that have an influence on the prediction of risk level in a patient.

Moreover, the objective of creating a practical classification ensemble model framework was achieved which also yields better and more accurate predictive results. The framework can be applied in various datasets across a wide range of platforms where the prediction type would be classification.

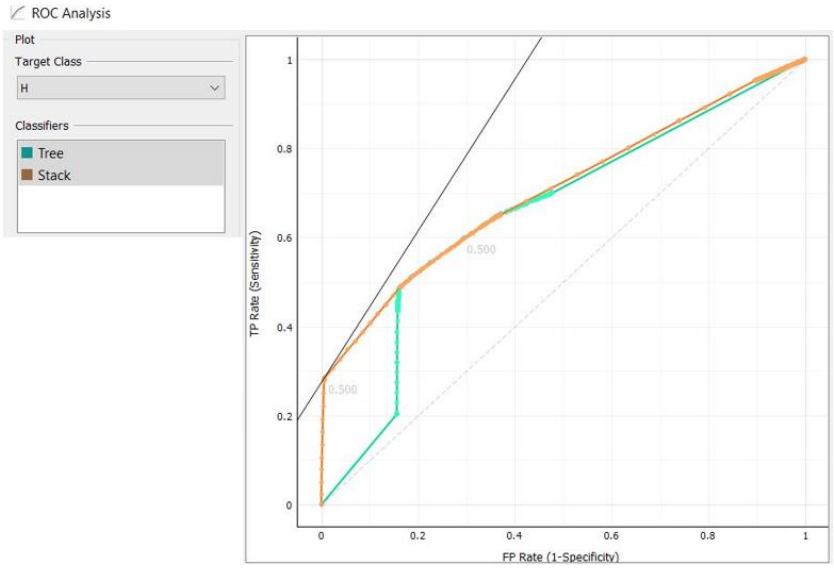

Fig. 5: Orange–ROC chart

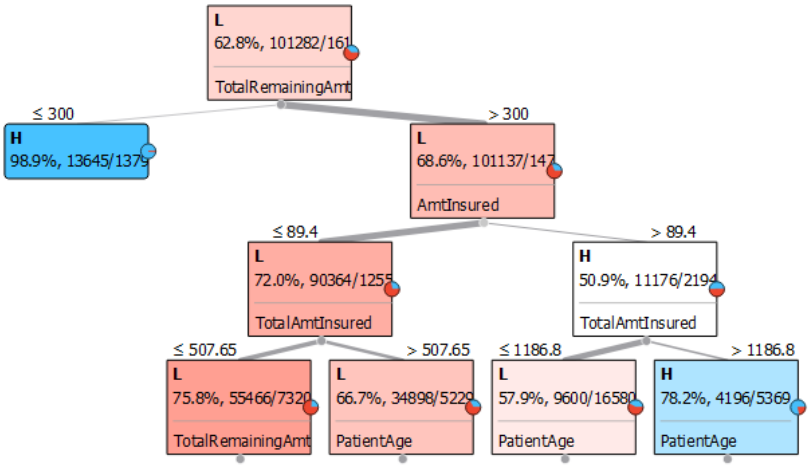

Fig. 6: Orange–tree overview

The concept of this approach was driven by the issue which has not been addressed whereby predictive models focus on enhancing and improving accuracy without addressing the issue of practicality and usability by practitioners who are not experts in this field. The conceptual framework presented focuses on key aspects in Phase 2 where Feature Engineering and Feature Selection was implemented-the Feature Selection approach, which uses a combination of the wrapper and embedded methods (through Decision Tree) to select the most significant features. As mentioned by (Alharthi, 2018), most research focuses on complex mathematical and statistical models to increase the accuracy while little focus has been given to bridge the gap through interpretability. With the proposed framework, it would potentially provide a greater exploration opportunity to professionals who are not in the field of analytics. Moving forward, further exploration of implementing enhanced algorithms can be performed using the ensemble method of stacking which can be used across the various domain, to produce more innovative outcomes. And the exploration of implementing and creating a fusion model can be explored with the concept of developing an enhanced model development framework for fusion ensemble modeling.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abdunabi TA (2016). A framework for ensemble predictive modeling. Ph.D. Dissertation, University of Waterloo, Waterloo, Canada.

Agarwal R (2019). The 5 feature selection algorithms every data scientist should know. Available online at: https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2

Alharthi H (2018). Healthcare predictive analytics: An overview with a focus on Saudi Arabia. Journal of Infection and Public Health, 11(6): 749-756. https://doi.org/10.1016/j.jiph.2018.02.005 **PMid:29526444**

Alonso SG, de la Torre Diez I, Rodrigues JJ, Hamrioui S, and Lopez-Coronado M (2017). A systematic review of techniques and sources of big data in the healthcare sector. Journal of Medical Systems, 41(11): 1-9. https://doi.org/10.1007/s10916-017-0832-2 **PMid:29032458**

Annamalai N, Kamaruddin S, Abdul Azid I, and Yeoh TS (2013). Importance of problem statement in solving industry problems. Applied Mechanics and Materials, 421: 857-863. https://doi.org/10.4028/www.scientific.net/AMM.421.857

Bates DW, Saria S, Ohno-Machado L, Shah A, and 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. https://doi.org/10.1377/hlthaff.2014.0041 **PMid:25006137**

Bruno G, Cerquitelli T, Chiusano S, and Xiao X (2014). A clustering-based approach to analyze examinations for diabetic patients. In the IEEE International Conference on Healthcare Informatics, IEEE, Verona, Italy: 45-50. https://doi.org/10.1109/ICHI.2014.14 **PMCid:PMC6353491**

Chandrasekar P, Qian K, Shahriar H, and Bhattacharya P (2017). Improving the prediction accuracy of decision tree mining with data preprocessing. In the IEEE 41st Annual Computer Software and Applications Conference, IEEE, Turin, Italy, 2: 481-484. https://doi.org/10.1109/COMPSAC.2017.146

Eapen AG (2004). Application of data mining in medical applications. M.Sc. Thesis, University of Waterloo, Waterloo, Canada.

Gore A (2012). The digital earth: understanding our planet in the 21st century. The Australian Surveyor, 43(2): 89-91. https://doi.org/10.1080/00050348.1998.10558728

Hu H, Li JY, Wang H, Daggard G, and Wang LZ (2008). Robustness analysis of diversified ensemble decision tree algorithms for microarray data classification. In the International Conference on Machine Learning and Cybernetics, IEEE, Kunming, China, 1: 115-120. https://doi.org/10.1109/ICMLC.2008.4620389

Jain R (2015). Predictive modeling for chronic conditions. M.Sc. Thesis, Florida Atlantic University, Boca Raton, USA.

Jović A, Brkić K, and Bogunović N (2015). A review of feature selection methods with applications. In the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Opatija, Croatia: 1200-1205. https://doi.org/10.1109/MIPRO.2015.7160458

Kincade K (1998). Data mining: Digging for healthcare gold. Insurance and Technology, 23(2): 2-7.

Koh HC and Tan G (2011). Data mining applications in healthcare. Journal of Healthcare Information Management, 19(2): 65-72.

Koller D, Schön G, Schäfer I, Glaeske G, van den Bussche H, and Hansen H (2014). Multimorbidity and long-term care dependency-A five-year follow-up. BioMed Central Geriatrics, 14(1): 1-9.
https://doi.org/10.1186/1471-2318-14-70
PMid:24884813 PMCid:PMC4046081

Lin YK, Chen H, Brown R, Li SH, and Yang HJ (2014). Healthcare analytics and clinical intelligence: A risk prediction framework for chronic care.
https://doi.org/10.2139/ssrn.2444025

Menahem E, Rokach L, and Elovici Y (2009). Troika–An improved stacking schema for classification tasks. Information Sciences, 179(24): 4097-4122.
https://doi.org/10.1016/j.ins.2009.08.025

Moturu ST, Johnson WG, and Liu H (2007). Predicting future high-cost patients: A real-world risk modeling application. In the IEEE International Conference on Bioinformatics and Biomedicine, IEEE, Fremont, USA: 202-208.
https://doi.org/10.1109/BIBM.2007.54

Raghupathi W and Raghupathi V (2014). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems, 2(1): 1-10.
https://doi.org/10.1186/2047-2501-2-3
PMid:25825667 PMCid:PMC4341817

Rahm E (2016). Big data analytics. IT-Information Technology, 58(4): 155-156. https://doi.org/10.1515/itit-2016-0024

Ramzai J (2019). Simple guide for ensemble learning methods. Available online at:
https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2

Raul A, Patil A, Raheja P, and Sawant R (2016). Knowledge discovery, analysis and prediction in healthcare using data mining and analytics. In the 2nd International Conference on Next Generation Computing Technologies, IEEE, Dehradun, India: 475-478.
https://doi.org/10.1109/NGCT.2016.7877462

Solutions V (2016). Improving predictions with ensemble model. Available online at:
https://www.datasciencecentral.com/profiles/blogs/improving-predictions-with-ensemble-model

Tekieh MH (2012). Analysis of healthcare coverage using data mining techniques. Ph.D. Dissertation, University of Ottawa, Ottawa, Canada.

Tuysuzoglu G, Birant D, and Pala A (2017). Ensemble methods in environmental data mining. IntechOpen, Rijeka, Croatia.
https://doi.org/10.5772/intechopen.74393

Wang Y, Kung L, and Byrd TA (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126: 3-13.
https://doi.org/10.1016/j.techfore.2015.12.019

Wolpert DH (1992). Stacked generalization. Neural Networks, 5(2): 241-259.
https://doi.org/10.1016/S0893-6080(05)80023-1

Yuvaraj N and SriPreethaa KR (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing, 22(1): 1-9.
https://doi.org/10.1007/s10586-017-1532-x