Filter Variable Selection Algorithm and Knowledge Discovery in Datasets

Donald Douglas Atsa'am

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
June 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy in Applied Mathematics and Computer Science.

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Applied Mathematics and Computer Science.

Asst. Prof. Dr. Ersin Kuset Bodur
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev
2. Prof. Dr. Hamza Erol
3. Prof. Dr. Efendi Nasiboğlu
4. Asst. Prof. Dr. Ersin Kuset Bodur
5. Asst. Prof. Dr. Müge Saadetoğlu

ABSTRACT

This dissertation addressed two aspects within the Data Mining field: filter variable selection, and knowledge discovery in datasets. A filter algorithm that serves to reduce the feature space in datasets, with special attention to healthcare data, was developed and tested. The algorithm binarizes the dataset, and then separately evaluates the risk ratio of each predictor with the response, and outputs ratios that represent the association between a predictor and the class attribute which translates to the importance rank of the corresponding predictor. The performance of the developed algorithm was compared against some existing feature selection algorithms on different datasets, using classification models. In the majority of the cases, the predictors selected by the new algorithm outperformed those selected by the existing algorithms. The proposed filter algorithm is therefore a reliable alternative for variable ranking in data mining classification with a dichotomous response.

In the aspect of knowledge discovery in datasets, the relationship between employees' psychological capital (PsyCap) and educational qualifications, and the relationship between employees' PsyCap and organizational tenure was mined. The PsyCap and demographic data of 329 employees in the hospitality industry were collected. The odds ratio (OR) technique was deployed to measure the associations which revealed that, employees with higher educational qualifications are 2.6 times more likely to have positive psychological capital than those with lower educational qualifications. It was also discovered that employees who have stayed longer periods within the service of an organization are 3.6 times more likely to be seen as having positive psychological capital compared with those who have stayed shorter periods. The results of the two associations are statistically significant at p-value = 0.002 < 0.05 and p-value = 0.004 < 0.05, respectively. These findings will guide business owners on the calibre of employees to hire, retrench, or retain during general recruitment or retrenchment.

Keywords: Data mining, Classification, Attribute selection, Odds ratio, Filter algorithm, Balanced classification accuracy

ÖZ

Bu çalışma veri madenciliği alanındaki iki konuya değinmiştir: nitelik altküme (değişken) seçimi ve bilgi keşfi. Özellikle sağlık alanındaki veriler kullanılarak, veri kümelerindeki değişken miktarını azaltmaya yarayan bir algoritma geliştirildi ve test edildi. Önerilen algoritma, veriyi ikili sayma sistemi durumuna getirir ve herbir değişkenin ayrı ayrı sınıf değişkenine göre risk oranını değerlendirir ve değişkenleri risk oranına bağlı olarak önem derecesine göre sıralar. Geliştirilen algoritmanın performansı, bilinen diğer sınıflandırma algoritmaları ile farklı modeller kullanılarak karşılaştırılmıştır. Vakaların çoğunda, önerilen algoritma mevcut algoritmaların sonuçlarından daha iyi sonuçlar vermiştir. Bu nedenle önerilen algoritma, veri madenciliği sınıflandırmasında değişken sıralama için güvenilir bir alternatiftir.

Veri setlerinde bilgi keşfi açısından, çalışanların psikolojik durumu ile eğitim kalitesi arasındaki ilişki ve çalışanların psikolojik durumu ile çalışanların görev süresi arasındaki ilişki incelenmiştir. Bu amaçla, konaklama sektöründe 329 çalışanın psikolojik durumu ve demografik verileri toplanmıştır.

Yüksek vasıflı niteliklere sahip çalışanların, pozitif psikolojiye sahip olma ihtimalinin düşük eğitim niteliklerine sahip olanlara oranla 2.6 kat daha fazla olduğunu ölçmek için, göreceli olasılıklar oranı tekniği uygulandı. Ayrıca, daha uzun süre çalışanların, kısa süreli çalışanlara kıyasla pozitif psikolojiye sahip olma ihtimalinin 3.6 kat daha fazla olduğu tespit edilmiştir.

İki ilişkinin sonuçları sırasıyla $p = 0.002 < 0.05$ ve $p = 0.004 < 0.05$ değerlerinde istatistiksel olarak anlamlıdır. Bu bulgular, işletme sahiplerine çalışanları konusunda işe alım, işten çıkarma veya genel işe alım veya genel işten çıkarma konusunda rehberlik edecektir.

Anahtar Kelimeler: Veri madenciliği, Sınıflandırma, Değişken seçimi, Göreceli olasılıklar oranı, Filtreleme algoritması, Dengelenmiş sınıflandırma doğruluğu

To Doose, Tor’bem, and Er’amodoo

ACKNOWLEDGEMENT

This is to appreciate my supervisor, Asst. Prof. Dr. Ersin Kuset Bodur, who gave me all the needed support, encouragement and time throughout the duration of my research work. Dr. Ersin exhibited towards me the kind of patience and simplicity that is very rare to find in a supervisor. She went an extra step to make sure we conducted a thorough research. I make bold to say I am a lucky person to have found her as my supervisor.

I acknowledge the Chair and lecturers of the Department of Mathematics who imparted useful knowledge to me during the Ph.D programme. I wish to thank, particularly, members of my Thesis Monitoring Committee and Jury Committee. Your contributions and constructive criticisms were all this work needed to achieve this magnitude of success. I am indebted to my wife, Barr. Doose Don Atsa’am, and to my sons, Tor’bem and Er’amodoo; my parents, other family members and in-laws, for their support and good will. I appreciate Mr. Sebastine Hon (SAN), Rev. Dr. Philip Tachin, Mr. Terver Akase, Mr. Tor Jooji, Mr. Andrew Ayabam, Prof. Rufus Sha’Ato, Engr. Joseph Amahundu, Dr. Dzer Hembe, Mr. Tersoo Yese, Dr. Uju Alola, Prof. N.V. Blamah, Dr. I. Agaji, Messrs. Terungwa Magen, Msughter Iyoryisa, Olakunle Dada, Charles Erukaa, Mtema Chin and all who contributed variously to the success of this programme. I am grateful to Mr. Michael Awuya, Miss Aver Adikpo, and members of St. Cyril’s Catholic Chaplaincy in Famagusta, and my flat mates at Noyan 7 Apartment. Your friendship was wonderful. I appreciate the Vice Chancellor of the Federal University of Agriculture, Makurdi, Nigeria - Prof. Richard Kimbir, for his support and good intentions. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

t_{00}	Total observations with *input* = 0 and *output* = 0
t_{01}	Total observations with *input* = 0 and *output* = 1
t_{10}	Total observations with *input* = 1 and *output* = 0
t_{11}	Total observations with *input* = 1 and *output* = 1
F^{-}	False Negative
F^{+}	False Positive
T^{-}	True Negative
T^{+}	True Positive
β_{ij}	t_{10}
δ_{ij}	t_{11}
ϕ_{ij}	t_{01}
φ_{ij}	t_{00}
B	Backward Elimination
BCA	Balanced Classification Accuracy
CI	Confidence Interval
Edu	Educational Qualification
F	Forward Selection
FNR	False Negative Rate
FPR	False Positive Rate
F.S.	Fisher Score
N	Nvmax

OR	Odds Ratios
OrgTenure	Organizational Tenure
P	Pearson’s Correlation
P.A.	Proposed Algorithm
PCQ	Psychological Questionnaire
PsyCap	Psychological Capital
RR	Risk Ratios
SE(logOR)	Standard Error of the Log Odds Ratio
TNR	True Negative Rate
TPR	True Positive Rate
V	VarImp
varImp	Variable Importance
VIM	Variable Importance Ranking

Chapter 1

INTRODUCTION

This dissertation focuses on two sub-areas of data mining: variable selection, and knowledge discovery in databases. In the aspect of variable selection, we will design and test a filter variable selection algorithm that uses risk ratios (RR), otherwise known as relative risk, to rank the importance of predictor variables for data mining classification problems, with special attention to healthcare data. In the knowledge discovery component, we will investigate the relationship between educational qualifications, organizational tenure, and psychological capital of employees in the hospitality industry.

Variable importance ranking is the process that assigns numeric values, or some other form of quantifiers, to individual predictors in a dataset, indicating the level of their importance in predicting the outcome. After such a ranking has been established, variables that rank low can be expunged from a predictive model without compromising goodness of fit or predictive accuracy. Variable selection is necessary in the era of big data where voluminous data is generated from healthcare activities, including diagnosis, epidemiology analysis, and patient medical history. These data often consist of many attributes, some of which are not needed in data mining classification, and thus the need to select only the relevant ones is imperative.

Data mining is the process of extracting useful but hidden information from existing data sources (Han & Kamber, 2000). Seven steps are involved in the process of data mining (Al-Radaideh & Nagi, 2012; Sumathi, Kannan, & Nagarajan, 2016). These include the cleaning step during which irrelevant data is removed, the integration step when data obtained from multiple sources are combined. Following the integration step is the data selection step, during which relevant data to be mined are chosen based on the task at hand. Other steps include, transforming the data into summarized forms to ease the mining process, and the mining step where useful patterns are extracted. The last two steps are pattern evaluation, and knowledge presentation. In pattern evaluation, the validity of the extracted knowledge is assessed, while the knowledge presentation step makes available to the public the useful information extracted. According to Wu (2013), three major techniques of data mining are: clustering, association rule mining, and classification and prediction. Clustering is concerned with grouping similar data objects together into same classes called clusters (Lazhar & Yamina, 2016). This technique is a form of unsupervised learning because class labels are not known ahead of time. Clustering allows observations to be organized into a hierarchy consisting of similar events. The overall objective of clustering is to obtain very accurate clusters with minimum inter-cluster similarity and maximum intra-cluster similarity.

Association rule mining technique extracts the relationship that exists among data items in a dataset (Kantardzic, 2011). In order to generate association rules, the frequency of occurrence between two or more items is considered. A common application of association analysis is the market basket analysis where the relationship among frequently purchased items in a market store is determined (Wu, 2017). Association rule mining is an unsupervised learning task as the relationship that might exist between data objects is not known in advance (Kantardzic, 2011). Classification is concerned about developing models that accurately distinguish one data class from another (Al-Radaideh & Nagi, 2012). After this is done, the developed model is then used to predict the class of objects whose class is not known. Apart from predicting the class label of data objects, this technique is also used in predicting missing data values in a given dataset. Classification models often take the form of IF-THEN rules, mathematical formulae, neural networks, or decision trees (Lazhar & Yamina, 2016).

Consider a dataset $D = \{ (X_1, Y_1), ..., (X_m, Y_m) \}$ consisting of m observations, where $X = (X_1, ..., X_n)$ are predictor variables having dimension $X \in R^n$ and $Y \in C$ where C is a class label. In data mining, classification is defined mathematically as a mapping of the form $t : R^n \to C$ where t is a classifier (Genuer, Poggy, & Tuleau-Malot, 2010). One of the ways of measuring the performance of a classifier is by evaluating its classification accuracy; that is, how accurately it can predict the classes of a set of vectors whose classes are unknown (Freenay, Doquire, & Verleysen, 2013). According to Tharwat (2018), classification accuracy evaluation methods are divided into two categories: scalar metrics and graphical methods. Scalar metrics compute accuracy by taking the ratio of correctly classified observations versus total number of observations in the validation set. For binary classification problems, scalar values representing accuracy are obtained from a confusion matrix, which is a tabulation of actual and predicted classes for each sample (Tharwat, 2018; Lever, Krzywinski, & Altman, 2016). In graphical methods, such as a receiver operating characteristics curve, accuracy is plotted on a x, y-axis to represent the tradeoffs between the cost of correct or wrong classification into class 0 or 1 (Lever, Krzywinski, & Altman, 2016). Another method of evaluating classification accuracy is the k-fold cross-validation, also referred to as re-sampling (Anguita, Ghelardoni, Ghio, Oneto, & Ridella, 2012). This method divides the dataset into k subsets, uses k-1 subsets to train the classifier and then tests its performance on one subset. The process is done iteratively, reshuffling subsets until accuracy has been evaluated on all vectors (Anguita, et. al, 2012; Jung & Hu, 2015).

Variable selection has been identified as an important step towards constructing classification models that achieve higher accuracy (Freenay, Doquire, & Verleysen, 2013). Inclusion of variables with little or no modeling value in machine learning negatively affects the predictive power of classifiers. The algorithm to be developed in this research, using risk ratios, is expected to offer a good alternative to existing filter methods of variable selection. The RR, just like odds ratios (OR), is a statistical measure of the association between binary variables across two different groups, where one group is referred to as the independent group while the other as the dependent group (Schmidt & Kohlmann, 2008; Last & Wilson, 2004). While odds ratios are known to overestimate the strength of association, the RR technique does not exhibit this demerit (Tamhane, Westfall, Burkholder, & Cutter, 2016). Additionally, Odds ratios have the property of reciprocity, which allows for the direction of an association to be changed by taking the inverse of the OR estimate (Tamhane et al., 2016). It turns out that RR does not exhibit this property. For the purposes of variable importance ranking, the direction of association is usually from independent variable to dependent variable and not vice versa; therefore, the lack of reciprocity in RR is a good property to be explored for use in feature selection.

Existing literature indicates that risk ratios have been deployed previously for different purposes within the healthcare domain. For example, Rohde, Dimcheff, Blumberg, Saint, & Langa (2014) used RR to evaluate the extent to which red blood cells transfusion strategies are associated with the risk of infection among patients. The study, conducted on 7456 patient records, concluded that irrespective of the strategy used, blood transfusion was not associated with reduced risk of infection, generally. However, transfusion strategies were found to be associated with a minimized risk of specifically dangerous infections. In a related study, Capistrant, Moon, & Glymour (2012) used RR to investigate the relationship between caregiving and risk of hypertension incidence among American older adults. The research, conducted on 5708 Americans aged 50 years and above, held that caregiving for a spouse is associated with the possibility of becoming hypertensive by the caregiver in the long run. The association between diabetes and the possibility of prostate cancer incidence was investigated by Tseng (2011) using RR. The research reported risk ratios representing the extent of this association as 5.83, 2.09, and 1.35 for ages 40–64, 65–70, and 75 years and above, respectively, on a sample of 1 million Taiwanese patients. The aforementioned applications of RR in the healthcare domain give evidence that this technique holds good prospects for further deployment in this area. To our knowledge, risk ratios have so far been applied only in cohort and specific studies, with results limited in scope and generalization potentials. Against this backdrop, this research will explore the possibility of using the RR as the basis for developing a generic variable importance ranking algorithm. The algorithm will facilitate reduction of the dimension space of any healthcare dataset in order to enhance predictive accuracy and efficiency.

Relying on the usefulness of the RR measure, this study will construct an algorithm that first binarizes data values of predictors in a dataset with a dichotomous response. Next, the algorithm evaluates the RR of each predictor with the response, and then outputs a value that signifies the relative importance of that predictor in determining the response. Computed values of RR will indicate the strength of the association, with larger values meaning strong association and, thus, high importance.

Meanwhile, knowledge discovery which is the second aspect of this dissertation is the process that analyzes and models datasets in order to identify novel but hidden patterns from the datasets (Holzinger & Jurisica, 2014). Knowledge mining will be performed on psychological capital (PsyCap, Luthans, Youssef, & Avolio, 2007) dataset drawn from a sample of employees in the hospitality industry (Atsa'am & Bodur, 2019a). This will be done with the aim of discovering the relationship between employees' psychological capital and educational qualifications on the one hand, and psychological capital and organizational tenure on the other hand. According to Antunes, Caetano, & Cunha (2017), psychological capital is a measure of the positive abilities of a human being that enable them excel in chosen endeavours. Luthans, Youssef-Morgan, & Avolio (2015) identified four components of PsyCap to include hope, efficacy, resilience, and optimism. The Psychological Capital Questionnaire (PCQ) is the tool used in measuring PsyCap level of an individual (Luthans et al., 2007). The PCQ will be distributed to a number of employees in tourism-related organizations in order to generate the experimental dataset. The educational qualification and organizational tenure information of the employees will also be captured. The Odds Ratio (OR) technique (Kleinbaum & Klein, 2010) will then be deployed to evaluate the association between psychological capital and educational qualifications; and psychological capital and organizational tenure. The OR is a statistical measure that evaluates the relationship between two variables across groups (Sperandei, 20014). The knowledge that will be mined from the PsyCap dataset will establish whether higher academic qualifications confer positive psychological capital on hospitality employees, and whether the duration of service by an employee within an organization has effect on their PsyCap.

Chapter 2

LITERATURE REVIEW

2.1 Chapter Overview

This chapter covers review of existing variable selection techniques relevant to this research. The materials to be used in the design of the proposed algorithm and for testing its effectiveness are also presented. Equally, relevant literature regarding psychological capital will also be discussed. Section 2.2 discusses feature selection generally, including the various categories. Section 2.3 narrows the discussion of feature selection to the filter category, which falls under the supervised filter selection methods. Section 2.4 presents some unsupervised feature selection methods relevant to this research. In Section 2.5, the definition of the research tools to be adopted in the Methodology is carried out. These include classification tools and statistical techniques that will serve in the proposed algorithm design and the knowledge discovery experiment. Section 2.6 explains the four components of psychological capital (PsyCap) and reviews existing literature on employees PsyCap and the work environment.

2.2 Feature Selection

Consider a dataset with predictor variables $X = \{X_1, X_2, ..., X_n\}$ in a high dimensional space R^n, and a class variable Y. The objective of feature selection is to find a subset k of n where $k < n$, producing classification models with high predictive accuracy (Freenay et al., 2013). Dadaneh, Markid, & Zakerolhosseini

(2016) reported four categories of features as (a) irrelevant, (b) weakly relevant and redundant, (c) weakly relevant but not redundant, and (d) strongly relevant features. A feature is considered irrelevant if it contains no useful information needed in modeling the problem at hand. Weakly relevant and redundant features contain little information needed in classification; however, there exist(s) one or more features within the same feature space that contains similar information with such feature. Weakly relevant but not redundant features do not have any major role to play in the classification model; however, there exists no other feature with similar information. Relevant features contain the required amount of information needed to model the problem domain. Feature selection is performed in order to achieve dimensionality reduction, which ultimately aims at producing a smaller subset that consists of only relevant variables for machine learning (Chandrashekar & Sahin, 2014; Dadaneh et al., 2016). With big data, the variable space is often large such that features with no classification value are equally included within the dimension (Bermejo, Ossa, Gamez, & Puerta, 2012). Limiting the number of features included in a model to only the relevant ones has several advantages: machine learning algorithms train faster, model complexity and overfitting are reduced, and predictive accuracy is enhanced (Frenay et al., 2013; Bagherzade-Khiabani et al., 2016).

Basically, there are three methods of feature selection: filters, wrappers, and embedded (Javed, Babri, & Saeed, 2014; Cateni, Colla, & Vannucci, 2014).

2.2.1 Filters

These methods use statistical techniques to select subset of variables independent of any machine learning algorithm (Huang, Wulsin, Li and Guo, 2009). One of the bases for selecting a feature is determined by the score of its correlation with the outcome variable. Some basic statistical tests for correlation include the Chi-square, ANOVA (Analysis of Variance), and the Pearson's correlation (Chandrashekar & Sahin, 2014). These and many more modernized statistical approaches have been utilized in procedures for selecting relevant dataset features for machine learning tasks. Two categories of filter methods; namely, the rankers and subset selectors have been identified in Bagherzade-Khiabani et al. (2016). While rankers generate numeric values indicating the importance of a predictor in determining the class, selectors are concerned about generating subsets of variables that collectively produce accurate model outcomes. It should be noted that rankers typically generate variable importance ranking without suggesting which variables to be included or excluded into a model, unlike subset selectors. It is the user's responsibility to determine a cut-off point of the variables to be included into model construction, using the ranking as guide. Performance diagnostics of the previous model determines if further exclusion or inclusion of predictors is needed until a perfect classification model is obtained. The fact that filters do not depend on any machine learning algorithm, they generally serve in the preprocessing step of data mining (Crone & Kourentzes, 2010). After the best attribute subset has been filtered out, any learning algorithm at the disposal of the modeler can be deployed for modeling (Hu, Bao, Xiong, & Chiong, 2015). Compared to other feature selection methods, filters exhibit the advantages of being computationally inexpensive and less prone to overfitting (Javed et al., 2014; Hu et al., 2015).

2.2.2 Wrappers

In wrapper methods, the learning procedure and feature selection are done by the same algorithm (Huang et al., 2009). Subset selection is achieved by means of statistical resampling where different variable subsets are routinely trained before arriving at a subset that produced better classification results (Bagherzade-Khiabani et al., 2016). Unlike filters, wrappers are computationally intensive due to the fact that several models with all possible subsets have to be built before producing the best subset (Frenay et al., 2013). Apart from the disadvantage relating to computational cost, results of wrappers lack generality since they are limited to specific machine learning algorithms. However, because features are optimized for specific training algorithm, wrappers produce models with better performance than filters (Bagherzade-Khiabani et al., 2016). Xue, Yao, & Wu (2018) identified three common procedures of wrappers; namely, forward selection, backward elimination, and recursive elimination. In the forward selection procedure, the algorithm starts with a model having no feature and iteratively adds features to the model until a point when addition of a new feature does not improve model performance. Backward elimination starts with all variables and iteratively removes insignificant features until variable removal does not improve model performance. The recursive elimination procedure iteratively constructs models using different variable subsets. At each iteration, the procedure sets aside the best or worst feature then builds the next model with the remaining features. The process continues until all features may have been utilized, then the algorithm ranks the variables based on the order they were eliminated.

2.2.3 Embedded

In this approach, the feature selection activity is integrated into the training process of the machine learning algorithm (Javed et al., 2014). The main distinctive quality of the embedded method from the wrapper method is that, while wrappers consist of two separate procedures for subset selection and training, embedded methods perform feature selection and training within the same procedure. Computational time requirement in embedded methods is smaller than that in wrappers since feature selection and training are done hand-in-hand in the former methods of feature selection (Lazar et al., 2012). Furthermore, with embedded technique, process parameters obtained during training are updated iteratively and they evolve by virtue of the efficiency of the model being constructed (Cateni, Colla, & Vannucci, 2017).

2.3 Filter Feature Selection Approaches

In this section, the review of some filter feature selection approaches is conducted. Filter methods belong to the category referred to as supervised variable selection (Dadaneh et al., 2016). This is so because the user is actively involved in the selection process which is done as a preprocessing activity. Generally, filter methods are concerned about measuring the strength of the association among variables; which could be predictor-to-predictor or predictor-to-class. According to Javed et al. (2014), metrics of association are of three categories: correlation measures, information-theoretic measures, and probabilistic measures. The correlation-based measures evaluate the linear relationship among two variables; and predictor variables exhibiting strong correlation with the class are selected for modeling. The information-theoretic metrics evaluate the mutual information contained in two different variables (Meyer, Schretter, & Bontempi, 2008). If two variables are found to exhibit similar characteristics, one of them can be eliminated from the data mining classification models in order to check redundancy. The probabilistic metrics measure the dependence between a predictor and the class using probability distributions (Cateni et al., 2014). These metrics generate estimates ranging over [0, 1] and higher values indicate the importance of the predictor in modeling.

- **The Fisher Index**

This metric computes the importance of the i-th explanatory variable in a dataset with binary response as shown in equation (1):

$$F_i = \left| \frac{\overline{X_1}(i) - \overline{X_0}(i)}{d_1^2(i) + d_0^2(i)} \right| \tag{1}$$

where $\overline{X}_1(i)$ is the mean and $d_1(i)$ is the standard deviation of the i-th variable with outcome 1; $\overline{X}_0(i)$ is the mean and $d_0(i)$ is the standard deviation of the i-th variable with outcome 0 (Maldonado & Weber, 2009).

- **The t-test**

According to Rice (2007) and Cateni et al. (2014), the t-test calculates the importance of the i-th predictor variable as shown in equation (2):

$$t_i = \frac{\left|\overline{X_1}(i) - \overline{X_0}(i)\right|}{\sqrt{\frac{d_1^2(i)}{n_1} + \frac{d_0^2(i)}{n_0}}} \tag{2}$$

where $\overline{X}_1, \overline{X}_0, d_1, d_0$ are as defined in equation (1), where n_0 and n_1 are total observations in the class 0 and 1, respectively.

- **The Kullback Liebler Distance (KL-distance)**

This is a filter method that evaluates the relative entropy between predictor variables, measuring the difference in their probability distributions (Cateni et al., 2014). For $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_n\}$ both discrete, the KL-distance is defined as

$$KL(X,Y) = \sum_i X_i \log_2 \left(\frac{X_i}{Y_i} \right) \tag{3}$$

See (Kullback & Leibler, 1951). It should be pointed out that the KL-distance is not symmetric; and in situations when X and Y are continuous, an integral replaces the sum in equation (3).

- **The Wilcoxon Rank Sum Test**

The Wilcoxon rank sum is a form of non-parametric test for two populations with no requirement for specific distributions (Li, Liu, Tung, & Wang, 2004). For a feature X, collection $C = \{A, B\}$ of classes A and B, the Wilcoxon measure, $w(X, C)$ is obtained using the following steps:

i. The values $v_1, v_2, ..., v_{n(A)+n(B)}$ of X under consideration across all samples in C are sorted in ascending order.

ii. To each value v_i in (i) above, assign a rank, denoted by $r(v_i)$, such that ties are handled by taking the average as in equation (4).

$$rank(v_i) = \begin{cases} i & \text{if } v_{i-1} \neq v_i \neq v_{i+1} \\ \dfrac{\sum_{k=0}^{n}(j+k)}{n+1} & \text{if } v_{j-1} < v_j = ... = v_i = ... = v_{j+n} < v_{j+n+1} \end{cases} \tag{4}$$

iii. Then, the sum of the ranks for the class having fewer samples is returned as the Wilcoxon index (Li et al., 2004). That is, $w(X, C) = \sum_{v \in \operatorname{argmin}} r(v)$. In situations where both classes contain same number of samples, any class is chosen arbitrarily.

- **Signal-to-Noise Measure**

Given two classes, C_1 and C_2, in a sample dataset, the signal-to-noise measure is based on this argument: a feature that is relevant must contribute in separating data points in C_1 from those in C_2 (Li et al., 2004). If a feature is irrelevant, it will contribute little or nothing in separating data points in C_1

from those in C_2. Specifically, this measure holds that if the values of a feature are substantially dissimilar in both samples in C_1 and C_2, then that feature is likely to be of more relevance than another feature which has similar values in both C_1 and C_2. For a feature f, the signal-to-noise concept is evaluated using the formula in equation (5) (Golub et al., 1999; Li et al., 2004).

$$S(f, C_1, C_2) = \frac{\left| \overline{X}_{C_1} - \overline{X}_{C_2} \right|}{d_{C_1} + d_{C_2}} \tag{5}$$

where $\overline{X}_{C_1}$ and $\overline{X}_{C_2}$ are the means of data points in class C_1 and C_2, respectively; d_{C_1} and d_{C_2} are standard deviations in C_1 and C_2, respectively. According to Li et al. (2004), if $S(f, C_1, C_2) > S(f', C_1, C_2)$ then the feature f is considered better than the feature f'.

- **Feature Selection Based on Conditional Mutual Information**

  Fleuret (2004) developed a feature selection procedure that relies on conditional mutual information. The technique, designed to operate on binary data, iteratively selects features which maximize mutual information with the class. At each iteration, features similar to the ones already selected are skipped which guarantees that selected features convey unique information and are weakly dependent on each other. Each time an iteration takes place and a feature is added to the subset, a score table is updated and the next score is calculated using equation (6).

$$S(n) = \min_{l<k} \hat{I}(Y; X_n \mid X_{v(l)}) \tag{6}$$

where $S(n)$ is the score updated at each iteration, X_n is the feature being evaluated currently, $X_{v(l)}$ is the set of features already selected.

- **Pearson's Correlation**

  This is one of the correlation-based filter methods, and the importance of a feature is computed as presented in equation (7) (Chandrashekar & Sahin, 2014).

$$P_i = \frac{Cov(X_i, Y)}{\sqrt{Var(X_i) \times Var(Y)}} \tag{7}$$

  where X_i is the i-th predictor variable and Y is the outcome label, $Cov()$ is the covariance and $Var()$ is the variance. The Pearson's correlation evaluates the importance of predictor variables using each predictor's individual linear dependence with the outcome.

Tran, Afanador, Buydens, & Blanchet (2014) examined three methods of filter variable selection including the Variable Importance in the Projection (VIP), Beta CI, and the selectivity ratio. The VIP determines importance of variables by measuring the proportion of the explained variance of each predictor and the covariance between each predictor and the class. Usually, the average VIP equals 1, and variables that produced VIP scores greater than 1 are selected as important. The Beta CI technique evaluates the confidence interval bounding the coefficients of regression for each variable. A variable is considered important if its Beta CI does not overlap zero. In the selectivity ratio method, the sum of squares of each variable is measured by taking the ratio of the explained variance versus the residual variance, and the resultant value signifies the importance of the corresponding variable.

.Murtaugh (2009) reported two approaches to filter variable selection which include stepwise and all subsets procedures. According to Murtaugh (2009), the stepwise procedures use some form of quantitative measure, such as p-values and Akaike Information Criteria (AIC), to compare models. Explanatory variables are sequentially added and/or deleted until a point is reached based on the threshold value of the quantitative measure. In the all subsets procedures, explanatory variables are grouped in all possible subgroupings and the subsets that produced the most appropriate value of the quantitative measure are selected.

2.4 Unsupervised Variable Selection

The second category of feature or variable selection methods considered in this research is the unsupervised variable selection. These methods are termed so because the selection process takes place without interference from the user.

2.4.1 Automatic Variable Selection

The R programming language, developed by R Core Team (2017), has a package called leaps that consists of a function, regsubsets, used for automatic selection of best variables (Lumley, 2017). Irrespective of the machine learning algorithm being deployed, the function can be used to achieve variable selection in either of three ways: by specifying the maximum number of best variables to return, by forward selection, and by backward elimination. In order to select the best subset of a particular size, the number of desired variables is specified in the nvmax argument as illustrated in the following syntax.

> bestSubset = *regsubsets*($y \sim X_1 + X_2 + \cdots + X_n$, *data* = dataset, *nvmax* = number)

> summary(bestSubset)

The syntax above selects the variables considered best by the regsubsets function and assigns the same to the user-defined variable, bestSubset. The y in the syntax represents class variable; X_1, $X_2,\ldots,X_n$ are the explanatory variables, while dataset is the name of the data frame holding the experimental dataset. Another alternative is to deploy the regsubsets function to perform forward selection or backward elimination in selecting subsets according to their importance, using the following syntax.

> bestSubset = *regsubsets*(y : $X_1 + X_2 + \cdots + X_n$, *data* = dataset, *method* =
“forward”)

> summary(bestSubset)

This syntax applies to forward selection and is substantially the same with that for backward elimination, except that “forward” is replaced with “backward” in the method argument. When this is executed, the function selects and returns variables considered the best for modeling. It should be noted that the user has no control over the number of variables the function will return. Aside selecting the best subsets, regsubsets ranks selected variables according to importance by indicating against each variable one or more asterisks. The more the number of asterisks assigned to a variable, the better the attribute.

2.4.2 Variable Importance Measure (varImp)

The R language consists of another package, known as caret, for Classification And REgression Training (Kuhn, et al., 2017). One of the functions within the caret package is the varImp (variable importance), which implements variable importance ranking for different machine learning algorithms, such as Logistic regression and Random Forest. To evaluate the importance of variables in a Random Forest model using varImp, the importance of a predictor variable X_j, $j = 1,...,n$ is calculated on the out-of-bag (OOB) data sample for each tree that was not used for tree construction. Initially, the predictive accuracy of the OOB sample is evaluated. Then, the values of X_j in the OOB are permuted; keeping all other predictor variables unchanged. The predictive accuracy of the shuffled data values is also measured and the mean predictive accuracy across all trees is reported. By doing so, the importance of a variable in predicting the response is quantified by evaluating the difference of how much including or excluding that variable decreases or increases accuracy (Liaw & Weiner, 2002; Strobl, Boulesteix, Zeileis, & Hothorn, 2007; Kuhn et al., 2017). This difference is referred to as the Mean Decrease Accuracy (MDA), and is computed by the formula shown in Equation (8) (Wang, Yang, & Luo, 2016; Hur, Ihm, & Park, 2017).

$$I(X_j) = MDA(X_j) = \frac{1}{n}\sum_{t=1}^{n}\frac{\sum_{i \in OOB} I(y_i = b(X_i)) - \sum_{i \in OOB} I(y_i = a(X_i^j))}{|OOB|} \tag{8}$$

where n is the total number of trees and t is a particular tree, $t = 1,...,n$. In Equation (8), $y_i = b(X_i)$ is the predictive accuracy for OOB instance X_i before permuting X_j and $y_i = a(X_i^j)$ is the predictive accuracy for OOB instance X_i after permuting X_j, while $|OOB|$ is the number of data samples not used in tree construction. In the case of Logistic regression models, the varImp function evaluates the importance of a predictor variable using the absolute value of the t-statistic for that predictor.

2.5 Modeling Tools

The tools relevant to this dissertation will be defined in this section. These include the statistical techniques that will be deployed for design of the proposed algorithm and investigation of the psychological capital dataset. Review of classification tools and methods to be adopted in the Methodology will also be conducted.

2.5.1 Risk Ratios and Odds Ratios

The formal definitions of risk ratios and odds ratios are given as: Let t_{11}= total data points where $X = 1$ and $Y = 1$, t_{10} = total data points where $X = 1$ and $Y = 0$, t_{01}= total data points where $X = 0$ and $Y = 1$, and t_{00} = total data points where $X = 0$ and $Y = 0$ for a binary independent variable X and a binary dependent variable Y. Then, the risk ratio is given by

$$RR = \frac{t_{11} / (t_{11} + t_{10})}{t_{01} / (t_{01} + t_{00})} = \left(\frac{t_{11}}{t_{11} + t_{10}} \right) \times \left(\frac{t_{01} + t_{00}}{t_{01}} \right) \tag{9}$$

(Last & Wilson, 2004; Schmidt & Kohlmann, 2008; Andrade, 2015); and the odds ratios is given by

$$OR = \frac{t_{00} \times t_{11}}{t_{01} \times \mathrm{t}_{10}} \tag{10}$$

(Szumilas, 2010; Hancock & Kent, 2016). The definitions in Equations (9) and (10) are represented in tabular form as shown in Table 1.

Table 1: Tabular definition of RR and OR

	$Y = 1$	$Y = 0$	**Total**
$X = 1$	t_{11}	t_{10}	$t_{11} + t_{10}$
$X = 0$	t_{01}	t_{00}	$t_{01} + t_{00}$

For RR, the independent variable, X is referred to as the exposure, with 0 and 1 as the unexposed and exposed, respectively. On the other hand, the dependent variable, Y is referred to as the incidence or risk of an event among the various exposure groups, with 0 and 1 representing event failure and success, respectively (Last & Wilson, 2004). Relative risk measures the ratio of the incidence of an event among data points within the exposed group compared with the incidence of that same event in the unexposed group (Schmidt & Kohlmann, 2008; Andrade, 2015). Exposure in this context could be any criterion of measurement by which data is generated. In both RR and OR, possible values range from 0 to infinity, where $RR/OR = 1$ signifies that no association exists between X and Y, $RR/OR < 1$ indicates a negative association between X and Y, and $RR/OR > 1$ shows that X and Y are positively associated (McNutt, Wu, Xue, & Hafner, 2003; Szumilas, 2010; Pandis, 2012; Andrade, 2015; Hancock & Kent, 2016).

The OR is a useful tool for evaluating the strength of association between binary variables across two different groups (Grimes & Schulz, 2008). This measure has been applied in many studies. One of these is the study by Jin, Chen, & Wang (2018) who examined how OR can be used to detect Differential Item Functioning (DIF) when conducting some tests with experimental datasets. The study, which compared the performance of OR with two other approaches: logistic regression and Mantel-Haenszel methods, showed that OR has better tendency to control false positive rates than the other methods when there is high percentage of DIF items in favour of particular groups within the dataset. VanderWeele & Vansteelandt (2010) used odds ratios for mediation analysis in epidemiology when the outcome is dichotomous. The role of a mediator variable between an exposure, outcome, and covariates was analyzed using odds ratios. The research further proposed a technique for estimating the direct and indirect effects of a mediator in the interaction between the exposure and the outcome, using odds ratios. Odds ratios are usually estimated on binary data.

If the OR technique is to be applied on continuous data, the data must first be dichotomized, which leads to information loss and reduction in precision of inference. In order to check this shortcoming, Sroka & Nagaraja (2018) proposed a method that uses the log odds link function to directly analyze continuous data without first dichotomizing the same. The experiment deployed three distributions for count data, namely; geometric, Poisson, and negative binomial, to prove that better precision of OR estimate could be obtained even on continuous data. Chen, Cohen, & Chen (2007) showed in their research that if age as a variable is dichotomized, a biased OR result will emerge. In the experiment results, the authors reported that when age was left as a continuous variable, including it as a confounder between causes of risks and outcomes produced good OR results. However, when age was converted to categorical data, the resulting estimate was biased. The study concluded that if it is necessary that age must be dichotomized, researchers should be cautious about choosing cut-points based on the size of empirical OR. Tamhane, Westfall, Burkholder & Cutter (2016) compared and contrasted odds ratios versus prevalence ratios (PR), both of which are measures of association between independent and dependent variables. The study examined the weaknesses and strengths of both measures and it turned out that OR usually overestimates strength of association compared to PR. However, OR unlike PR, has the desirable property of reciprocity, which allows for computation of the OR for group 2 by simply taking the reciprocal of the OR for group 1.

The validity of OR estimate is evaluated using confidence interval (CI) and p-values (Park, 2013). The CI is calculated by first evaluating the standard error of the log odds ratio $(SE(\log \text{OR}))$ as shown in equation (11)

$$SE(\log \mathrm{OR}) = Sqrt\left(\frac{1}{t_{00}} + \frac{1}{t_{01}} + \frac{1}{t_{10}} + \frac{1}{t_{11}}\right) \tag{11}$$

The mostly used CI is the 95%, which according to Bland & Altman (2000) is calculated as follows:

$$\left.\begin{aligned} LowerLimit &= exp[\log(\mathrm{OR}) - 1.96(SE(\log \mathrm{OR}))] \\ UpperLimit &= exp[\log(\mathrm{OR}) + 1.96(SE(\log \mathrm{OR}))] \end{aligned}\right\} \tag{12}$$

An interval that excludes the null value, 1, is statistically significant. Furthermore, if the lower and upper limits of the confidence intervals produce p -value less than 0.05, OR result is again said to be statistically significant.

2.5.2 Classification Tools

Logistic regression is a modeling tool used in examining the association between a categorical dependent variable and one or more independent variables of a set of observations (Stolzfus, 2011). This regression type is anchored on the logistic function where values must lie between 0 and 1, corresponding to class labels (Kleinbaum & Klei, 2010). The probabilities indicating the possibility of an observation belonging to a certain class are modeled using the logistic equation, $\log\left(\frac{P}{1-P}\right) = b_0 + b_1 X_1 + ... + b_n X_n$ where P is the probability of success, $P/(1-P)$ is the odds, b_0 is the intercept, $b_1,...,b_n$ are parameter estimates, and $X_1,...,X_n$ are data values corresponding to each independent variable (Liu, Li, & Liang, 2014; Sperandei, 2014). Depending on the threshold under consideration, the value of the logistic model determines whether an observation belongs to class 0 or class 1. Meanwhile, Random forest is a machine learning tool that combines several tree predictors$\{ h(X, v_k), k = 1,...,\mathrm{n} \}$, where X is an input vector and $\{v_k\}$ are independent random vectors within the same distribution across all trees in the forest

(Breiman, 2001; Genuer et al., 2010). In order to determine the class of an input vector X, each tree casts a single vote and the class with more votes is selected (Breiman, 2001).

The predictive accuracy and the balanced classification accuracy (BCA) of binary classifiers are defined by

$$\text{Accuracy} = \frac{T^{+} + T^{-}}{T^{+} + T^{-} + F^{+} + F^{-}} \tag{13}$$

and

$$\text{BCA} = 0.5 \times \left(\frac{T^{+}}{T^{+} + F^{-}} + \frac{T^{-}}{T^{-} + F^{+}} \right) \tag{14}$$

respectively, where, T^{+} is the number of correctly classified observations in class 1, T^{-} is the number of correctly classified observations in class 0, F^{+} is the number of observations in class 0 but wrongly classified in class 1, and F^{-} is the number of observations in class 1 but wrongly classified in class 0 (Catena et al., 2014). The four quantities in Equations (13) and (14) are represented in a confusion matrix as shown in Table 2.

Table 2: Confusion matrix

	Actual = 0	Actual = 1
Predicted = 0	T^{-}	F^{-}
Predicted = 1	F^{+}	T^{+}

The formula in Equation (14) is the appropriate measure of classifier accuracy while dealing with imbalanced datasets; that is, when the number of observations in class 0 and class 1 are not the same (Cateni et al., 2014; Tharwat, Moemen, & Hassanien, 2017). According to Tharwat (2014), other metrics useful in examining various aspects of predictive accuracy of classifiers on imbalanced datasets can be calculated from the confusion matrix. These metrics are considered below:

- Sensitivity. Also referred to as True Positive Rate (TPR) or recall, is the ratio of correctly predicted samples in class 1 to the total number of samples in class 1. This is given by $TPR = T^{+} / (T^{+} + F^{-})$. In other words, it evaluates how often the classifier correctly classifies observations that are actually in class 1 as being in class 1.
- False Positive Rate (FPR). This is calculated as $FPR = F^{+} / (F^{+} + T^{-})$, and it shows how often the classification model wrongly predicts an observation that is originally in class 0 as being in class 1.
- Specificity. Also referred to as True Negative Rate (TNR) or inverse recall, is the ratio of correctly predicted samples in class 0 to the total number of samples in class 0. This is calculated as $TNR = T^{-} / (T^{-} + F^{+})$, and it shows how often the model classifies an observation that is truly in class 0 and being in class 0.

- False Negative Rate (*FNR*). This evaluates how often the classification model classifies an observation that is actually in class 1 as being in class 0. It is given by $FNR = F^{-} / (F^{-} + T^{+})$.

These four metrics, together with Equation (14), are insensitive to class distributions in the dataset. This means that irrespective of whether the number of observations across class 0 and class 1 are unequal, these metrics will produce unbiased accuracy estimates.

2.6 Psychological Capital

2.6.1 Psychological Capital Explained

Psychological capital (PsyCap) is a measure of the positive capabilities of an individual which consists of four components: hope, efficacy, resilience, and optimism (Antunes, et al., 2017; Luthans et al., 2007). Efficacy is the psychological quality of an individual that is reflected in their ability to carry out pending tasks diligently within a given time frame (Luthans & Youssef, 2017; Stajkovic & Luthans, 1998). Employees with positive efficacy exhibit the initiative to source for resources to achieve set goals within time limits (Breevaart, Bakker &, Demerouti, 2014). According to Kobau et al. (2011), optimism is a quality that associates positive events with personal and permanent causes, while interpreting negative events as external, temporary, and contextual. It enables an individual to view things in a positive light. Optimistic individuals are flexible and pragmatic, always focusing on positive outcomes while pursuing desired goals (Carver & Scheier, 2002; Alarcon, Bowling, & Khazon, 2013). Hope, in the opinion of Snyder et al. (1991), is a state that motivates wishful thinking geared towards successfully achieving a desired goal. Hope has a distinctive feature from other components of PsyCap because of its planned path and the desire to achieve set goals with a positive mindset. Resilience is the ability of an individual to recover from, or adapt to stress and adversity which could arise from family, workplace, financial or relationship problems (Lee & Chu, 2016). Positive resilience enables individuals to persevere in unfavourable conditions, which ultimately affect job performance and organizational outcomes (Tugade et al., 2004). Collectively, these four components constitute the psychological capital and have the positive effect on an individual, manifesting in dedication to duty, job performance, job satisfaction, and self-development (Avey, Luthans, Smith, & Palmer, 2010).

The commonly used tool for measuring psychological capital is the Psychological Capital Questionnaire (PCQ) (Luthans et al., 2007). The PCQ is a 6-point scale questionnaire, consisting of 24 items which employees utilize to supply self-information that assist in assessing their PsyCap. This measure has been validated by Luthans et al. (2015) as an authentic tool, and has been adopted in many climes for psychological capital measurement (Antunes et al., 2017).

2.6.2 Psychological Capital and the Work Environment

Several studies have been conducted to evaluate the relationship between employees' psychological capital and the work environment. One of such is the research by Simons & Buitendach (2013) which was undertaken to determine the relationship between employees' psychological capital and their commitment to the organization they work for. The study used the PCQ, demographic, work engagement and organizational commitment questionnaires to collect data on a sample of 106 call center employees in South Africa. The result showed that positive relationship exists between PsyCap, work engagement and organizational commitment. That is, employees with positive psychological capital are well committed and positively attached to the organization they work for.

Another research by Leon-Perez, Antino, & Leon-Rubio (2016) found out that there exists a negative relationship between psychological capital and burnout on the one hand, and a positive relationship between psychological capital and quality of service on the other hand. The study was conducted on a sample of 798 workers in a Spanish vehicle safety and emissions inspection company. The psychological capital of the respondents was measured using the Psychological Capital Questionnaire, Quality of Service (QoS) was assessed using the QoS questionnaire, while burnout, also referred to as stress, was measured using the questionnaire developed by Shirom-Melamed (Leon-Perez et al., 2016). The research concluded that positive psychological capital among employees has high prospects in stress reduction and improved quality of service, which positively affects productivity.

In a related research, Avey et al. (2010) investigated the relationship between positive psychological capital and employee's well-being over time. The study was conducted on 280 participants drawn from a Midwestern university. Their psychological capital was measured using the PCQ; their psychological well-being was measured using the General Health Questionnaire and the Index of Psychological Well-Being. The research findings indicated that positive PsyCap enhances employee well-being, and psychological well-being has a proportionate effect on job satisfaction.

Luo, Wang, & Yi (2017) researched on the role psychological capital plays between corporate culture and job performance. The survey, conducted on 377 workers of a petrochemical company, revealed that corporate culture has a positive effect on the four components of psychological capital; and psychological capital has prospects to intermediate between corporate culture and job performance among employees. A study conducted by Durrah, Alhamoud, & Khan (2016) investigated the relationship between PsyCap and job performance on the one hand, and the mediating role of job satisfaction among PsyCap and job performance. The research utilized the

Psychological Capital Questionnaire, job performance and job satisfaction questionnaires to collect data on 110 instructors from the Philadelphia University. The result showed that positive PsyCap is positively related to job performance, and job satisfaction mediates the relationship between PsyCap and job performance. In another survey, Sun, Zhao, Yang, & Fan (2011) investigated how psychological capital impacts on job embeddedness and performance. A sample size of 1000 nurses in a university hospital in China provided the research data; and the findings indicated a strong association between psychological capital, job embeddedness and work performance. The research suggested that by improving the psychological capital of nurses, their willingness to remain in the current work place as well as their job output will be greatly enhanced.

The reviewed PsyCap literature suggests that several studies have been conducted on psychological capital and various aspects of work engagement. However, we have not come across any existing literature on the relationship between psychological capital and educational qualification, or psychological capital and organizational tenure. This identified gap will be addressed in this dissertation.

Chapter 3

PROPOSED FILTER VARIABLE SELECTION ALGORITHM

3.1 Chapter Over

In this chapter, the design methodology of the proposed filter variable selection algorithm is presented. Experiments are conducted to evaluate the performance of the proposed algorithm in comparison with some existing variable selection algorithms; results are presented and discussed. Section 3.2 presents the experimental datasets to be used in the experiments, while Section 3.3 covers the design of the proposed algorithm. Section 3.4 reports on experiment procedures and results are shown. In Section 3.5, the results obtained in the experiments are discussed.

3.2 Experimental Datasets

A number of datasets, mostly from the healthcare domain, were deployed to demonstrate the effectiveness of the proposed algorithm in feature selection. Since the proposed algorithm places emphasis on a dichotomous response, each experimental dataset considered in this experiment has a binary outcome. The considered datasets are listed below:

- Psychological Capital (PsyCap). This dataset carries psychological capital (PsyCap) information of some workers in the hospitality industry. Each worker’s PsyCap was assessed on the four components of psychological capital (hope, efficacy, resilience, and optimism), using the questionnaire presented in (Paek, Schuckert, Kim, & Lee, 2015). The dataset has a binary class variable, where 0 and 1 represent negative and positive PsyCap, respectively. The detailed dataset variables are presented in Table A.1.

- Diabetes in Pima Indian Women (Diabetes). The dataset consists of 332 observations about diabetes test results of Indian women of Pima indigene. The population sample was those from 21 years and above, residing in Arizona. This dataset, accessible through the R language “MASS” package, reported in Venables & Ripley (2002), is named Pimat.te within the package, and was originally sourced from (Smith, Everhart, Dickson, Knowler, & Johannes, 1998). The dataset has a binary response variable named “type”, where 0 and 1 signify non-diabetic and diabetic, respectively. Details of the dataset variables are in Table A.2.
- Survival from Malignant Melanoma (Melanoma). This dataset, available in the R package “boot”, records information on the survival of patients from malignant melanoma (Canty & Ripley, 2017). The patients had surgery at the Department of Plastic Surgery of the University Hospital, Odense, Denmark, between 1962 and 1977. Several measurements were taken and reported as predictor variables, with a binary class “ulcer”, where 1 indicates an ulcerated tumour and 0, non-ulcerated. Find detailed dataset variables in Table A.3.
- Spam E-mail Data (Spam). The dataset consists of e-mail items with measurements relating to total length of words written in capital letters, numbers of times the “$” and “!” symbols occur within the e-mail, etc.; and a binary class variable, “yesno”, with 1 classifying an e-mail as spam and 0 otherwise. The dataset, titled spam7, can be accessed in the R package “DAAG” (Maindonald & Braun, 2019). Details of the dataset variables are presented in Table A.4.

- Biopsy Data of Breast Cancer Patients (Cancer). Named biopsy in the R package, “MASS” in Venables & Ripley (2002), the dataset measures the biopsies of breast tumours on a number of patients. The dataset was obtained from the University of Wisconsin Hospital, Madison, with known binary outcome named “class”, where 0 = benign and 1 = malignant. Find details of the dataset variables in Table A.5. Some characteristics of the experimental datasets are presented in Table 3.

Table 3: Properties of the experimental datasets

Dataset	**Predictors**	**Records**	**Class = 0**	**Class = 1**
PsyCap	20	329	68	261
Diabetes	7	332	223	109
Melanoma	6	205	115	90
Spam	6	4601	2788	1813
Cancer	9	683	444	239

3.3 Design of Proposed Algorithm

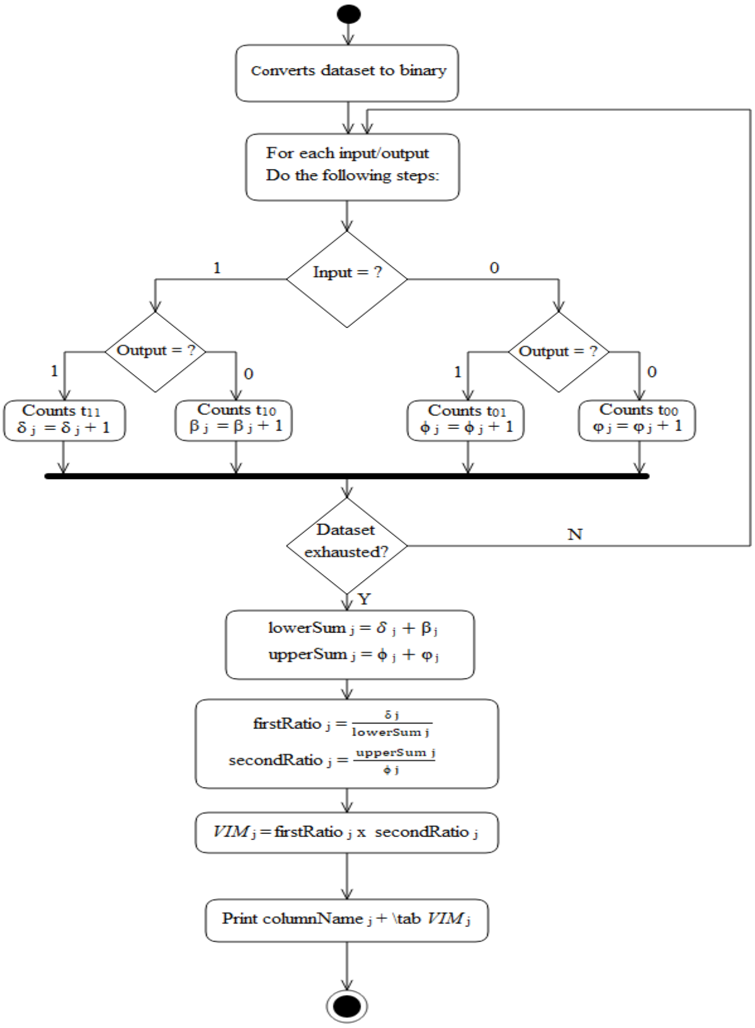
In this section, we will consider X_j, $j = 1,...,n$ as a set of predictors in a high dimensional space R^n. In most cases, especially with big data, some of these predictors are irrelevant, duplicative, and, thus, not needed in machine learning tasks (Chen, Mao, & Li, 2014). Usually, the objective is to reduce the number of predictors to k where $k < n$, such that k consists of the most relevant explanatory variables needed in classification. This is the objective the proposed algorithm seeks to achieve.

Let *RawData* denote a dataset consisting m observations, n predictors, and the outcome variable y_i. Let $RawData[i, j]$ denote a data point at row i, column j where $i = 1,...,m$ and $j = 1,...,n$. This algorithm will require a normalized dataset on the interval [0, 1], also referred to as min-max normalization (Pandey & Jain, 2017; Jain, Shukla, & Wadhvani, 2018). The proposed algorithm, presented in Appendix B, will take the following steps:

- The first step of the proposed algorithm, as presented in Listing 1, is to binarize the dataset. It is a requirement that both independent and dependent variables carry only binary values for the risk ratio measure to be deployed. On purpose, we did not design the algorithm to print the output of the binary dataset. This is to guard against users inadvertently using the binary dataset for model construction. The binary data is only useful for RR computation, after which classification models are fit on the original dataset.
- In the second step, listed in Listing 2, the algorithm counts, for each predictor X and the class Y, all occurrences where $(X, Y) = (1,1), (1,0), (0,1)$ and $(0,0)$. Just as in step 1, these computations are kept behind the scene, without printing any output visible to the user.
- The third step, listed in Listing 3, applies the risk ratio formula of Equation (9) on the values computed in Listing 2 to produce the variable importance rankings. This algorithm outputs the importance rankings of the variables in the order the predictors appear in the dataset. For a better view of the results, the user may decide to arrange the output in ascending or descending order. It is upon the judgment of the modeler to determine the cutoff point of those variables to include in a model.

- The statement on line 44 of the algorithm, in Appendix B, will output the names of predictor variables and their RR values, separated by a tab, each on a separate line. Each RR value constitutes the importance rank of the corresponding predictor, signifying the extent to which it is associated with the class.

The processes involved in feature ranking by the proposed algorithm are shown in Figure 1 and the pseudo code below.



Converts dataset to binary
For each input/output Do the following steps:
Input = ?
1
0
Output = ?
1
0
Output = ?
1
0
Counts t_{11} $\delta_j = \delta_j + 1$
Counts t_{10} $\beta_j = \beta_j + 1$
Counts t_{01} $\phi_j = \phi_j + 1$
Counts t_{00} $\varphi_j = \varphi_j + 1$
Dataset exhausted?
N
Y
$\text{lowerSum}_j = \delta_j + \beta_j$
$\text{upperSum}_j = \phi_j + \varphi_j$
$\text{firstRatio}_j = \frac{\delta_j}{\text{lowerSum}_j}$
$\text{secondRatio}_j = \frac{\text{upperSum}_j}{\phi_j}$
$VIM_j = \text{firstRatio}_j \text{ x } \text{secondRatio}_j$
Print columnName_j + \tab VIM_j

Figure 1: Activity diagram of the proposed algorithm

Algorithm 1: Pseudo code

```
START
Convert dataset to binary, that is, round all values < 0.5 to 0 and
> = 0.5 to 1
FOR each input/output, DO the following:
IF INPUT is 1
AND OUTPUT is 1 THEN
Count $t_{11}$ that is $d_j = d_j + 1$
ELSE Count $t_{10}$, that is $b_j = b_j + 1$
END IF
IF INPUT is 0
AND OUTPUT is 1 THEN
Count $t_{01}$ that is $f_j = f_j + 1$
ELSE Count $t_{00}$, that is $j_j = j_j + 1$
END IF
NEXT input/output
IF All input/output are exhausted, compute the following:
FOR each variable $j = 1$ to $n$
        $\text{lowerSum}_j = d_j + b_j$
        $\text{UpperSum}_j = f_j + j_j$
        $\text{firstRatio}_j = \frac{d_j}{\text{lowerSum}_j}$
        $\text{secondRatio}_j = \frac{\text{upperSum}_j}{f_j}$
         $VIM_j = \text{firstRatio}_j ´ \text{secondRatio}_j$
          PRINT $\text{columnName}_j$ and space, and $\text{VIM}_j$
NEXT variable
```

```
STOP
```

A higher value of *VIM* for a predictor signifies strong association with the class, and consequently indicates its importance in classification. This algorithm is summarized in Equation (15).

$$VIM_j = \left(\frac{\sum_{i=1,j=1}^{m,n} \delta ij}{\sum_{i=1,j=1}^{m,n} \delta ij + \sum_{i=1,j=1}^{m,n} \beta ij} \right) \times \left(\frac{\sum_{i=1,j=1}^{m,n} \phi ij + \sum_{i=1,j=1}^{m,n} \varphi ij}{\sum_{i=1,j=1}^{m,n} \phi ij} \right) \tag{15}$$

where VIM_j is the importance ranking of the jth predictor, $j=1,...,n$, d_{ij} is the total number of observations with *input* = 1 and *output* = 1, β_{ij} is the total number of observations with *input* = 1 and *output* = 0, ϕ_{ij} is the total number of observations with *input* = 0 and *output* = 1, and φ_{ij} is the total number of observations with *input* = 0 and *output* = 0 (Bodur & Atsa’am, 2019).

Worst-case Computational Time Complexity. The worst case time complexity of the proposed algorithm is given as follows. In Listing 1, the algorithm requires two loops to scan through an $n \times m$ dataset to binarize the data points. The time complexity of Listing 1 is therefore, $O(n \times m)$. Listing 2 requires three loops to scan through the dataset in order to count $t_{11}, t_{10}, t_{01}, t_{00}$. The time complexity for Listing 2 is therefore, $O(n \times m^2)$. Listing 3 requires one loop to compute RR, its time complexity is $O(n)$. Combining the time complexities of Listings 1, 2 and 3, we have $O(n \times m) + O(n \times m^2) + O(n) = O(n \times m^2)$. Therefore, the worst case computational time complexity of the proposed algorithm is $O(n \times m^2)$.

3.4 Experiment and Results

3.4.1 Execution of the Proposed Algorithm on the Datasets

The proposed algorithm was executed on all the datasets in order to rank the variables according to importance. The existing varImp function and the regsubsets methods (nvmax, forward, backward) were also deployed to rank the variables. Equally, the Fisher score and Pearson's correlation were deployed. This was done in order to compare the effectiveness of the proposed algorithm against existing methods of variable selection.

Two machine learning algorithms, namely Logistic Regression and Random Forest, were used in the experiment for model construction, evaluation of goodness of fit and predictive accuracy. Samples of variable ranking results by various algorithms on the datasets are shown in Tables 4-8. In each table, the ranking by the proposed algorithm is placed side-by-side with that of an existing algorithm. The column labeled 'Importance' in Tables 4-8 shows the extent to which the corresponding variable is useful in model construction. When comparing the importance of two variables, the variable with a higher value of importance is selected.

Table 4: Variable ranking of PsyCap dataset

Proposed Algorithm		**Pearson**	
Variable	Importance	Variable	Importance
H2	11.2302	H2	0.6235

Table 4: Variable ranking of PsyCap dataset (Continued)

Proposed Algorithm		**Pearson**	
Variable	Importance	Variable	Importance
S4	4.7628	S4	0.5322
H1	3.7997	H3	0.5099
S2	2.8553	S2	0.4897
S1	2.4419	H1	0.4597
⋮	⋮	⋮	⋮
O1	1.0942	O4	0.0940
R1	0.7726	R1	-0.0826

Table 5: Variable ranking of Diabetes dataset

Proposed Algorithm		**varImp**	
Variable	Importance	Variable	Importance
Glu	3.8438	Glu	43.4587
Npreg	3.7508	Age	21.0085
Bmi	3.5803	Bmi	16.4553
Ped	3.4098	Skin	10.4756
Age	2.4550	Npreg	7.9362
Skin	2.0050	Pped	7.1898
Bp	1.2672	Bp	-1.6153

Table 6: Ranking of Melanoma dataset

Proposed Algorithm		**Fisher Score**	
Variable	Importance	Variable	Importance
thickness	2.5556	Thickness	0.4200
sex	1.5262	Time	0.1500
age	1.2159	Status	0.1500
year	1.1561	Sex	0.0580
status	0.6324	Age	0.0320
time	0.5242	Year	0.0022

Table 7: Variable ranking of Spam dataset

Proposed Algorithm		**Backward Elimination**	
Variable	Importance	Variable	Importance
n000	16.9310	n000	******
dollar	10.7743	Dollar	*****
crl.tot	7.6959	Bang	****
money	3.8445	Money	***
bang	0.7689	crl.tot	**
make	0.6990	Make	*

Table 8: Variable ranking of Cancer dataset

Proposed Algorithm		**Forward Selection**	
Variable	Importance	Variable	Importance
V2	87.9331	V6	********

Table 8: Variable ranking of Cancer dataset (Continued)

Proposed Algorithm		**Forward Selection**	
Variable	Importance	Variable	Importance
V3	54.2460	V2	*******
V4	52.9456	V1	******
V6	52.0167	V8	*****
V7	35.0317	V7	****
V5	30.6527	V3	***
V9	29.7238	V5	**
V8	26.2148	V4	*
V1	15.1406	V9	

The Table 4 shows variable importance ranking results of PsyCap dataset based on the proposed algorithm and Pearson's correlation. The importance values of the proposed algorithm are computed as risk ratios of the association between a predictor variable and the outcome. In Table 5, the variables of the Diabetes dataset are ranked according to the proposed algorithm and the varImp method. The Table 6 presents importance ranking of Melanoma dataset variables by the proposed algorithm and the Fisher score, while Table 7 shows the Spam dataset variable importance ranking using the proposed algorithm and backward elimination. In Table 8, the variable importance ranking results of Cancer dataset using the proposed algorithm and forward selection method are reported.

Performance evaluation of the proposed algorithm in comparison with existing algorithms was done in two steps. First, the goodness of fit of models developed using variables selected by the new algorithm and existing ones was examined, and secondly, the predictive accuracy evaluation was carried out.

3.4.2 Goodness of Fit Evaluation

The goodness of fit test was assessed on two metrics: deviance and Mean Squared Error (MSE). In Logistic regression models, two deviance types are reported: null deviance and residual deviance (Chapela, 2013). The residual deviance is calculated cumulatively as predictors are added to the model. The difference between the final residual deviance and the null deviance explains the goodness of fit of a model. When comparing two models, the model with the smallest deviance is said to have better fit. The MSE is a parameter-free measure that gives information on the difference between actual and predicted values (Wang & Boyik, 2009). Lower values of MSE for a model indicate better fit. A sample result of the goodness of fit test of the various models is presented in Table 9.

The goodness of fit results presented in Table 9 show that the subsets selected by the proposed algorithm competed favorably with those selected by the existing varImp algorithm. In Table 9, the Subset Size column shows the number of variables that were selected in each dataset; the Deviance and MSE columns give values that were obtained from subsets selected by the proposed algorithm and the existing varImp algorithm for each of the five datasets.

Table 9: Goodness of fit evaluation

varImp				**Proposed Algorithm**		
Dataset	Subset Size	Deviance	MSE	Subset Size	Deviance	MSE
PsyCap	8	231.9	1.4	8	204.1	1.2

Table 9: Goodness of fit evaluation (Continued)

varImp				**Proposed Algorithm**		
Dataset	Subset Size	Deviance	MSE	Subset Size	Deviance	MSE
Diabetes	3	87.5	0.83	3	92.7	0.73
Melanoma	5	49	1.2	5	37.3	1.1
Spam	3	1017.7	1.3	3	929	1.4
Cancer	5	638.4	1.7	5	631.4	1.1

3.4.3 Predictive Accuracy Evaluation

The results of the predictive accuracy test of models constructed with subsets selected by the proposed algorithm compared with those constructed with variables selected by existing algorithms were examined. Before fitting the models, each dataset was split into 80% and 20% train and test sets, respectively. The train sets were used for model construction, while the test sets were used to evaluate the predictive power of the models. Typically, the predictive accuracy is computed using the Equation (13). However, the Equation (13) assumes that classes of the dataset are balanced. This is usually not the case in real life as could be seen in Table 3, where the number of observations in class 0 is not same as that in class 1 across all experimental datasets. For imbalanced datasets, the balanced classification accuracy (BCA) defined in the Equation (14) is applied to calculate predictive accuracy. In this dissertation, the BCA was used throughout the experiments for predictive accuracy.

The proposed algorithm was executed on all the datasets to obtain importance rankings of predictor variables. After generating the rankings, the best subsets were selected for modeling using Random Forest and Logistic regression classification. Two criteria were adopted in arriving at best subsets. The first option was to sequentially select all variables with ranking values close to each other until there is an unusual decline with subsequent variables down the group. The second option was to keep adding variables with reasonably high ranking values until further additions do not improve model performance. Existing ranking algorithms, namely regsubsets (nvmax, forward, and backward), varImp, Fisher score, and Pearson's, were equally executed on the datasets. The best subsets generated by these algorithms were selected for modeling. The balanced classification accuracy of each model was computed on the test sets of the datasets. The performance evaluation process is summarized in Figure 2.

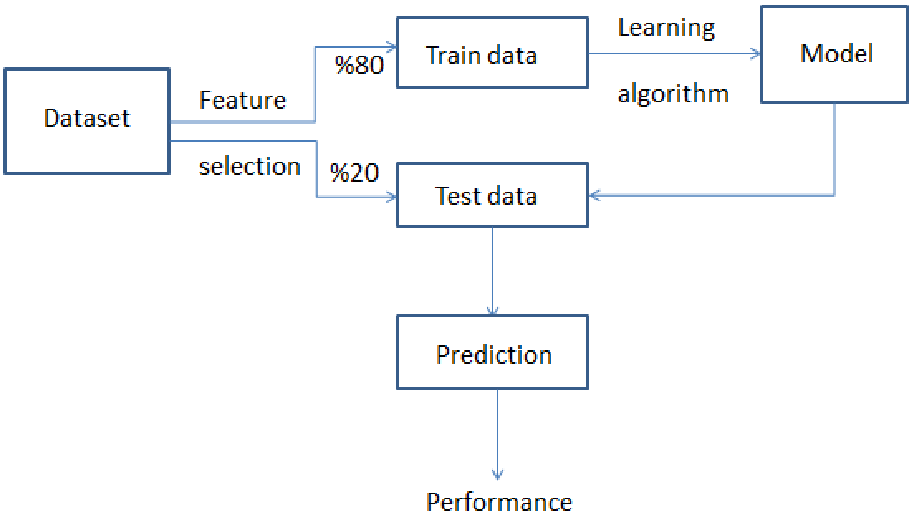


Figure 2: Performance evaluation process

An example of how the BCA was computed from actual values of the PsyCap dataset and predicted values by a Random Forest model, using subset selected by the proposed algorithm is shown below. The Table 10 shows prediction results on the 66 data samples in the test set of the PsyCap dataset.

Table 10: Prediction results of PsyCap Dataset

	Actual = 0	Actual = 1
Predicted = 0	10	3
Predicted =1	2	51

On Table 10, the definition of confusion matrix in Table 2 and the BCA formula in the Equation (14) were applied to calculate BCA as:

$$BCA = 0.5 \times \left[\frac{51}{51+3} + \frac{10}{10+2} \right] = 0.5 \times \left[\frac{51}{54} + \frac{10}{12} \right]$$

$$\approx 0.5 \times [0.9444 + 0.8333] \approx 0.5 \times [1.7777] \approx 0.89 \approx 89\%$$

Similar computations were carried out on all prediction results generated by various models in the experiments, yielding the performance accuracies presented in Tables 11-15. The Subset size row represents the number of variables suggested by the each algorithm as the best for modeling, the BCA row shows the balanced accuracy. The Table 11 indicates the predictive accuracy comparison of various ranking algorithms on the PsyCap dataset, while Table 12 reports results of predictive accuracies on the Diabetes dataset. Relatedly, Table 13 reports various predictive accuracies generated by each ranking algorithm on the Melanoma dataset, while Table 14 presents accuracy results on the Spam dataset. In Table 15, the predictive accuracies on the Cancer dataset for the various ranking algorithms are reported. In Tables 11-15 and in Figures 3-4, the abbreviations F, B, N, V, F.S., P and P.A. mean forward selection, backward elimination, Nvmax, varlmp, Fisher score, Pearson`s Correlation and Proposed Algorithm, respectively.

Table 11: Ranking methods performance comparison on the PsyCap dataset using Random forest

Ranking Method	**F**	**B**	**N**	**V**	**F.S.**	**P**	**P.A.**
Subset size	8	8	5	5	6	8	5
BCA (%)	83	83	88	82	88	87	89

Table 12: Ranking methods performance comparison on the Diabetes dataset using Logistic regression

Ranking Method	**F**	**B**	**N**	**V**	**F.S.**	**P**	**P.A.**
Subset size	3	3	3	5	4	2	3
BCA (%)	74	80	81	82	77	77	83

Table 13: Ranking methods performance comparison on the Melanoma dataset using Logistic regression

Ranking Method	**F**	**B**	**N**	**V**	**F.S.**	**P**	**P.A.**
Subset size	4	4	4	5	3	3	5
BCA (%)	71	71	70	66	66	72	73

Table 14: Ranking methods performance comparison on the Spam dataset using Logistic regression

Ranking Method	**F**	**B**	**N**	**V**	**F.S.**	**P**	**P.A.**
Subset size	4	4	5	3	2	2	3
BCA (%)	68	68	72	72	71	71	71

Table 15: Ranking methods performance comparison on the Cancer dataset using Logistic regression

Ranking Method	**F**	**B**	**N**	**V**	**F.S.**	**P**	**P.A.**
Subset size	5	5	4	4	3	3	5
BCA (%)	97	97	92	91	97	97	98

As could be observed in Tables 11-15, the variable subsets selected by the proposed algorithm performed competitively with the selection by existing algorithms.

3.5 Discussion

A predictive accuracy test was conducted in order to determine how well the variables selected by both existing algorithms and the proposed algorithm can predict the outcome variable Y on the validation set. Output was determined as probabilities of the form $P(Y=1|X_i),$ where X_i is the data value for each predictor, $i = 1,...,n$. The boundary used in making a decision was 0.5. What the program typically did was that, if $P(Y=1|X_i)>0.5$ then $Y=1$ otherwise, $Y=0$. When this test was run on the different models generated by the various ranking algorithms, their respective predictive accuracies were obtained using Equation (14). The Figures 3 and 4 represent graphically that variables selected by the proposed algorithm in all datasets produced higher predictive accuracies, except in one instance, compared with the selections by the existing algorithms. Therefore, the new algorithm can be said to be a good choice of filter variable ranking in machine learning classification.

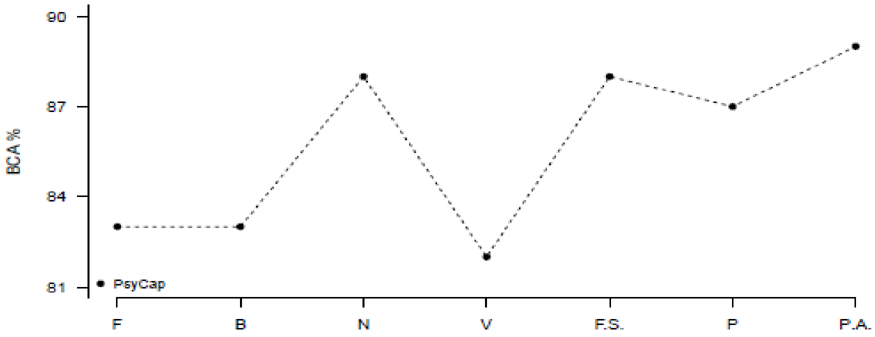


Figure 3: Balanced Classification Accuracy (BCA) results of the ranking algorithms on the PsyCap dataset

Apart from selecting variable subsets that resulted in good model performance, another plus for the proposed algorithm over the existing algorithms is the way ranking values are presented to the user. As could be observed in Table 4 and Table 5, one ranking value in each of these tables is a negative number. The negative ranking values were generated by the Pearson’s and varImp methods. It is important to point out that the proposed algorithm will not generate a negative value since RR values range from 0 to infinity.

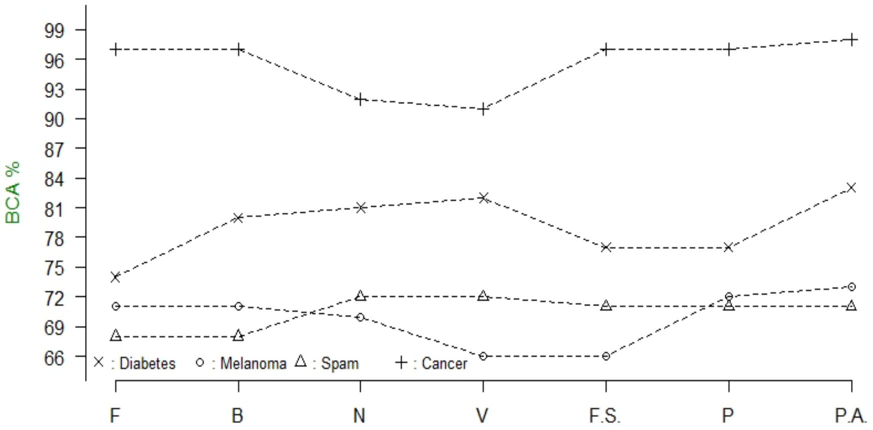


Figure 4: BCA results of the ranking algorithms

For quick insights into how much one predictor is more or less important than another, it would be better for all values to carry the same sign across the board. The

Pearson’s is a correlation-based method, which means all ranking values produced will fall within the interval [–1, 1]. In big data, where some datasets consist of a high number of features, say 100 and above, ranking the entire feature space within this interval may not give quick visual insights. This same argument is valid for the rankings presented in Table 7 and Table 8, where asterisks are used by backward elimination and forward selection methods to represent variable importance. By using discrete asterisk to convey information about the importance of a variable, some vital information might certainly be truncated. The RR deployed in the proposed algorithm produces continuous values over the range 0 to infinity. This range seems more appropriate for representing ranking values when the feature space is large and to curtail information loss.

Chapter 4

KNOWLEDGE DISCOVERY ON EMPLOYEES PSYCHOLOGICAL CAPITAL DATASET

4.1 Chapter Over

In this chapter, the psychological capital dataset is investigated in order to discover the association between employees PsyCap and educational qualification; and employees PsyCap and organizational tenure. Section 4.2 is about the methodology adopted in the investigation. This includes presentation of the experimental dataset and various experiments conducted to mine desired knowledge. In Section 4.3, the results obtained from the experiments are presented and discussed.

4.2 Methodology

4.2.1 Experimental Dataset

Between July and September, 2018, the PCQ was distributed to 500 workers in the hospitality industry in Abuja, Lagos, and Makurdi, all in Nigeria. Reponses were rated on a 6-point scale, ranging from 1 = strongly disagree to 6 = strongly agree. The questions that form the PCQ were adopted from Paek et al. (2015) as shown in Table A.1

Information on each employee’s highest educational qualification was also collected. Each participant was required to select the highest qualification that applied to them from a range of 4 options; namely, primary, secondary, post-secondary (below

Bachelor’s), Bachelor’s degree and postgraduate.

Information relating to how long an employee has served within the current organization was equally collected. This is termed “organizational tenure”, and was categorized into four; namely, “less than one year”, “1 – 3 years”, “4 – 6 years”, and “7 years and above”. Each participant was required to choose one category that applied to their years of service in the organization. There was no requirement for prior ethics approval.

A total of 329 questionnaires had complete responses and thus, formed the experimental dataset. Among the 329 observations, 150 [46%] came from employees of some 4-star and 5-star hotels across Abuja and Lagos, 102 [31%] were from civil servants of Tourism Section, Benue State Ministry of Arts, Culture and Tourism, Makurdi; and 77 [23%] came from staff of the Elegushi Private Beach, Lagos. Demographic breakdown of the participants showed that 222 [67%] and 107[33%] were males and females respectively. A total of 300 [91%] fell between the ages of 18 to 41 years, while 29 [9%] were 42 years and above.

In terms of position held in the industry, 212 [64.4%] belonged to the lower cadre (Receptionist, Waiter, Security, etc), 96 [29.2%] belonged to the middle cadre (Supervisor, Manager, Assistant Director, etc), and 21 [6.4%] belonged to top management (Director, General Manager, etc). No respondent had stayed less than a year in the current organization [0.0%], while a total of 255 [77.51%] participants had stayed between 1 – 3 years in their organization. A total of 73 [22.19%] participants indicated that their tenure in the organization was between 4 – 6 years as at the time of completing the questionnaire. Lastly, 1 [0.3%] participant had served their current organization for a period of 7 years and above. The demographic breakdown of the employees is shown in Table 16.

Table 16: Demographic breakdown of employees

Hospitality units	**Number of employees**	**Gender**	
Hotel	150	Male	222
Beach	77	Female	107
Ministry	102		
Age		**Cadre**	
18-41	300	Lower	212
42 and above	29	Middle	96
		Top	21
Tenure (in years)			
< 1	0		
1-3	255		
4-6	73		
7 and above	1		

4.2.2 Experiments

4.2.2.1 Internal Consistency and Binary Variables

In order to confirm reliability, the Cronbach's internal consistency (Vaske, Beaman, & Sponarski, 2017) of the PsyCap, educational qualification, and organizational tenure measures were evaluated and each produced Cronbach's $\alpha = 0.86$.

Let it be recalled that the OR measure requires binary independent and dependent variables for computation. In order to binarize the dependent variable, employees with total score of 65 out of 100 were classified as having 'positive PsyCap'; while those with total scores from 0 to 64 were classified as 'negative PsyCap'. On the 6-point Likert scale of the PCQ validated by Luthans et al. (2015), 4 = 'somewhat agree'. When transformed to percentage, 4 out of 6 is equivalent to 67%.

Considering that some few questions in the PCQ are on reverse scale, a threshold of 65% was chosen to differentiate negative and positive PsyCap. Respondents possessing educational qualifications below the Bachelor’s degree were classified as ‘lower educational qualification’; while those with Bachelor’s degrees and or a postgraduate certificate were classified as ‘higher educational qualification’. This threshold was informed by the outcome of a study by Carnevale, Smith & Strohl (2010) who identified two categories of educational qualifications on the basis of wage earnings.

The study held that employees with qualifications lower than the Bachelor’s degree earn lower wages within same range, while their counterparts with Bachelor’s degrees and above earn higher pays. Two binary variables were then defined for the dataset as ‘Edu’ and ‘PsyCap’; where Edu is an independent variable and PsyCap is the dependent variable, each taking only 0 and 1 values. The 0 represents negative psychological capital and lower educational qualification, while 1 represents positive psychological capital and higher educational qualification.

Equally, the organizational tenure independent variable was dichotomized and then named ‘OrgTenure’. Workers that stayed a duration of 0 – 3 years were categorized as having ‘short OrgTenure’, while those that stayed between 4 years and above were categorized as having ‘long OrgTenure’. Therefore, 0 and 1 represent ‘short OrgTenure’ and ‘long OrgTenure’, respectively.

4.2.2.2 PsyCap and Educational Qualification

Cross tabulating the 329 observations of the PsyCap dataset based on the binary variables, PsyCap and Edu, Table 17 was formed.

Table 17: Cross tabulation of PsyCap and Edu variables

	Edu = Lower=0	Edu = Higher=1	**Total**
PsyCap = Negative=0	22	46	68 (21%)
PsyCap = Positive=1	40	221	261 (79%)
Total	62 [19%]	267 [81%]	329 (100%)

The Equation (10) was applied to Table 17, and the OR was obtained as,

$$\text{OR} = \frac{22 \times 221}{46 \times 40} = \frac{4,862}{1,840} = 2.642 \approx 2.6 \tag{16}$$

Note that $log_e(\text{OR}) = log_e(2.6) = 0.97$. To validate the OR result, Equation (11) was applied to obtain $SE(log\text{OR})$ as shown in the following equation (17):

$$SE(\log \text{OR}) = Sqrt\left(\frac{1}{22} + \frac{1}{46} + \frac{1}{40} + \frac{1}{221}\right) = Sqrt(0.097) = 0.311 \tag{17}$$

Then, applying Equation (12), the lower and upper 95% CI limits were obtained.

$$\left.\begin{array}{l} \text{Lower Limit 95\%CI} = exp(0.97 - 1.96(0.311)) = exp(0.36) = 1.43 \\ \text{Upper Limit 95\%CI} = exp(0.97 + 1.96(0.311)) = exp(1.58) = 4.85 \end{array}\right\} \tag{18}$$

Thus, 95% CI of the OR result, 2.6, is 1.43 to 4.85, yielding a p-value = 0.002 < 0.05.

4.2.2.3 PsyCap and Organizational Tenure

The 329 observations of the PsyCap dataset were cross tabulated based on the binary variables, PsyCap and OrgTenure, forming Table 18.

Table 18: Cross tabulation of PsyCap and Org tenure variables

	OrgTenure = Short = 0	OrgTenure = Long = 1	**Total**
PsyCap = Negative = 0	62	6	68 (21%)
Psy Cap = Positive = 1	193	68	261 (79%)
Total	255 [78%]	74 [22%]	329 (100%)

By applying the OR formula in equation (10) to Table 18, the following result was obtained.

$$OR = \frac{62 \times 68}{193 \times 6} = \frac{4,216}{1,158} = 3.641 \approx 3.6 \tag{19}$$

Note that $log_e\left(\text{OR}\right) = log_e\left(3.6\right) = 1.29$. To validate the OR result, Equation (11) was applied to obtain $SE\left(log\text{OR}\right)$ as shown in below Equation (20):

$$SE(\log\text{OR}) = Sqrt\left(\frac{1}{62}+\frac{1}{6}+\frac{1}{192}+\frac{1}{68}\right) = Sqrt\left(2017\right) = 0.4502 \tag{20}$$

The formula in Equation (12) was then deployed to obtain lower and upper 95% CI limits as shown in the Equation (21).

$$\left.\begin{array}{l}\text{Lower Limit 95\%CI} = exp(1.29 - 1.96(0.4502)) = exp(0.41) = 1.51 \\ \text{Upper Limit 95\%CI} = exp(1.29 + 1.96(0.4502)) = exp(2.17) = 8.78\end{array}\right\} \tag{21}$$

Thus, 95% CI of the OR result 3.6, is 1.51 to 8.78, producing a *p*-value = 0.004 < 0.05.

4.3 Results and Discussion

The OR value obtained in the Equation (16) means that employees with higher educational qualifications are 2.6 times more likely to have positive psychological capital than employees with lower educational qualifications (Atsa’am & Bodur, 2019b). In other words, employees perceived to have positive psychological capital are 2.6 times more likely to be holders of at least a Bachelor’s degree compared with those having negative psychological capital. Going by the criteria for statistical significance given in Szumilas (2010) and Hancock & Kent (2016), the 95% CI of 1.43 to 4.85, with a p-value = 0.002 < 0.05, show that this result is statistically significant. This implies that whichever sample population is deployed to calculate OR of psychological capital and educational qualification, the result must fall between 1.43 and 4.85, 95 out of every 100 times.

Meanwhile, the interpretation of the odds ratio in the Equation (19) is that, an employee who has stayed longer periods in an organization is 3.6 times more likely to have positive psychological capital than one who has stayed shorter periods (Atsa’am & Bodur, 2019b). In another way, this could be put as; employees perceived to have positive psychological capital are 3.6 times more likely to have stayed longer in an organization compared with those having negative psychological capital. We are 95% confident that whenever the relationship between PsyCap and organizational tenure is evaluated on whichever population sample, the result will fall between 1.51 and 8.78. Considering the criteria in Szumilas (2010) and Hancock & Kent (2016), this result is statistically significant because the null value, 1, is not included within this interval; furthermore, the interval produced p-value = 0.004 < 0.05.

Chapter 5

CONCLUSION

This dissertation addressed two aspects within the Data Mining field: filter variable selection and knowledge discovery in datasets. In the first sub-area, a filter variable ranking algorithm that relies on risk ratios to evaluate the association between a predictor and the class was developed and tested. Under the second sub-area, the odds ratio measure was deployed to investigate the relationship between employees' psychological capital and educational qualifications, and the relationship between employees' psychological capital and organizational tenure.

In the era of big data, where voluminous, high-dimensional data are constantly being generated from healthcare delivery activities, it is necessary to pay more attention to the problem of variable selection. The majority of the attributes that come with historical or daily data are usually not necessary in modeling. When such unimportant attributes are not eliminated before model construction, many metrics of model diagnostics, such as variance, deviance, degrees of freedom, and predictive accuracy, are negatively affected. Furthermore, machine learning algorithms train slower, and constructed models are over-fitted and more complex to interpret if irrelevant predictors are included. The ranking algorithm developed in this research, which performs competitively with some existing algorithms, will be a useful tool for dimensionality reduction in healthcare data to guard against these unwanted results in classification. As could be observed in Chapter 3, this algorithm demonstrates that it is more appropriate for healthcare datasets than other domains. Better performance was recorded in the cancer, PsyCap, diabetes, and melanoma datasets compared with the spam e-mail dataset. The algorithm achieves a variable importance ranking by employing the statistical measure of risk ratio to evaluate the association between a predictor and the response. Predictors exhibiting a strong association with the class will be selected for classification, while those with a weak association will be excluded. The algorithm does not include a means of determining a threshold of which variables to include in a model. It is left to the discretion of the modeler to apply trial and error in adding or removing variables based on the ranking and performance of previous models. In future research, the algorithm should be extended to be able to determine a cut-off point of important variables algorithmically. Also, the possibility of implementing this algorithm in a way that makes it compatible with open-source languages, such as R, should be explored. As a candidate filter method, the algorithm is independent of any machine learning tool. It is meant to effect variable selection as a preprocessing activity, after which any modeling tool can be applied for model fitting proper. The algorithm is generic; thus, it can execute on any healthcare dataset, provided it is numeric with a dichotomous response.

Meanwhile, a gap was identified in existing literature to the effect that no work had been done previously to establish the relationship between employees' psychological capital and their level of educational qualifications. Before now, it was not known whether the extent to which an employee is educated has positive or negative effect on their PsyCap level. It was also not clear whether the length of service of an employee within an organization has positive or negative impact on their PsyCap. In order to investigate these, the psychological capital data of a sample of 329 workers in the hospitality industry was collected, including their educational qualifications and length of service information in years. The dataset variables were dichotomized, generating three binary variables: PsyCap, educational qualification, and organizational tenure. The odds ratio was deployed to evaluate the association between PsyCap and educational qualification. It turned out that employees with higher educational qualifications are 2.6 times more likely to have positive psychological capital than employees with lower educational qualifications. This result was statistically significant at the 95% CI of 1.43 to 4.85, with a *p*-value = 0.002 < 0.05. Equally, the association between PsyCap and organizational tenure was evaluated using odds ratio. The result showed that employees who have stayed longer with an organization are 3.6 times more likely to be seen as having positive PsyCap than those who have stayed shorter periods. This was also statistically significant at the 95% CI of 1.51 to 8.78, with a *p*-value = 0.004 < 0.05. The implication of these findings is that business owners should be conscious of the educational background of candidates to hire. Priority should be given to those with higher qualifications since those possessing them have better tendencies of exhibiting positive PsyCap; and positive PsyCap influences job performance of an employee. Also, the findings imply that business owners should make concert efforts to retain workers that have stayed longer periods in service. During general retrenchment, those employees with shorter organizational tenure should be considered so that the psychological capital asset of the company could be preserved. It is not clear whether other factors such as poor employee welfare have potentials to mediate negatively between the association between PsyCap and educational qualifications, and or organizational tenure. This is left open for future research.

REFERENCES

Alarcon, G.M., Bowling, N.A., Khazon, S. (2013). Great expectations: A meta-analytic examination of optimism and hope. *Personality and Individual Differences, 54,* 821-827, DOI:10.1016/j.paid.2012.12.004.

Al-Radaideh, Q.A., & Nagi, E.A. (2012). Using data mining techniques to build a classification model for predicting employees performance. *International Journal of Advanced Computer Science and Applications, 3(2),* 144-151.

Andrade, C. (2015). Understanding relative risk, odds ratio, and related terms: As simple as it can get. *J. Clin. Psychiatry, 76*, 857-861, DOI:10.4088/JCP.15f10150.

Antunes, A.C., Caetano, A., & Cunha, M.P. (2017). Reliability and construct validity of the Portuguese version of the psychological capital questionnaire. *Psychological Reports, 120,* 520-536.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The 'K' in k-fold cross validation. In *ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (25-27 April, 2012, pp. 441-445). Bruges, Belgium: i6doc.com publ.

Atsa'am, D.D., & Bodur, E.K. (2019a). A data mining classifier for predicting employees' psychological capital. In *Proceedings of the 4th International*

Conference on Computational Mathematics and Engineering Sciences (CMES-2019, In Press), Antalya, Turkey.

Atsa'am, D.D., & Bodur, E.K. (2019b). Knowledge mining on the association between psychological capital and educational qualifications among hospitality employees, *Current Issues in Tourism,* DOI: 10.1080/13683500.2019.1597026.

Avey, J.B., Luthans, F., Smith, R.M., & Palmer, N.F. (2010). Impact of positive psychological capital on employee well-being over time. *Journal of Occupational Health Psychology, 15,* 17-28, DOI:10.1037/a0016998.

Bagherzade-Khiabani, F., Ramezhankani, A., Aiziz, F., Hadaegh, F., Steyerberg, E.W., & Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology, 71,* 76-85.

Bermejo, P., Ossa, L., Gamez, J.A.. & Puerta, J.M. (2012). Fast wrapper feauturesubset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems, 25,* 35-44.

Bland, J.M., & Altman, D.G. (2000). Statistics notes: The odds ratio. *BMJ, 32,* 1468.

Bodur, E.K., & Atsa'am, D.D. (2019). Filter variable selection algorithm using risk ratios for dimensionality reduction of healthcare data for classification. *Processes,7,* 222, DOI:10.3390/pr7040222.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32, DOI:10.1023/A:1010933404324.

Canty, A., & Ripley, B. (2017). *boot: Bootstrap R (S-Plus) functions. R package version 1.3-20.* Retrieved March 7, 2019, from https://cran.r-project.org/web/packages/boot/boot.pdf

Capistrant, B.D.; Moon, J.R.; Glymour, M.M. (2012). Spousal Caregiving and Incident Hypertension. *Am. J. Hypertens. 25*, 437–443, DOI:10.1038/ajh.2011.232.

Carnevale, A.P., Smith, N., & Strohl, J. (2010). *Help wanted: Projections of jobs and education requirements through 2018.* The Georgetown University Center on Education and the Workforce. Retrieved March 4, 2019, from https://cew.georgetown.edu/cew-reports/help-wanted.

Carver, C.S., & Scheier, M.F. (2002). Control processes and self-organization as complementary principles underlying behavior. *Personality and Social Psychology Review, 6,* 304-315.

Catena, S., Colla, V., & Vannucci, M. (2014). A hybrid feature selection method for classification purposes. In *UKSim-AMSS 8th European Modeling Symposium on Mathematical Modeling and Computer Simulation* (EMS2014, 39-44), Pisa, Italy: IEEE Computer Society.

Catena, S., Colla, V., & Vannucci, M. (2017). A fuzzy system for combining filter features selection methods. *International Journal of Fuzzy Systems, 19*(4), 1168-1180.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering, 40,* 16-28.

Chapela, J.G. (2013). Things that make us different: Analysis of deviance with time-use data. *J. Appl. Stat., 40,* 1572-1585, DOI:10.1080/02664763.2013.789097.

Chen, H., Cohen, P., & Chen, S. (2007). Biased odds ratios from dichotomization of age. *Statistics in Medicine, 26,* 3487-3497.

Chen, M., Mao, S., Liu, Y. (2014). Big data: A Survey. *Mob. Netw. Appl., 19,* 171-209, DOI:10.1007/s11036-013-0489-0.

Crone, S.F., & Kourentzes, N. (2010). Feature selection for time series prediction – A combined filter and wrapper approach for neural networks. *Neurocomputing, 73,* 1923-1936.

Dadaneh, B.Z., Markid, H.Y., & Zakerolhosseini, M.A. (2016). Unsupervised probabilistic feature selection using ant colony optimization. *Expert Systems With Applications, 53,* 27-42, DOI:10.1016/j.eswa.2016.01.021.

Durrah, O. Alhamoud, A., & Khan, K. (2016). Positive psychological capital and job performance: The mediating role of job satisfaction. *International Scientific Research Journal, 72*, 241-225, DOI:10.21506/j.ponte.2016.7.17.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research, 5,* 1531-1555.

Frenay, B., Doquire, G., & Verleysen, M. (2013). Is mutual information adequate for feature selection in regression? *Neural Networks, 48,* 1-7.

Genuer, R.; Poggy, J.; Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31,* 2225–2236, DOI:10.1016/j.patrec.2010.03.014.

Golub, T.R., Slonim, D.K., Tamayo,P., Huard, C., Gaasenbeek,M., Mesirov, J.P., Coller, H., Loh,M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., & Lander, E.S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286,* 531-537.

Grimes, D.A., & Schulz, K.F. (2018). Making sense of odds and odds ratios. *Obstetrics & Gynecology, 111,* 423-426.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques.* Morgan Kaufmann Publishers, Massachusetts.

Hancock, P., & Kent, P. (2016). Interpretation of dichotomous outcomes: Risk, odds, risk ratios, odds ratios and number needed to treat. *Journal of Physiotherapy, 62,* 172-174, DOI:10.1016/j.jphys.2016.02.016.

Holzinger, A., & Jurisica, I. (2014). Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In A. Holzinger & I. Jurisica (Eds.), *Interactive knowledge discovery and data mining in biomedical informatics: State-of-the-art and future challenges,* 1-18. Heidelberg: Springer.

Hu, Z., Bao, Y., Xiong, T., & Chiong, R. (2015). Hybrid filter–wrapper feature selection for short-term load forecasting. *Engineering Applications of Artificial Intelligence, 40,* 17-27.

Huang, S.H., Wulsin, L.R., Li, H., & Guo, J. (2009). Dimensionality reduction for knowledge discovery in medical claims database: Application to antidepressant medication utilization study. *Computer Methods and Programs in Biomedicine, 93,* 115-123.

Hur, J., Ihm, S., Park, Y. (2017). A variable impacts measurement in random forest for mobile cloud computing. *Wirel. Commun. Mob. Comput. 2017,* 1–13, DOI:10.1155/2017/6817627.

Jain, S., Shukla, S., Wadhvani, R. (2018). Dynamic selection of normalization techniques using data complexity measures. *Expert Syst. Appl., 106*, 252-262, DOI:2018.04.008.

Javed, K., Babri, H.A., & Saeed, M. (2014). Impact of a metric of association between two variables on performance of filters for binary data. *Neurocomputing, 143,* 248-260.

Jin, K., Cheng, H., & Wang, W.(2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement, 42*(8), 613-629.

Jung, Y., & Hu, J. (2015). A k-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics, 27*(2), 1-13.

Kantardzic, M. (2011). *Data Mining Concepts, Models, Methods, and Algorithms* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Kleinbaum, D.G., & Klein, M. (2010). *Logistic Regression: A Self Learning Text* (3rd Ed). New York: Springer, 73-101.

Kobau, R., Seligman, M.E., Peterson, C., Diener, E., Zack, M.M., Chapman, D., & Thompson, W. (2011). Mental health promotion in public health: perspectives and strategies from positive psychology. *American Journal of Public Health, 108,* e1-e9, DOI:10.2105/AJPH.2010.300083.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Benesty, M., Lescarbeau, R., et al. (2017). *Caret: Classification and regression training, R package version 6.0-77.* Retrieved December 12, 2018, from https://CRAN.R-project.org/package=caret

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Last, A., & Wilson, S. (2004). Relative Risks and Odds Ratios: What's the Difference? *J. Fam. Pract. 53,* 108–108.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A. Molter, C., Schaetzen, V., Duque, R., Bersini, H., & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(4), 1106-1119.

Lazhar, F., & Yamina, T. (2016). Mining explicit and implicit opinions from reviews. *International Journal of Data Mining, Modelling and Management, 8(1),* 75-92.

Lee, C., & Chu, K. (2016). Understanding the effect of positive psychological capital on hospitality interns' creativity for role performance. International Journal of *Organizational Innovation, 8,* 213-222.

Leon-Perez, J.M., Antino, M., & Leon-Rubio, J.M. (2016). The role of psychological capital and intragroup conflict on employees' burnout and quality of service: A multilevel approach. *Frontiers in Psychology, 7,* 1-11, DOI:10.3389/fpsyg.2016.01755.

Lever, J., Krzywinski, M., & Altman, N.S. (2016). Points of significance: Classification evaluation. *Nat. Methods*, *13*, 603–604, DOI:10.1038/nmeth.3945.

Li, J., Liu, H., Tung, A., & Wong, L. (2014). The practical bioinformatician. In L. Wong (Ed.), *Data mining techniques for the practical bioinformatician* (pp. 35-70). 5 Toh Tuck Link, Singapore: World Scientific Publishing.

Liaw, A., & Wiener, M. (2002). Classification and regression by Randomforest. *R News*, *2*, 18–22.

Liu, D., Li, T., & Liang, D. (2014). Incorporating logistic regression to decision-theoretic rough sets for classifications. *International Journal of Approximate Reasoning, 55*, 197-210, DOI:10.1016/j.ijar.2013.02.013.

Lumley, T. (2017). *Leaps: Regression subset selection. R package version 3.0.* Retrieved December 12, 2018, from https://CRAN.R-project.org/package=leaps

Luo, H., Wang, Y., & Yi, L. (2017). The intermediate effect of psychological capital between culture and performance. In *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting*, DOI:10.1177/1541931213601704.

Luthans, F., & Youssef-Morgan, C.M. (2017). Psychological capital: An evidence-based positive approach. *Annual Review of Organizational Psychology and*

Organizational Behavior, 4, 339-366, DOI:10.1146/annurev-orgpsych-032516-113324.

Luthans, F., Youssef, C. M., & Avolio, B. J. (2007). Psychological capital: Investing and developing positive organizational behavior. In D.L. Nelson & C.L. Cooper (Eds), *Positive Organizational Behavior*. London: SAGE Publications Ltd. 9-24.

Luthans, F., Youssef-Morgan, C. M., & Avolio, B. J. (2015). *Psychological Capital and Beyond*. England: Oxford University Press.

Maindonald, J.H., Braun, J.W. (2019). *DAAG: Data analysis and graphics data and functions. R package version 1.22.1*. Retrieved March 7, 2019, from https://CRAN.R-project.org/package=DAAG

Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. *Information Sciences, 179,* 2208-2217.

McNutt, L., Wu, C., Xue, X., & Hafner, J.P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am. J. Epidemiology*, *157*, 940-943, DOI:10.1093/aje/kwg074.

Meyer, P.E., Schretter, C., & Bontempi, G. (2008). Information-theoretic feature selection in microarray aata using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing, 2*(3), 261-274.

Murtaugh, P.A. (2009). Performance of several variable-selection methods applied to real ecological data. *Ecology Letters 12*, 1061-1068, DOI:10.1111/j.1461-0248.2009.01361.x.

Paek, S., Schuckert, M., Kim, T.T., & Lee, G. (2015). Why is hospitality employees' psychological capital important? The effects of psychological capital on work engagement and employee morale. *International Journal of Hospitality Management, 50,* 9-26, DOI: 10.1016/j.ijhm.2015.07.001.

Pandey, A., & Jain, A. (2017). Comparative analysis of Knn algorithm using various normalization techniques. *Int. J. Comp. Netw. Inf. Secur., 11*, 36-42, DOI:10.5815/ijcnis.2017.11.04.

Pandis, N. (2012). Risk ratio vs odds ratio: Statistics and research design. *Am. J. Orthod. Dentofac. Orthop., 142,* 890-891, DOI:10.1016/j.ajodo.2012.08.003.

Park, H. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *J Korean Acad Nurs. 43*, 154-164.

R Core Team (2018). *R: A language and environment for statistical computing*; *R foundation for statistical computing:* Vienna, Austria. Retrieved December 11, 2018, from https://www.R-project.org/

Rice, J.A. (2007). *Mathematical Statistics and Data Analysis* (3rd Ed.). Belmont, CA: Thomson Books/Cole.

Rohde, J.M.; Dimcheff, D.E.; Blumberg, N.; Saint, S.; Langa, K.M. (2014). Health care-associated infection after red blood cell transfusion: A systematic review and meta-analysis. *J. Am. Med. Assoc.* 2014, *311*, 1317-1326, DOI:10.1001/jama.2014.2726.

Schmidt, C.O., & Kohlmann, T. (2008). When to Use the Odds Ratio or the Relative Risk? *Int. J. Public Health, 53*, 165–167.

Simons, J.C., & Buitendach, J.H. (2013). Psychological capital, work engagement and organisational commitment amongst call centre employees in South Africa. *SA Journal of Industrial Psychology, 39,* 1-12, DOI:10.4102/ sajip.v39i2.1071.

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S. (1998). Using the ADAP learning algorithm to forecast the onset of dia*betes mellitus.* In *Proceedings of the Annual Symposium on Computer Application in Medical Care (*Orlando, Florida, FL, *7-11 November, 1998),* Washington, DC: IEEE Computer Society Press.

Snyder, C.R., Harris, C., Anderson, J.R., Holleran, S.A., Irving, L.M., Sigmon, S.T., Yoshinobu, L., Gibb, J., Langelle, C., & Harney, P. (1991). The will and the ways: Development and validation of an individual-differences measure of hope. *Journal of Personality and Social Psychology, 60,* 570-585.

Sperandei, S. (2014). Lessons in biostatistics: Understanding logistic regression analysis. *Biochemia Medica, 24,* 12-18.

Sroka, C.J., & Nagaraja, H.N. (2018). Odds ratios from logistic, geometric, Poisson, and negative binomial regression models. *BMC Medical Research Methodology, 18*(112), 1-11.

Stajkovic, A., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin, 124,* 240-261, DOI: 10.1037//0033-2909.124.2.240.

Stoltzfus, J.C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine, 18,* 1099-1104, DOI:10.1111/j.1553-2712.2011.01185.x.

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform, 8*(1), 25, DOI:10.1186/1471-2105-8-25.

Sumathi, K., Kannan, S., & Nagarajan, K. (2016). Data mining: Analysis of student database using classification techniques. *International Journal of Computer Applications, 141(8),* 22-27.

Sun, T., Zhao, X.W., Yang, L.B., & Fan, L.H. (2011). The impact of psychological capital on job embeddedness and job performance among nurses: A structural equation approach. *Journal of Advanced Nursing, 68,* 69-79, DOI:10.1111/j.1365-2648.2011.05715.x.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 19,* 227-229.

Tamhane, A.R., Westfall, A.O., Burkholder, G.A., & Cutter, G.R. (2016). Prevalence odds ratio versus prevalence ratio: Choice comes with consequences. *Stat. Med. 35*, 5730–5735, DOI:10.1002/sim.7059.

Tharwat, A. (2018). Classification assessment methods. *Appl. Comput. Inf.* DOI:10.1016/j.aci.2018.08.003.

Tharwat, A., Moemen, Y.S., & Hassanien, A.E. (2017). Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines. *Journal of Biomedical Informatics, 68,* 132-149, DOI:10.1016/j.jbi.2017.03.002.

Tran, T.N., Afanador, N.L., Buydens, L.M.C., & Blanchet, L. (2014). Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemometrics and Intelligent Laboratory Systems, 138,* 153-160.

Tseng, C. (2011). Diabetes and risk of prostate cancer: A study using the national health insurance. *Diabetes Care, 34,* 616-621. DOI:10.2337/dcl0-1640.

Tugade, M.M., Fredrickson, B.L., & Barrett, L.F. (2004). Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of Personality, 72,* 1161-1190.

Vanderweele,T.J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology, 172*(12), 1339-1348.

Vaske, J.J., Beaman, J., & Sponarski, C.C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences,* 39(2), 163-173, DOI:10.1080/01490400.2015.1127189.

Venables, W.N., Ripley, B.D. (2002). *Modern Applied Statistics with S* (4th Ed). New York: Springer. pp. 331-349.

Wang, H., Yang, F., & Luo, Z. (2016). An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinform, 17,* 60, DOI:10.1186/s12859-016-0900-5.

Wang, Z., & Bovik, A.C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag., 26,* 98-117, DOI:10.1109/MSP.2008.930649.

Wu, L. (2017). Production adoption rate prediction in a competitive market. *IEEE Transactions on Knowledge and Data Engineering, 30*(2), 325-338.

Wu, X. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97-107.

Xue, Yao, & Wu (2018). A novel ensemble-based wrapper method for feature selection using extreme learning machine and genetic algorithm. *Knowledge and Information Systems, 57*(2), 389-412.

APPENDICES

Appendix A: Variables of Experimental Datasets

Table A.1: Psychological capital dataset variables

Variable Name	Description	Data Type	Possible Values
S1	I am confident when analyzing a long-term problem to find a solution	Integer	1,2,3,4,5,6
S2	I am confident when presenting my work area in meetings with authorities	Integer	1,2,3,4,5,6
S3	I am confident when participating in discussions relating to my employer’s strategy	Integer	1,2,3,4,5,6
S4	I am confident when helping to set targets/goals in my work area	Integer	1,2,3,4,5,6
S5	I am confident when meeting people outside my work environment to discuss problems	Integer	1,2,3,4,5,6
O1	I would resign to fate if anything goes wrong at my work place	Integer	1,2,3,4,5,6
O2	I think positively on all issues relating to my official duties	Integer	1,2,3,4,5,6
O3	I have a positive feeling about future happenings regarding my work	Integer	1,2,3,4,5,6
O4	Things always go the wrong way against my expectations at work	Integer	1,2,3,4,5,6
O5	I carry out my official duties with a mindset that success is sure	Integer	1,2,3,4,5,6
H1	In the event of any difficult situation at work, I can devise several alternative means to overcome it	Integer	1,2,3,4,5,6
H2	I am currently pursuing my goals zealously	Integer	1,2,3,4,5,6
H3	There exist several potential solutions to any problem I am faced with now	Integer	1,2,3,4,5,6
H4	I can think of many ways to reach my current goals	Integer	1,2,3,4,5,6
H5	At this time, I am meeting the work goals I have set for myself	Integer	1,2,3,4,5,6
R1	When I have a setback at work, I have trouble recovering from it and moving on	Integer	1,2,3,4,5,6
R2	I can perform my duties independently if necessary	Integer	1,2,3,4,5,6
R3	I work diligently in my stride when tackling difficult tasks at my job	Integer	1,2,3,4,5,6
R4	I am capable of handling difficult situations at work because of past experience	Integer	1,2,3,4,5,6
R5	I can multitask at the same time while performing my duties	Integer	1,2,3,4,5,6
PsyCap	Class	Boolean	0,1

Table A.2: Diabetes in Pima Indian women dataset variables

Variable Name	Description	Data Type
npreg	Number of pregnancies	Integer
glu	Plasma glucose concentration in an oral glucose tolerance test	Integer
bp	Diastolic blood pressure (mm Hg)	Integer
skin	Triceps skin fold thickness (mm)	Integer
bmi	Body mass index (weight in kg/(height in mm)\^2)	Real
ped	Diabetes pedigree function	Real
age	Age in years	Integer
type	Yes or No, for diabetic according to WHO criteria	Boolean

Table A.3: Survival from malignant Melanoma dataset variables

Variable Name	Description	Data Type
Time	Number of days survived since the operation	Integer
Status	Status indicating whether the patient died after the operation. 1 = died from melanoma, 2 = survived, 3 = died from another cause	Integer
Sex	Patient's gender	Boolean
Age	Patient's age at operation	Integer
Year	Year of operation	Integer
Thickness	Thickness of tumour in mm	Real
Ulcer	Whether ulceration present or not	Boolean

Table A.4: Spam e-mail dataset variables

Variable Name	Description	Data Type
crl.tot	Total length of words appearing in capitals	Integer
dollar	Number of times the symbols \ and $ occur	Integer
bang	Number of times the symbol ! Occurs	Integer
money	Number of times the word 'money' occurs	Integer
n000	Number of times the string '000' occurs	Integer
make	Number of times the word 'make' occurs	Integer
yesno	Class variable with 1 = spam, 0 = not spam	Boolean

Table A.5: Biopsy of breast cancer patients dataset variables

Variable Name	**Description**	**Data Type**
V1	Thickness of clump	Integer
V2	Cell size uniformity	Integer
V3	Cell shape uniformity	Integer
V4	Marginal adhesion	Integer
V5	Single epithelial cell size	Integer
V6	Bare nuclei	Integer
V7	Bland chromatin	Integer
V8	Normal nucleoli	Integer
V9	Mitoses	Integer
Class	Outcome, whether 0 = benign or 1 = malignant	Boolean

Appendix B: Proposed Algorithm

Algorithm 2. Proposed Algorithm

```
• Step 1. Binarizing the Dataset
//Listing 1. This step converts all input values to binary
    For $j$ = 1 to $n$ //counts columns
    For $i$ = 1 to $m$ //counts rows
                    IF RawData $[i, j]$ < 0.5 Then
                    RawData $[i, j]$ = 0 //round down values to 0
                    ELSE RawData $[i, j]$ = 1 //round up values to 1
                END IF
        Next $i$
    Next $j$
• Step 2. Counts Occurrences of $t_{11}, t_{10}, t_{01}, t_{00}$
    //Listing 2. This step counts $t_{11}$, $t_{10}$, $t_{01}$, $t_{00}$ for each predictor
    RawData = Array $[1...m][1...n]$ As Integer //2-dim array of rows/columns
Class = Array $[1...m]$ As Integer //1-dim array for class
    $\delta j = 0 : \beta j = 0 : \theta j = 0 : \varphi j = 0$ As Integer //initialize sums of $t_{11}, t_{10}, t_{01}, t_{00}$
    For $j$ = 1 to $n$ // holds column index position for predictors
        For $i$ = 1 to $m$ //holds row index position for predictors
            For $y$ = 1 to $m$ //holds row index position for class
                IF $i$ = $y$ THEN //compares input and output index
                IF RawData $[i, j]$ = 1 AND Class $[i, j]$ = 1 THEN
                        $\delta j = \delta j + 1$ //counts $t_{11}$
ENDIF
    IF RawData $[i, j]$ = 1 AND Class $[i, j]$ = 0 THEN
                        $\beta j = \beta j + 1$ //counts $t_{10}$
ENDIF
    IF RawData $[i, j]$ = 0 AND Class $[i, j]$ = 1 THEN
                        $\phi j = \phi j + 1$ //counts $t_{01}$
ENDIF
    IF RawData $[i, j]$ = 0 AND Class$[i, j]$ = 0 THEN
                        $\varphi j = \varphi j + 1$ //counts $t_{00}$
ENDIF
ENDIF
            Next $y$
        Next $i$
    Next $j$
• Step 3. Computes RR for each Column
```

```
//Listing 3. This step computes Risk Ratios for each column
//temporary variables
$\text{lowerSum}_j$, $\text{upperSum}_j$ As Intger
$\text{firstRatio}_j$, $\text{secondRatio}_j$, $\text{RR}_j$ As Real
For $j$ = 1 to $n$
    $\text{lowerSum}_j = \delta j + \beta j$
    $\text{upperSum}_j = \phi j + \varphi j$
    $\text{firstRatio}_j = \frac{\delta j}{\text{lowerSum}j}$
    $\text{secondRatio}_j = \frac{\text{upperSum}j}{\phi j}$
    $\text{RR}_j$ = $\text{firstRatio}_j$´ $\text{secondRatio}_j$ //computes RR
    Print $\text{columnName}_j$ + \tab $\text{RR}_j$ +\ enter
Next $j$
```