Applied Information Technology And Computer Science Vol. 2 No. 1 (2021) 318-325
© UniversitiTun Hussein Onn Malaysia Publisher's Office

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

A Comparative Analysis of Data Mining Techniques for Mental Health Classification Problem

Muhammad Ali Shahrudin[1], Nazri Mohd Nawi[2]*, Radiah Mohamad[1]

[1]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, MALAYSIA

[2]Soft Computing and Data Mining Centre (SMC),
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, MALAYSIA

DOI: https://doi.org/10.30880/aitcs.2021.01.01.024
Received 17 March 2021; Accepted 18 March 2021; Available online 31 May 2021

Abstract: Mental health is a very dangerous illness that can bring death to the patient as well as people around them if it is not treated properly at early stage of the illness. Mental health is dangerous because it involves the person psychological and it can make the patients either aggressive or passive. Early detection is crucial in helping experts and practicians to classify them from as normal person. This paper will discuss the use of data mining techniques for classifying the tech worker with mental health problems. We select different methods of data mining techniques to find the most accurate results related to mental health. The technique that will be using is decision tree, Naïve Bayes, and neural network. Based on the three techniques, we will find which of the technique has better accuracy, better precision, and less time taken.

Keywords: Mental health classification, data mining, precision, accuracy

1. Introduction

People that have a mental health problem tend to have moderate to severe depression, bipolar disorder, and schizophrenia, and other psychotic disorders. They are estimated to have a life expectancy of 10-20 years shorter than usual. This is caused by physical health conditions. A lot of people with mental disorder did not get essential health services which offer health promotion, screening, and treatment for physical and mental health conditions. For those who can get the health care services, the care is in poor quality and physical health conditions are often ignored[1]. This research focuses on comparing some of the data mining techniques in order to get an accurate result for the factors that cause the mental health problem among tech workers and categorize them. The results of classification will identify workers that are badly affected to a mental health issue in the workplace. Furthermore, it is hard to know which factor that caused tech worker in having mental health. The objectives of this

*Corresponding author: nazri@uthm.edu.my research is to study, analysis and simulate the comparison results of data mining techniques for tech workers with mental health problem classification based on accuracy, precision, and time taken. Furthermore, this research is to investigate the mental health of tech workers by using data mining techniques in finding which techniques that can give the most accurate result about the mental health issues.

The rest of the paper is organized as follows: Section II describes the related work on profiling and classification models. Section III presents the proposed classification model for mental health. Finally, Section IV concludes the work and highlights a direction for future research.

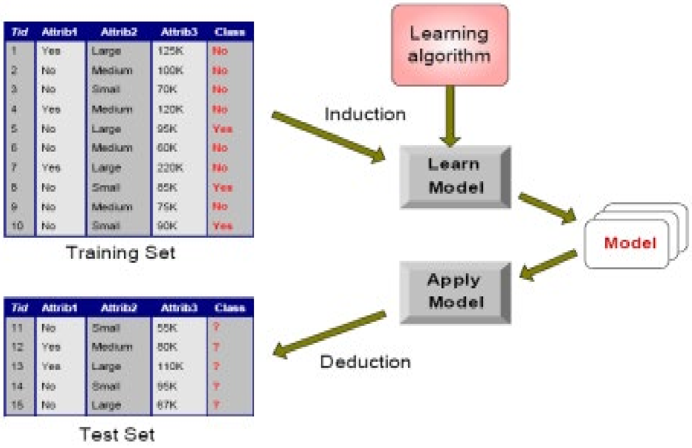
2. Related Work

2.1 Mental Health

Mental, physical, and social wellbeing are crucial components of life that are firmly joined and greatly reliant. It turns out to be always obvious that emotional well-being is important to the general prosperity of people, social orders, and nations. Unfortunately, in many pieces of the world, psychological wellness and mental issue are not respected with anything like a similar significance as physical wellbeing. Rather, they have been to a great extent disregarded or ignored. Halfway, therefore the world is experiencing an expanding weight of mental issue, and an extending "treatment hole". Today, somewhere in the range of 450 million individuals experience the ill effects of a psychological or conduct issue, yet just a little minority of them get even the most fundamental treatment. In creating nations, most people with a severe mental issue are left to adapt as well as can be expected with their private weights, for example, sorrow, dementia, schizophrenia, and substance reliance. All around, many are deceived for their ailment and become the objectives of disgrace and separation. Further increments in the number of sufferers are likely in perspective on the maturing of the populace, declining social issues, and common turmoil. Effectively, mental issue speaks to four of the 10 driving reasons for inability around the world. This developing weight adds up to an enormous expense as far as human hopelessness, handicap and financial misfortune Mental and social issue are evaluated to represent 12% of the worldwide weight of malady, yet the emotional wellness spending plans of most of the nations establish under 1% of their all-out wellbeing consumptions [1].

2.2 Classification Techniques

Classification is one of the techniques in data mining that provides the capability of processing a variety of data. Data mining is the most significant in Machine Learning (ML) where it can make analyses, establish a relationship between multiple features easier, more successful, and more efficiencies in systems and the design of machines [2]. It can handle a big amount of data, predict a categorical class label, and can classify database on training set and class labels and be used for classifies a new set of data [3]. The classification methods are strong for modeling interactions. Every each of the methods can be used in any situation where can use while the other may not and the other way around. Every one of the techniques has its own pros and cons [4]. There are different types of data set that any of classification technique can be applied like data of patients, financial data according to performance, customer segmentation, business modeling, marketing, credit analysis and biomedical and drug response modeling. Moreover, classification techniques can show how the data work and how can it decide and grouped when a new set of data is available. Each one of the techniques has a condition that needed in order for them to be selected. The algorithm can also be used to detect natural disasters like cloud bursting, earthquakes, etc. [2][4]. Data classification is defined as two-step process shown in Figure 1.



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Figure 1: illustrating classification task[2]

(a) Decision Tree

Decision tree classifies instances is created by sorting them based on the values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represent a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature value [5]. The issue of developing ideal two-fold choice trees is a NP-complete issue and along these lines, theoreticians have looked for efficient heuristics for building close ideal choice trees. The component that best partitions the preparation information would be the root hub of the tree. There are various strategies for finding the component that best partitions the preparation information [6]. Correlation of individual techniques may even now be significant when choosing which metric ought to be utilized in a specific dataset. A similar methodology is then rehashed on each segment of the partitioned information, making sub-trees until the preparation information is separated into subsets of a similar class [6].

(b) Naïve Bayes

Autonomy supposition. Hence, there were additionally many intriguing works of exploring Naive Bayes since it has been one of the well-known machine learning strategies for a long time. Its effortlessness makes the structure appealing in different assignments and sensible exhibitions are gotten in the errands even though this learning depends on a ridiculous Naive Bayes. Particularly, appears that naive Bayes can perform shockingly well in the classification tasks where the probability itself calculated by the naive Bayes isn't significant [7]. It is an independent component model. It calculates probabilities to hypothesis. Furthermore, it is robust to noise in the information. This classifier accepts that the presence/absence of a specific element is lineal to any component's presence/absence. It is a managed learning classifier. It is easy to develop and simple to translate. So clients who are do not have earlier learning in classification can likewise build Naive Bayes classifier [8]. Bayesian approaches are a fundamentally important DM technique. Given the probability distribution, Bayes classifier can provably achieve the optimal result. Bayesian method is based on the probability theory.

(c) Artificial Neural Networks

Artificial Neural networks is non-linear statistical data modeling tools that be used to model a complex relationship between input and output and to find patterns in data. The difference between data warehouse and an ordinary database is the actual manipulation and cross-fertilization of the data can help the user to have more informed decisions [2]. A Radial basis function (RBF) neural network can be considered as a special three-layered network. The input nodes pass the input values to the internal nodes that formulates the hidden layer. The nonlinear responses of the hidden nodes are weighted to calculates the final outputs of the network in the third (output) layer. A typical hidden node in an RBF network is characterized by its center, which is a vector with dimension equal to the number of inputs to the node [9]. The artificial neural network is an understanding and emulating from the human brain to broader the issues if copying human abilities such as speech and can be use in various filed. There is various application for neural network that involve recognizing patterns and making simple decisions about them. For example, in real time situation, a neural network can be use in airplanes as a basic autopilot where inputs units read signal from the various cockpit instruments and output as unit modifying the plane's controls appropriately to keep in the course [10]. Table 1 shows the comparison between three techniques which shows the advantages and disadvantages between the three techniques.

Table 1: Comparison of the selected data mining techniques

TECHNIQUE	ADVANTAGES	DISADVANTAGES
DECISION TREE	1) It produces an accurate result. 2) It takes less memory to large program execution. 3) It takes less model build time. 4) 4. It has a short time of searching time	1) 1.Empty branches 2) 2.Insignificant branches 3) 3.Overfitting
NAÏVE BAYES	1) To improve the classification performance by removing the irrelevant features. 2) Good performance 3) 3. It is a short computational time.	1) The naïve Bayes classifier requires a very large number of records to obtain good results. 2) It is instance-based or lazy in that they store all of the training samples.
ARTIFICAL NEURAL NETWORK	1) Easy to use, with few parameters adjust. 2) Reprogramming is not needed. 3) Easy to implement 4) Wide range of problem in real life.	1) Requires high processing time if large data. 2) Hard to determines how many neurons and layers is needed. 3) Learning can be slow.

3. Methodology/Framework

This section shows the methodology that been used in this research. It consists of 6 phases and all activities and workflow that are carried out in this research is showed in Figure 2. The research starts with the study of three data mining techniques that will be used in this research. The study selects the suitable techniques for the dataset in order to get a good result. The techniques that been selected are Two-Class Boosted Decision Tree, Two-Class Naïve Bayes Machine, and Two-Class Neural Network. Specifications and properties of materials, equipment, and other resources used in this study also been described in this section. Next, a dataset that will be used for tech worker mental health is taken from Kaggle websites where it is gathered from raw data and need to be clean. Only related attributes for the experiment were selected and all unwanted data were deleted. Not all the empty data can be fixed, therefore Azure Machine Learning is used to fix missing data as shown in Figure 3.

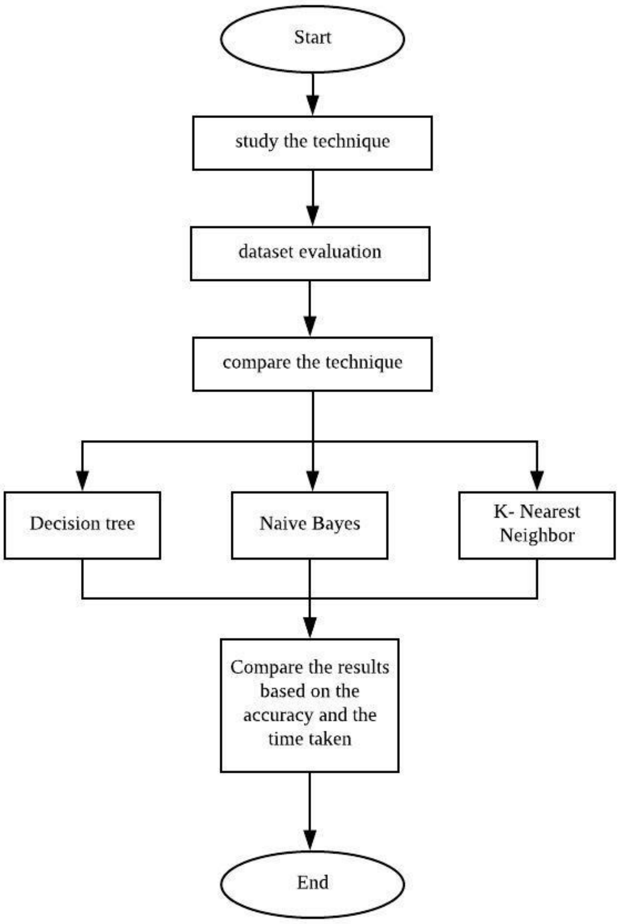
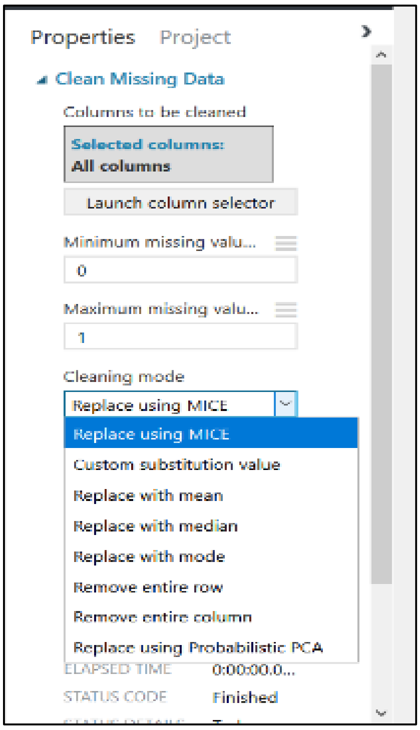


Figure 2: Research Methodology (CRISP-DM)



Clean Missing Data

Columns to be cleaned

Selected columns:
All columns

Minimum missing valu...
0

Maximum missing valu...
1

Cleaning mode
Replace using MICE

ELAPSED TIME 0:00:00.0...
STATUS CODE Finished

Figure 3: Clean Missing Data Model

The datasets were taken from the Kaggle website which contains all attributes that need to pre-process for the purpose of this experiment. Moreover, there are a lot of inconsistence of values in datasets that need to be clean and repair. Some of the pre-process techniques have been applied for that reason which are;

(a) attributes selection - Since the datasets contain a lot of attributes that some of them not suitable for the experiment and need to remove and clean. The process of selecting attributes is done manually by choosing carefully attributes that need to be included in the experiments.
(b) In the normalization process, we organize and manage the dataset to make sure the accuracy of the data in the experiment. It is a process for reducing the redundancies of data in the database. It is also to change the numeric columns in the dataset to use a common scale without altering differences in the range of values in the data.

The performance of selected data mining techniques for this paper were compare and analyses based on some measurement metric. Accuracy is one metric for evaluating classification models. Informally**,** accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \textit{Eq.1}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \qquad \textit{Eq.2}$$

where TP=True Positive, TN=True Negative, FP=False Positives, and FN=False Negatives.

Precision identifies the frequency with which a model was correct when predicting the **positive class**. In short, precision is variation when it is measure at the same part repeatedly with the same value. That is:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad \textit{Eq. 3}$$

4. Results and Discussion

4.1 Results

This section discusses the results that obtained from the Microsoft Azure Machine Learning studio by comparing the performance of the selected classification techniques on the tech worker dataset. The performance of classifiers is based on some selected classifiers such as Decision Tree, Naïve Bayes, Neural Network, and all models are tested on the same datasets. By using the Microsoft Azure Machine Learning Studio, it will ease the experiment. The "Clean missing data" object will help the simulation by filling the missing data with values that suitable based on other data or it will remove the entire row for the experiment. Then the "Split Data" object will split the dataset into 80% which contains a training set that will make the classifiers learn about the data and the expected result. The other 20% of data will be the testing set that needs to be use by classifiers for predicting the trained model to get a result. "Train Model" object will train the 80% of the data with the classification techniques to produce a set of train prediction. "Score Model" object is to generate a prediction based on the train model and 20% of the data. "Evaluate Model" object will measure the accuracy and the precision of the prediction based on the train model and transform it into a graph.

4.2 Discussions

The result of the overall performance for all models on the mental health dataset have been generated by using Microsoft Azure Machine Learning Studio. The performance results will show either the worker with mental health is badly handled response to a mental health issue in the workplace. The accuracy of the experiment can be used as a decision based on the accuracy of the result. The accuracy for each of the experiments is shown in Table 2. As can be seen, the Naïve Bayes experiment has the highest accuracy with 0.689 while the Decision tree experiment managed to achieve the second with the accuracy of 0.588 and follows by Neural Network with an accuracy of 0.568. All three of the experiment are using the same dataset. While for the precision of the experiment, the Naïve Bayes experiment has the highest precision with 0.389 and Neural Network was the second with the precision of 0.280, and Decision Tree was the last with the precision of 0.268. All three of the experiment are using the same dataset. For the time taken for each experiment to run completely, the Naïve Bayes experiment has the shortest time taken with 23 seconds as compared to Neural Network which took 31 seconds and follows by the Decision Tree with52 seconds. Table 2 and Table 3 show the overall result of the experiment. It shows the differences and which one has a better result among the three techniques. Accompanying discussions that further explain observations of the results are usually placed immediately below the results paragraph.

Table 2: The result of the experiments

	ACCURACY	PRECISION	TIME TAKEN
DECISION TREE	0.588	0.268	52 seconds
NAÏVE BAYES	**0.689**	0.389	**23 seconds**
NEURAL NETWORK	0.568	0.280	31 seconds

Table 3: The result form the Evaluate model

	DECISION TREE	NAÏVE BAYES	NEURAL NETWORK
TRUE POSITIVE	11	7	14
FALSE NEGATIVE	31	35	28
FALSE POSITIVE	30	11	36
TRUE NEGATIVE	76	95	70
RECALL	0.262	0.167	0.333
F1 SCORE	0.265	0.233	0.304

This paper shows the performance results of three classifiers on the mental health dataset by using the Microsoft Azure Machine Learning. The main purpose of the experiment is to classify either the worker with mental health is badly handled response to a mental health issue in the workplace. The results show that Naïve Bayes had performed the best as compared to decision tree and Neural Network classification techniques.

5. Conclusion

Mental health is a very dangerous illness that can bring death to the patient as well as people around them if it is not treated properly at early stage of the illness. The early prevention is crucial in identifying the potential workers that effected with mental health. Data mining techniques had gained its popularity in classification problems and it helps many applications such as healthcare, manufacturing etc. in helping them to make a decision based on the classification results. This paper compares and analyses the performance of three data mining techniques on mental health datasets and the results show that Naïve Bayes had performed the best as compared to others. During this research, there are a few problems that been identified such as: (a) lacking of knowledge about the data mining method and data evaluation. This is because more time is needed in order to understand more about the research itself.

(b) finding the most accurate dataset or a suitable dataset is a barrier to have a great result because the dataset also plays a big role in the research. However, there are several suggestions for future improvement that related to this research in getting a better result which are: (a) understanding on how the data mining work will help to get a better result, (b) the use of multiple datasets can help to get more accurate results and (c) understanding the tools that will be needed to do the research can also help to produce a better result.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support and encouragement throughout the process of conducting this study.

References

[1] W. P. Brown, "WHO guidelines: Management of physical health conditions in adults with severe mental disorders," *WHO*, 2018. [Online]. Available: https://www.who.int/mental_health/evidence/guidelines_severe_mental_disorders_web_note_2018/en/.

[2] G. Kesavaraj and S. Sukumaran, "A Study On Classification Techniques in Data Mining," 2013.

[3] S. S. Ahmed and Z. Road, "Survey on Classification Algorithms for Data Mining : (Comparison and Evaluation)," vol. 4, no. 8, pp. 18–25, 2013.

[4] S. Archana and K. Elangovan, "Survey of Classification Techniques in Data Mining," vol. 2, pp. 65–71, 2014.

[5] T. N. Phyu, "Survey of Classification Techniques in Data Mining," vol. I, 2009.

[6] A. I. Rev, E. Software, P. E. Pintelas, and I. D. Zaharakis, "unc orre cted unc orre cted," 2007.

[7] S. Kim, K. Han, H. Rim, and S. H. Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," vol. 18, no. 11, pp. 1457–1466, 2006.

[8] C. Engineering, "A Study of Classification Techniques of Data," pp. 13779–13786, 2017.

[9] H. Sarimveis, P. Doganis, and A. Alexandridis, "A classification technique based on radial basis function neural networks," vol. 37, pp. 218–221, 2006.

[10] S. S. Nikam, "ORIENTAL JOURNAL OF A Comparative Study of Classification Techniques in Data Mining Algorithms," 2015.