© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

http://penerbit.uthm.edu.my/ojs/index.php/jscdm
e-ISSN : 2716-621X

Journal of Soft Computing and Data Mining

Depression Prediction Using the Classification and Regression Tree (CART)

Khalisa Helwani Hasni[1], Suhaila Mohd Yasin[1*]

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

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2022.03.01.003
Received 01 February 2022; Accepted 01 April 2022; Available online 01 June 2022

Abstract: Depression is a mood disorder that involves the continuing feeling of sadness and loss of interest. The crucial life events for an individual, such as losing a job may lead to depression. However, the feelings of grief and sadness are clinically diagnosed as part of depression only if the symptoms persist for at least two weeks. Eventually, depression can last for several weeks, months, or years. Some symptoms of depression may overlap with other somatic illnesses and cause difficulty in diagnosing it. This research aims to use the developed forecast model to predict future depression cases and it uses classification and regression tree (CART) of data mining approach, to predict or classify whether an individual suffers from depression or not. The dataset that was used in this research is the depression dataset from the Dataset of Students' Mental Health at an international university in Japan. This dataset consists of 268 numbers of instances and it has 10 attributes. In addition, to acquire the results, the machine learning software that was used is R Studio and the language that was used is R Programming. Besides that, evaluation metrics were used to evaluate the performance of the forecasted model and the evaluation metrics that were used were accuracy, precision and recall. From the research, it shows that the value for accuracy is 0.50(50%), precision is 1.00 (100%) and recall is 0.50 (50%). Following that, it shows that this forecasting model has the highest value of precision which is 1.00(100%). Furthermore, from the data, it also shows that teenagers in the age range from 18-22 are most likely to get depression and they also have the intention of suicide. Lastly, in the future, this research could be continued with more training on different datasets and more different techniques could be used. Besides that, this research could be improved by adding other algorithms to best understand the strengths and weaknesses of other techniques.

Keywords: Depression, data mining, training and testing, classification, regression tree

1. Introduction

Data mining in healthcare has been an emerging field of great importance over the past few years [1]. This approach is constantly increasing and becoming more popular. Data mining uses pattern recognition techniques to extract important patterns and trends from data, for example, with the aim of predicting the outcomes [2]. The data mining tools were first developed to help scientists find meaningful relationships or patterns from a tremendous amount of data that, if done, would consume significant time and resources. Besides, healthcare organizations have extensive information available to scientists or researchers to research and analyze the data provided. To predict future data so that the final result can be estimated, big data needs to be interpreted correctly by them.

Depression is the most common form of mental illness that relates to impaired overall functioning, decreased quality of life, and various health problems [4], [5]. Depending on the number and severity of symptoms, a depressive episode

**Corresponding author: ysuhaila@uthm.edu.my* can be categorized as mild, moderate, or severe. An individual suffering from mild persistent depression tends to display a depressed mood, loss of interest and enjoyment, and increased fatigability [9]. In addition, the individual who suffers from severe depressive disorder tends to have a severe depressive episode, loss of self-esteem or feelings of uselessness or guilt, and having suicidal thoughts that are likely to be prominent.

Based on the study from different sources [2], the researcher conducted research to predict possible future depression cases by using synthetic data. The method that was used in this research was the C4.5 algorithm. Some problems may arise from using that method, such as the inaccuracy of the data collected. Then, this will lead to another problem: the result for the prediction may not be precise. Lastly, the problem that may arise is that the most important attribute or symptoms might not be identified.

The main objectives of this research were to design a depression forecasting model using the Classification and Regression Trees (CART) method of data mining, implement the forecasting model for prediction tasks, and evaluate the performance of the developed forecast model. The scope of the research is the data that will be used in the data set of students' mental health at an international university in Japan [6]. This data consists of 268 numbers of instances and 10 attributes.

The expected outcome of this research is that important attributes for depression prediction will be identified. By identifying these important attributes, new knowledge on depression is discovered, and then they can be used for depression prediction by using the developed forecast model. Besides that, the trends from the data sets can be identified. This trend will be beneficial for future forecasting efforts. The other expected outcome of this research is to use the forecast model developed by using CART of data mining to predict the upcoming depression disorder. This prediction may be helpful for the individual to know whether they suffer from depression or not. Lastly, from this study, the performance of the developed forecast model for depression prediction can be identified by evaluating the forecast model.

2. Literature Review

2.1 Studies on Depression

Kevin Daimi [2] contributes to depression diagnosis by applying the classification of data mining to predict patients who will most likely develop depression or are currently suffering from depression [1]. For this existing research, this researcher used synthetic data created using a Java program. The algorithm used is the C4.5 decision tree, and WEKA was implemented and slightly improved the version, namely C4.5 revisions 8 and also referred to as J4.8. The results of implementing the depression classification model were obtained using J4.8. This research showed that 92.5% of the instances were correctly classified.

Van Loo and her team [7] used two data mining techniques: ensemble recursive partitioning and Lasso generalized linear models (GLMs), followed by k-means cluster analysis. It is used to search for subtypes based on index episode symptoms predicting subsequent MDD courses in the World Mental Health (WMH) surveys. This research showed that the high-risk cluster accounted for 52.9% – 69.7% of high persistence and severity. The result also suggests that data mining methods can make useful MDD subtyping distinctions.

Richard Dinga [3] evaluated the predictive value of a wide range of clinical, psychological, and biological characteristics for predicting the course of depression and aimed to identify the best set of predictors. A penalized logistic regression was used to predict depression courses and evaluate the predictive value of individual variables. This research showed the patients with and without a unipolar depression diagnosis at 2-year follow-up with 0.66 AUROC and 62% balanced accuracy.

Subhagat Chattopadhyay [8] attempts to mathematically model how psychiatrists clinically perceive the symptoms and diagnose depression states. A hybrid system of Mamdani's Fuzzy logic controller (FLC) on a Feet Forward Multilayer Neural Net (FFMNN) has been developed. This study concluded that the hybrid controller could diagnose and grade depression with an average accuracy of 95.50%.

2.2 Comparative Study of the Existing Research

Table 1 shows a detailed comparison of the parameters used in the research between the existing research and the proposed research.

Table 1 - Comparative study of the existing research

Dataset	Evaluation Metrics	Prediction Algorithms
Synthetic Data [1]	Accuracy, Precision, and Recall	C4.5 Algorithm
World Mental Health (WMH) Surveys [7]	Persistence	k – means cluster
Clinical Data [3]	Accuracy	Machine Learning

Real-world adult depression cases [8]	Accuracy	neuro-fuzzy Approach
Students' Mental Health Survey at an international university in Japan [6]	Accuracy, Recall, and Precision	Classification and Regression Trees (CART)

3. Methodology/Framework

The Knowledge Discovery in Database (KDD) process was used as the methodology in this research. The KDD process [10] refers to the overall process of finding, identifying, and extracting knowledge in data, and emphasizes applying particular data mining methods. Data mining algorithms will be used to extract this useful knowledge from data in the context of large databases. The goal of KDD is to discover relevant and new knowledge through the use of algorithms. There are nine steps involved in the KDD process. Figure 1 shows the KDD process methodology.

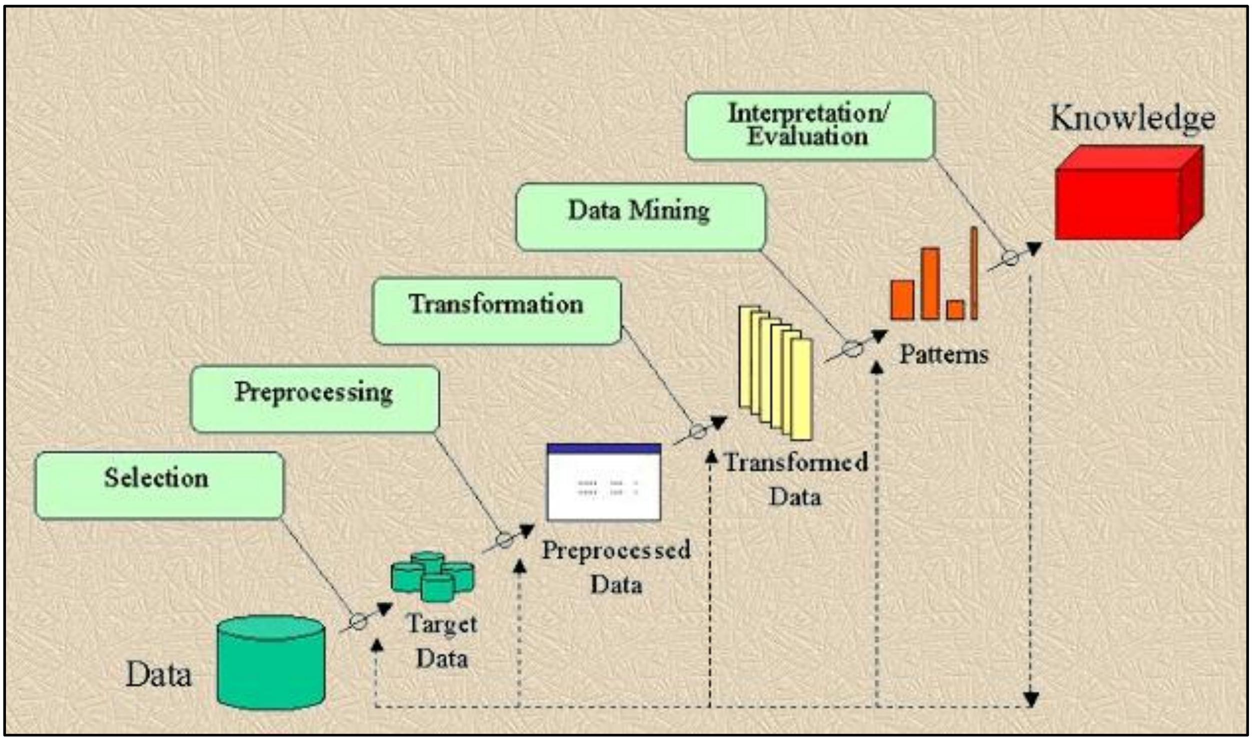
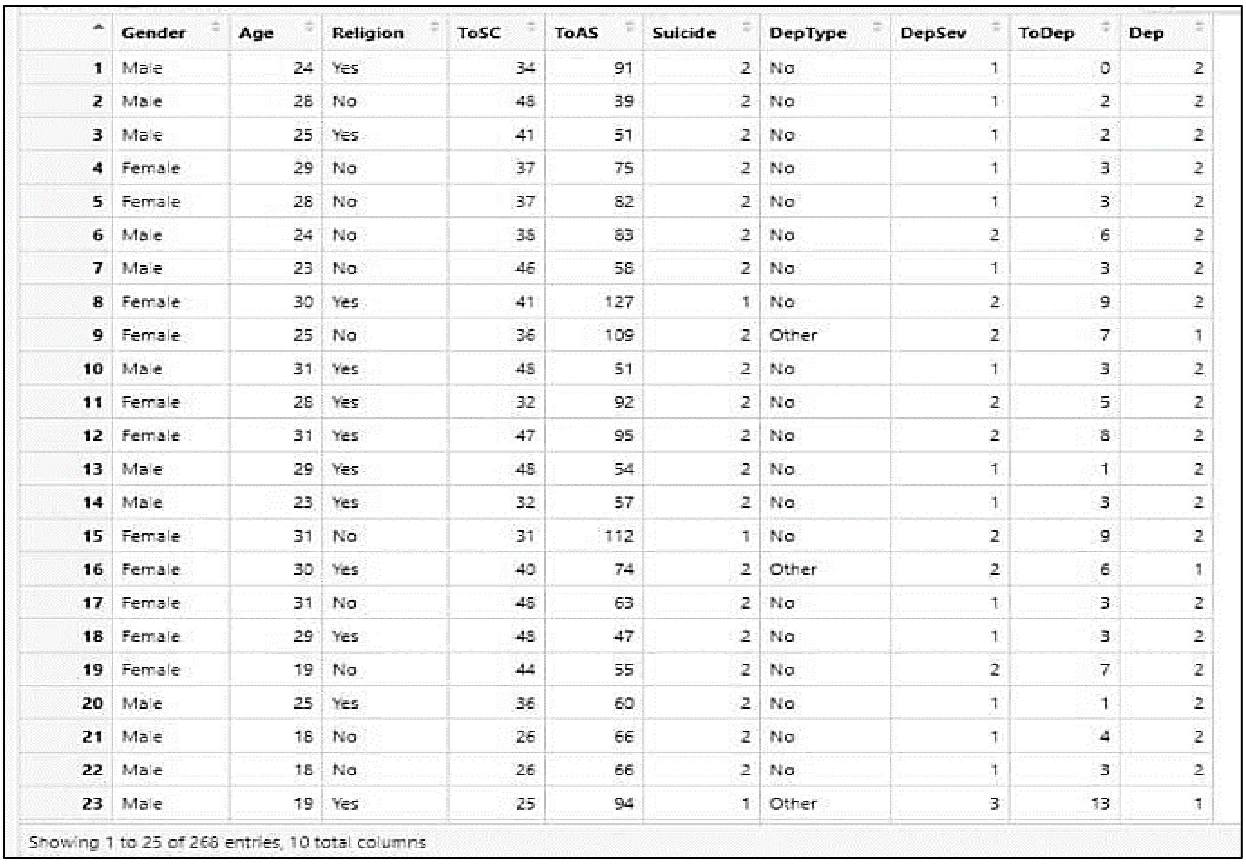


Fig. 1 - KDD Process [10]

3.1 Knowledge Discovery in Database (KDD) Process

KDD is an iterative process where evaluation measures may be enhanced; new data can be integrated and transformed in order to get more relevant results. Below are the steps in KDD process:

- Data selection, where the data that is used for this research is the data set of the students' mental health at an international university in Japan. This data consists of 268 numbers of instances and 10 numbers of attributes.



	Gender	Age	Religion	ToSC	ToAS	Suicide	DepType	DepSev	ToDep	Dep
1	Male	24	Yes	34	91	2	No	1	0	2
2	Male	28	No	48	39	2	No	1	2	2
3	Male	25	Yes	41	51	2	No	1	2	2
4	Female	29	No	37	75	2	No	1	3	2
5	Female	28	No	37	82	2	No	1	3	2
6	Male	24	No	38	83	2	No	2	6	2
7	Male	23	No	46	58	2	No	1	3	2
8	Female	30	Yes	41	127	1	No	2	9	2
9	Female	25	No	36	109	2	Other	2	7	1
10	Male	31	Yes	48	51	2	No	1	3	2
11	Female	28	Yes	32	92	2	No	2	5	2
12	Female	31	Yes	47	95	2	No	2	8	2
13	Male	29	Yes	48	54	2	No	1	1	2
14	Male	23	Yes	32	57	2	No	1	3	2
15	Female	31	No	31	112	1	No	2	9	2
16	Female	30	Yes	40	74	2	Other	2	6	1
17	Female	31	No	48	63	2	No	1	3	2
18	Female	29	Yes	48	47	2	No	1	3	2
19	Female	19	No	44	55	2	No	2	7	2
20	Male	25	Yes	36	60	2	No	1	1	2
21	Male	18	No	26	66	2	No	1	4	2
22	Male	18	No	26	66	2	No	1	3	2
23	Male	19	Yes	25	94	1	Other	3	13	1

Showing 1 to 25 of 268 entries, 10 total columns

Fig. 2 - Depression dataset from the data set of the student mental health at an international university in Japan [6]

- Data cleaning, where this phase deals with the missing values, cleaning the noisy data, where noise is the random or variance error in the data set.

- Data transformation, where during this stage, data will be transformed from one format to another format, which is more appropriate for data mining.

- Data mining, where the CART technique will be applied during this phase to extract functional patterns.

- Evaluation, during this phase, patterns will be evaluated to identify increasing patterns representing the knowledge based on given measures for this research. The measurements that will be used are accuracy, recall, and precision.

4. Result and Discussion

In this research, the objective was to implement the forecasting model using the CART method of data mining and evaluate the model's performance. In order to evaluate the performance of the model, three parameters were used: Accuracy, Precision, and Recall. A classification experiment is conducted using R Studio, and the language that is used is R programming. The dataset that was used for this experiment is the depression dataset. Meanwhile, a data splitting method is used in this experiment. The measurements to evaluate the performance of the forecasted model were determined using the formulas described below.

- **Accuracy**: The total number of samples correctly classified to the total number of samples classified. The formula for calculating accuracy is shown in Equation 1.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{1}$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

- **Recall**: It is the number of samples classified as positive divided by the total sample in the testing set positive category. The formula for calculating recall is shown in Equation 2.

$$Recall = TP / (TP + FN) \tag{2}$$

- **Precision**: It is the number of samples categorized positively classed correctly divided by the total samples that are classified as positive samples. The formula for calculating precision is shown in Equation 3.

$$Precision = TP / (TP + FP) \tag{3}$$

4.1 Result and Analysis Based on Depression Dataset

For the depression dataset, data splitting and data cleaning methods get the best performance result in Accuracy, Precision, and Recall. By using data cleaning, missing values can be identified. For this research, there is no missing value; hence only data splitting was used. The depression dataset that is used has 268 instances. The dataset is split with a ratio of 70:30. Seventy percent of the dataset was used as the training data, while 30 percent of the dataset was used as the testing data. Based on the 30 percent of 268 data, 78 instances are obtained. Table 2 shows the data splitting.

Table 2 - Train and test split dataset

Dataset	Train	Test
Dataset (Depression)	70%	30%

To create the model, a classification and regression tree algorithm is applied. The algorithm was trained using 190 instances datasets and tested with 78 representative datasets. For this research, numerical values are used to implement the model using the CART method. The attribute chosen is Class, which is the severity of depression, which indicates 1 for minimum, 2 for mild, 3 moderates, 4 for moderate-severe, and 5 for severe. Based on Table 3 , the value for both accuracy and recall is 0.50(50%), and the value for precision is 1.00(100%). Hence, it shows that the forecasting model has the highest precision value of 1.00(100%). Following that, the factor that may affect the result was the ratio for the data splitting, where the data may be split into a ratio of 70:30 and 80:20. The result may differ according to the ratio. Table 3 shows the performance result based on accuracy, precision and recall.

Table 3 - Performance results for depression

Evaluation Metrics	CART
Accuracy	0.50
Recall	0.50
Precision	1.00

4.2 Performance Comparison with Benchmark Datasets

For this research, the benchmark that is used is adopted from a previous work by Kevin Daimi, Using Data Mining to Predict Future Depression Cases (2014) [1]. The work uses accuracy, precision, and recall for the evaluation metrics. Table 4 shows the comparison between the previous paper and the proposed research.

Table 4 - Comparison of the evaluation metrics between previous and proposed research

Evaluation Metrics	Previous Research [1]	Proposed Research
Accuracy	0.833	0.50
Recall	0.803	0.50
Precision	0.858	1.00

In summary, this research reports the highest precision performance of 100%, followed by the accuracy and recall performance which both have the values of 50%, respectively. These evaluation metrics are important as accuracy refers to the number of data points that are predicted correctly, precision refers to the percentage of the relevant results, and recall refers to the percentage of total relevant results correctly classified by the algorithms. By comparing the result of the benchmark work or the previous paper, this research has the highest value for precision compared to the benchmark work. This benchmarking is used to compare the performance in order to achieve continuous improvement of the performance. Besides that, based on the data from the dataset also shows that teenagers aged 18-22 are most likely to get depression and have the intention of suicide.

5. Conclusion

In conclusion, this study is conducted to help design the forecasting model to predict the possible causes of depression. Besides that, this study is also conducted to evaluate the performance of the forecasting model by using the parameters mentioned before: accuracy, recall, and precision. This prediction is useful for a researcher to determine whether the dataset uses precise and accurate indicators for depression. In addition, the research is conducted using a depression dataset extracted from the Dataset of Students' Mental health from an international university in Japan (Nguyen et al., 2019). Data cleaning and splitting were conducted earlier in the experiment to obtain the best performance result. Then, the data set is split into training and testing data with a ratio of 70:30. Following that, the algorithm is trained using the trained data and tested using the test data. The result of this experiment was that it had the highest precision of 100%, followed by accuracy (50%) and recall (50%). By comparing the result of the benchmark paper, this research has the highest value for precision compared to the benchmark paper.

Acknowledgement

We 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] Daimi, K. & Banitaan, S. (2014), Using Data Mining to Predict Possible Future Depression Cases, International Journal of Public Health Science (IJPHS), vol. 3, no. 4, pp. 231 – 240.

[2] van Loo, H. M., Bigdeli, T. B., Milaneschi, Y., Aggen, S. H. & Kendler, K. S. (2020). Data mining algorithm predicts a range of adverse outcomes in major depression, Journal of Affective Disorders, no. 276, pp. 945 - 953, July 2020.

[3] Dinga, R., Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach, Translational Psychiatry, no. 241, 2018.

[4] Lim, J. L. & Ng, T. P. (2012). Anxiety and depression, chronic conditions, and quality of life in an urban population sample, Journal of Personality Assessment, no. 47, pp. 1047 - 1053, doi: 10.1007/s00127-011-0420-6.

[5] Rapaport et al. (2005). Quality of life impairment in depressive and anxiety disorder, Journal of Occupational Health Psychology, no. 162, pp. 1171 - 1178, doi: 10.1176/appi.ajp.162.6.1171.

[6] Nguyen, M. et al. (2019). A Dataset of Students' Mental Health and Help-Seeking Behaviors in a Multicultural Environment., MDPI Data Descriptor.

[7] van Loo, H. M., et al. (2014). Major Depressive Disorder Subtypes to Predict Long-Term Course, Wiley Digital Archives.

[8] Chattopadhyay, S. (2017). A neuro-fuzzy approach for the diagnosis of depression. *Applied computing and informatics*, *13*(1), 10-18.

[9] World Health Organization (WHO), Depression, World Health Organization, 30 January 2020 [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. [Accessed Dec. 12, 2020].

[10] Fayyad, U. (2001). Knowledge Discovery in Databases: An Overview. In: Džeroski, S., Lavrač, N. (eds) Relational Data Mining. Springer, Berlin, Heidelberg.