Research on Student Mental Health Problems Based on Machine Learning Method

Jiawen Luo
College of General Aviation and Flight, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Keywords: Student Mental Health Problems, Machine Learning, Prediction.

Abstract: Students suffer from mental health problems have less adaptability and efficiency, often show some unusual behaviors. It is important to find these problems early and eliminate them in time. Machine learning techniques can be fully used to predict mental health problems. Such applications could identify those students who have potential mental health problems from the results of student questionnaires. Three methods are proposed to handle student mental health problems. Their performances are compared to identify the best model and hyper-parameters. The data set includes the academic situation and mental health conditions of 101 university students. There are 11 attributes in the data set. Two new attributes "Total Mental Health Issue" and "CGPA Midpoint" are created in the research to better analysis the relation between academic performances and mental health conditions. Random Forest has the best performance with n_estimators is 100 and max_depth is 6. The research shows the feasibility of predicting students' mental health problems with machine learning techniques. This technology has a broad prospect and is beneficial to the whole society.

1 INTRODUCTION

Students live in campus, and study is their main task. They are also in an age of rapid social development and full of changes and competition. Many students struggle with the difficult challenges of individuating from their birth families and managing the pressure of their studies, while others may have to take care of a lot of employment and family obligations (Pedrelli et al 2015). For many students, going to college may be a stressful time. These pressures from different aspects lead to students' mental health problems. Students suffer from mental health problems have less adaptability and efficiency, often show a lot of unadaptable, uncoordinated, irrational and even wrong behaviors. Thus, it is important to find these problems in the early days and eliminate them in time to avoid causing serious adverse consequences.

The algorithms of machine learning can be entirely utilized to forecast mental health issues. When applied, such applications would benefit the society as a tool to monitor individual aberrant behavior (Srividya et al 2018). The results of student questionnaires can be used to determine recent changes in their mental state and generate corresponding labels. Some classifiers are built based on these labels to predict mental health conditions of each student. Such applications could help us identify those students who have potential mental health problems at early stages. Thus, psychological counseling can be implemented in a timely manner. Three methods are proposed to handle student mental health problems.

The literature review section reviews and summarizes some previous articles. In the methodology section, the data set and algorithms are introduced. The result section shows the result of this research. The conclusion section is the summary of the whole research.

2 RELATED WORK

On the basis of machine learning, characteristic extraction, and data resources, Rohizah Abd Rahman, et al proposed a crucial assessment analysis in Online Social Networks on mental health detection (Rahman et al 2020). The data analysis method, comparison, challenges, and restrictions of mental health detection were analyzed to determine its propriety. The research selected 22 articles from 2770 articles between 2007 and 2018. Big data in OSNs helps detect mental

Luo, J.
Research on Student Mental Health Problems Based on Machine Learning Method.
DOI: 10.5220/0012801400003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 285-290
ISBN: 978-989-758-705-4

health problems, which shows high potential in early detection as a data source. A. B. R. Shatte, et al. adopted a scoping analysis method to quickly define the realm of machine learning in mental health (Shatte et al 2019). The extraction of data included information on mental health applications, machine learning techniques, data types, and research results. Support vector machine, decision tree, and neural network were used. The application of machine learning has shown a series of benefits to mental health, but most of researches concentrate on the identification and treatment of mental health disorders. Machine learning applications still have plenty of scope to grow in other fields.

Jetli Chung and Jason Teo used the PRISMA methodology when collecting relevant research articles and studies by searching reliable databases (Chung and Teo 2022). Researchers' challenges and limitations were reflected in the research. In addition, specific suggestions on potential future research and development were also provided. Currently, there is no model that can predict a person's likelihood of having mental health issues. Machine learning techniques could improve logistic regression of the standard prediction modelling technique. Ashley E. Tate, et al aimed to evaluate whether machine learning techniques are superior to logistic regression and create a model to forecast mental health issues in mid-adolescence (Tate et al 2020). The research used nearly 500 predictors from register data and parental report. Finally, the best preforming model is not fit for clinical use. It is not necessary to seek more complex methods and forgo logistic regression for similar studies.

Sumathi M.R. and B. Poorna identified eight algorithms and evaluated the efficacy in diagnosing five basic mental health issues with various measures (Sumathi and Poorna 2016). In order to train and detect the accuracy of the algorithms, a data set of sixty cases was collected in the research. Ayako Baba and Kyosuke Bunji obtained data from 63% responses of about 6000 undergraduate students from a Japanese national university (Baba and Bunji 2023). The research compared the results of different machine learning models, including elastic net, logistic regression, XGBoost, random forest, and LightGBM. According to the comparison, the LightGBM model performed the best. Both conditions and analyses reached adequate performance in this model.

Konda Vaishnavi, et al identified five machine learning techniques, including KNN, Random Forest, Decision Tree, etc. (Vaishnavi et al 2021). The research used several accuracy criteria to assess the accuracy in identifying mental health problems. Finally, they acquired the most accurate model based on the Stacking technique with the prediction accuracy 81.75%. Jetli Chung and Jason Teo evaluated some popular machine learning algorithms (Chung and Teo 2023). Responses to a survey taken by Open Sourcing Mental Illness were included in the data set. Machine learning techniques included Gradient Boosting, Logistic Regression, KNN, Neural Networks, and Support Vector Machine, as well as an ensemble approach based on these algorithms. Gradient Boosting reached the highest accuracy, which was 88.80%, providing a highly promising approach toward automated clinical diagnosis for mental health professionals.

3 METHODOLOGY

Random Forest, Support Vector Machine, and Logistic Regression are three machine learning methods that the study suggests using. Their performances on this data set are compared in order to identify the best model.

The classification algorithm Support Vector Machine (SVM) uses interval maximization that separates data points of different classes by finding an optimal hyperplane. Data points are mapped into a high-dimensional space as the fundamental concept of SVM, which makes it easier to separate data points. SVM is a commonly used machine learning algorithm with high accuracy and generalization ability. Finding an ideal hyperplane that optimizes the separation between different categories of data points is the aim of SVM. This distance is known as the margin, and support vectors are the most closely linked data points to the hyperplane. The following stages can be used to explain the fundamentals of SVM, mapping data points to a high-dimensional space, finding an optimal hyperplane in a high-dimensional space to increase the separation of data points from the hyperplane in several categories., classifying data points into different categories according to the hyperplane, categorizing new data points. In SVM, the mapping of data points can be implemented using different kernel functions. Gaussian, polynomial, and linear kernel functions are examples of kernel functions that are frequently used. These kernel functions can map data points into higher-dimensional Spaces, making it easier to separate data points in higher-dimensional Spaces.

Random forest belongs to the category of ensemble learning, which creates a strongly supervised model by mixing weakly supervised

models. If one of the weak models produces a wrong prediction, the others can correct the error. Bagging adopts a random sampling with return, that is, a fixed number of samples are randomly collected from the training set, but after each sample is collected, the samples are returned. T weak learners are trained by T samples, and then strong learners are generated by combination strategy. Random forest is implemented on the basis of Bagging, and the weak learner is decision tree. The fundamental goal of a decision tree, a supervised classification model, is to choose a feature value that will result in the greatest information gain for tree segmentation. The node segmentation process of the decision tree is similar to a recursive process, which finds the most critical feature according to the information gain, splits it according to this feature, and splits the data nodes of the subtree in a similar way until the feature is exhausted or all the data on this node is the same label. The random forest chooses a few features on the node, whereas the ordinary decision tree chooses an ideal feature from all n sample features to segment the decision tree. To further improve the model's generalizability, an optimal feature is chosen for tree segmentation from among the characteristics that were randomly chosen.

Such a procedure is logistic regression, which involves establishing a cost function, iteratively solving the ideal model parameters using optimization techniques, testing the results to ensure the accuracy of our solution. Although the term "regression" appears in the name of the technique, logistic regression is essentially a classification method that is primarily applied to two classification issues. The generalized linear model is a family of regressions that includes multiple linear regression and logistic regression, both of which have many similarities. The dependent variables in this family of models are varied, but the model forms are essentially the same. It is multiple linear regression if the data is continuous. Logistic regression is used if the distribution has a binomial shape. Although binary variables are more widely used and simpler to understand, binary and multi-categorical dependent variables are both acceptable in logistic regression. Binary logistic regression is the most common method in practice.

Firstly the results of these three models are generated, which include the accuracy, precision, recall, and F1 score. Then the best model is identified. Next some hyper-parameters of the model are set as different values. Finally the best hyper-parameters and the best performance of the model are recorded.

4 RESULTS AND DISCUSSION

The data set for this research was collected by Google forms in a survey in 2020. The survey was aimed to find the relationship between the academic situation and mental health of university students. Both basic information and the mental health conditions of the total 101 students were recorded in the data set. Table 1 shows the attributes of the data set.

Table 1: The attributes of the data set.

No.	Attribute	Values
1	Date&Time	8/7/2020 15:14, 13/7/2020 10:34, etc
2	Gender	Male, Female
3	Age	18, 19, 20, etc
4	Course	Engineering, Law, IT, etc
5	Year	Year 1, Year 2, etc
6	CGPA	3.00-3.50, 3.50-4.00, etc
7	Marital Status	Yes, No
8	Depression	Yes, No
9	Anxiety	Yes, No
10	Panic Attack	Yes, No
11	Treatment	Yes, No

In the preliminary data processing, the distribution of students of different courses and genders who have mental health problems is firstly visualized. Fig. 1 shows the distribution of depression. The difficulties of engineering, IT, and BCS are higher that other courses, and students of these three courses suffer from the most depression. For all courses, female students are more likely to have depression than male student. The distributions of anxiety and panic attack also have similar situations. It can be concluded that female students and students of difficult courses such as engineering, IT, and BCS are more likely to suffer from mental health problems.

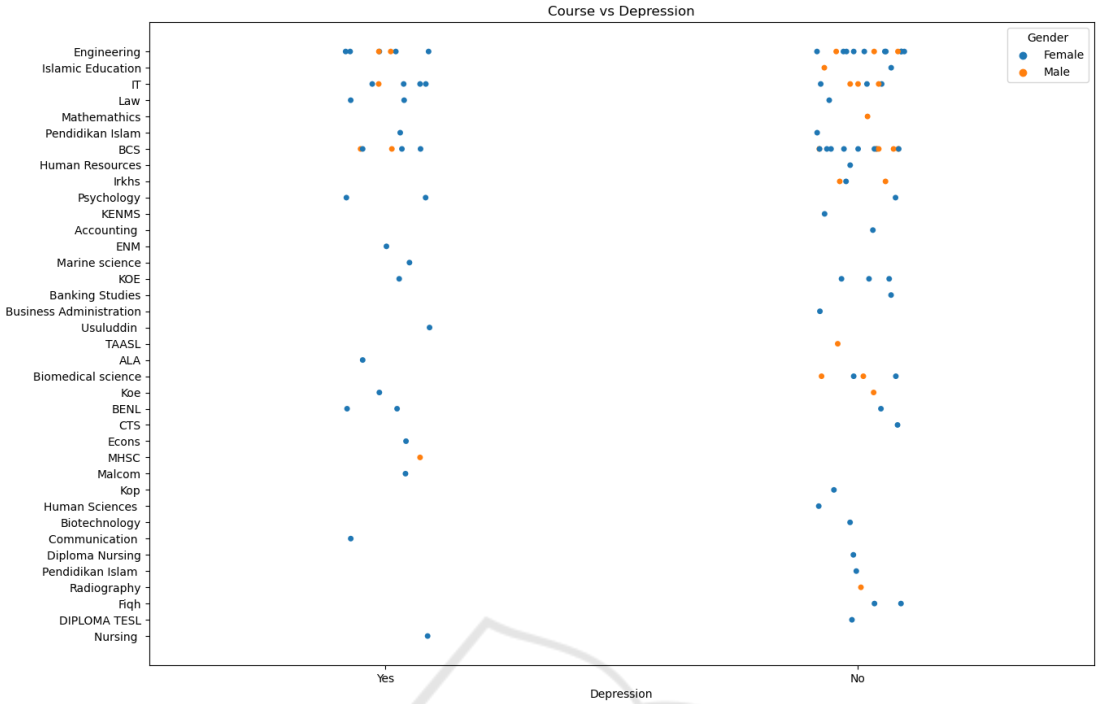

Figure 1: The distribution of course, gender, and depression (Picture credit: Original).

The distributions of students of different ages and years who have mental health problems are showed in Fig. 2, Fig. 3, and Fig. 4. Different figures correspond to different mental health problems. Students in year 1 aged between 18 and 20 suffer from all these three mental health problems. Students in year 4 do not experience anxiety, depression, or panic attack except for those who are aged 24. In year 3, elder students are more likely to have anxiety, while younger students are more likely to have depression. Students of all ages in year 2 experience depression and panic attack, but only those who are aged between 18 and 20 experience anxiety.

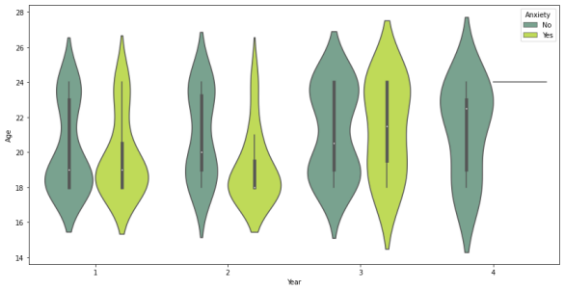

Figure 2: The distribution of year, age, and anxiety (Picture credit: Original).

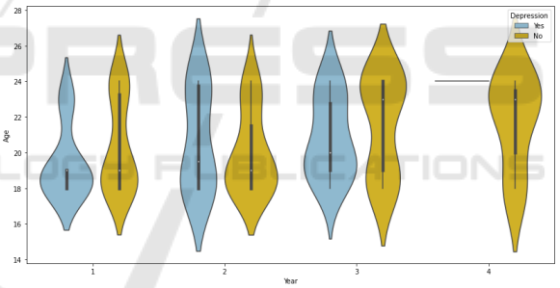

Figure 3: The distribution of year, age, and depression (Picture credit: Original).

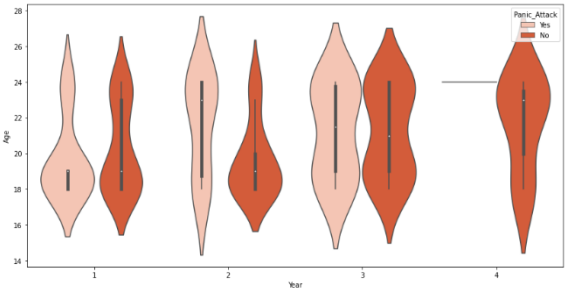

Figure 4: The distribution of year, age, and panic attack (Picture credit: Original).

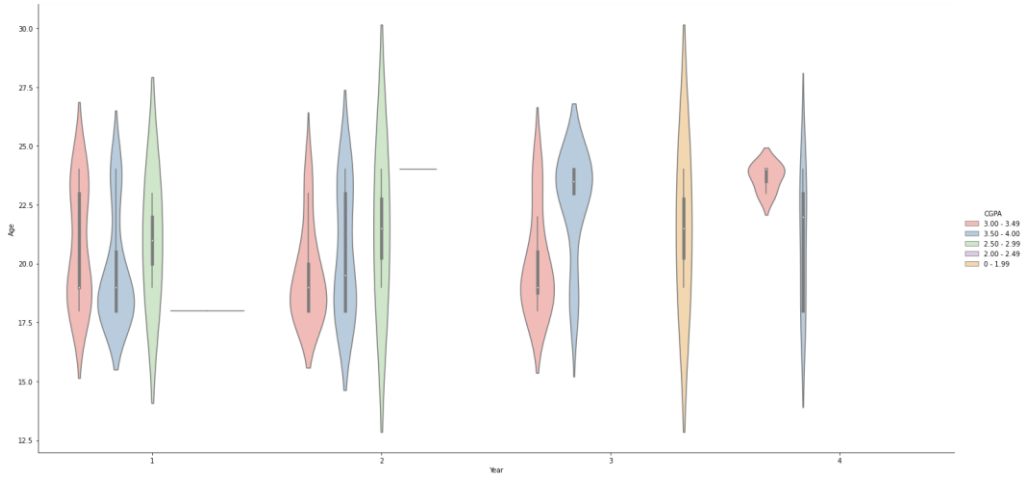

Figure 5: The distribution of year, age, and CGPA (Picture credit: Original).

Table 2: The results of three models.

	SVM	Random Forest	Logistic Regression
Accuracy	0.333333	0.904762	0.761905
Precision	0.333333	0.904762	0.761905
Recall	0.333333	0.904762	0.761905
F1 score	0.333333	0.904762	0.761905

The relation among students' CGPA, years, and ages is showed in Fig. 5. Combining the distributions of CGPA and mental health problems, it can be found that students in year 4 have the best academic performance, and they suffer from the least mental health problems. Top students in year 3 also perform well, but there are still many students have their CGPA under 2.0. That is why students in year 3 experience much more mental health problems than students in year 4. Students in year 1 and 2 have their CGPA all above 2.5, so they experience less mental health problems than students in year 3. It can be concluded that those students who have better academic performances suffer from less mental health problems.

Through the data processing, it can be found that the students' mental health conditions have a close relation with their academic performances. Thus, two new attributes are created to better analysis the relation between academic performances and mental health conditions. The "Total Mental Health Issue" attribute is combined with three attributes, "Depression", "Anxiety", and "Panic Attack". It can present the total mental health issue of each student. Another one is the "CGPA Midpoint" attribute from "CGPA". It transforms an interval value into a specific value which can increase efficiency when dealing with the data.

The accuracy, precision, recall, and F1 score of three models are all carried out and showed in Table 2. Considering false positives and false negatives are both important in this research, the F1 score is the decisive target. The value of SVM is the lowest, only 0.33. The value of Logistic Regression is much higher and reaches 0.76. Random Forest obtains the highest value, which is 0.90. Random Forest might be a good choice for a precise model.

In order to obtain the highest F1 score of Random Forest, four different values of n_estimators, 50, 100, 200, 300, and four different values of max_depth, 2, 4, 6, 8 are used. The values of various n_estimators and max_depth can be found in Table 3. For different values of max_depth, when n_estimators is 100 the F1 score is always higher than others. The highest point appears when max_depth is 6. The F1 score even reaches 1. Thus the best hyper-parameters are also identified, n_estimators is 100 and max_depth is 6.

The top Random Forest model has an extremely high F1 score. That is probably because the data set is too small and the model is over-fitting. The data could not present general conditions in daily life. When putting other data sets in the best model, the results may not be so satisfying. To solve this problem, some larger data sets should be adopted in the research.

Table 3: The values of different n_estimators and max_depth.

N_estimators Max_depth	50	100	200	300
2	0.714286	0.857143	0.809524	0.761905
4	0.904762	0.952381	0.904762	0.904762
6	0.952381	1	0.952381	0.952381
8	0.952381	0.904762	0.952381	0.904762

5 CONCLUSION

The research shows the feasibility of predicting students' mental health problems with machine learning techniques. Nowadays, students are under increasing pressure, more and more students suffer from mental health problems. Students have potential mental health problems can be identified at early stages and treated in time. This technology has a broad prospect and is beneficial to the whole society. However, the data set is too small, which could not present general conditions. The research just selects three models from machine learning techniques. There are still many machine learning models and even deep learning models. In the future, an online questionnaire will be made in order to collect more students' information and obtain a larger data set. The research will adopt more models, especially deep learning models.

REFERENCES

P. Pedrelli, M. Nyer, A. Yeung, C. Zulauf, and T. Wilens, "College Students: Mental Health Problems and Treatment Considerations," Academic Psychiatry, vol. 39, pp. 503–511, 2015.

M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms,"Journal of Medical Systems, vol. 42, no. 88, 2018

R. A. Rahman, K. Omar, S. A. M. Noah, M. S. N. M. Danuri, and M. Al-Garadi, "Application of Machine Learning Methods in Mental Health Detection: A Systematic Review." IEEE, 2020, pp. 183952 - 183964.

A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," Psychological Medicine, vol. 49, no9, pp. 1426 - 1448, 2019.

J. Chung, and J. Teo, "Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges," Applied Computational Intelligence and Soft Computing, pp. 9970363, 2022.

A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and Ralf Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," PLoS ONE, vol. 15, pp. e0230389, 2020.

Sumathi M.R., and B. Poorna, "Prediction of Mental Health Problems Among Children Using Machine Learning Techniques," International Journal of Advanced Computer Science and Applications, vol. 7, pp. 552-557, 2016.

A. Baba, and K. Bunji, "Prediction of Mental Health Problem Using Annual Student Health Survey: Machine Learning Approach," JMIR Ment Health, vol. 10, pp. e42420, 2023.

K. Vaishnavi, U. N. Kamath, B. A. Rao, and N. V. S. Reddy, "Predicting Mental Health Illness using Machine Learning Algorithms," AICECS, 2021.

J. Chung, and J. Teo, "Single classifier vs. ensemble machine learning approaches for mental health prediction," Brain Informatics, vol. 10, pp. 1, 2023