Prediction of Mental Health Among University Students

Fadhluddin Sahlan[1], Faris Hamidi[2], Muhammad Zulhafizal Misrat[3], Muhammad Haziq Adli[4], Sharyar Wani[5], Yonis Gulzar[6]

[12345]Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia

[6]Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa, Saudi Arabia

fadhluddinsahlan@gmail.com, fizalmisrat@gmail.com, hamidifaris@gmail.com, haziqadlizamzuri@gmail.com, sharyarwani@iium.edu.my, ygulzar@kfu.edu.sa

Abstract— Mental Illness or mental disorder is defined as a health condition that changes a person's thinking, feelings, or behaviour which can cause the person to distress and face difficulty to live normally. Unlike other diseases, mental illness doesn't only harm the affected individual, but also others around them. Early detection of mental disorders can help to avoid severe consequences. Mental wellness issues have been largely reported in university students. This paper explores the state of student well-being and uses various machine learning algorithms for the prediction process based on the data of entrepreneurial competency in university students. The results indicate the choice of major and gender has a significant impact on a student's well-being. Decision Tress perform better than KNN and SVM presenting an accuracy and F1-score of 0.64 and 0.61 respectively.

Keywords— Mental Health, Students, Decision Tree, K-Nearest Neighbors, Support Vector Machine.

I. Introduction

Mental illness or mental disorder is defined as a health condition that changes a person's thinking, feelings, or behaviour. It causes the person distress and difficulty to live a normal life. Mental disorder includes anxiety disorder, depression, eating disorder, personality disorder, post-traumatic stress disorder, psychotic disorder and many others [1]. Mental disorder can bring a lot of harm either towards the person who suffers from it or other people around. Extreme cases of mental disorder may lead to committing suicide, involvement in crime, or harming others [2]. This shows the severity of mental disorders and the importance of mental health care.

Mental health issues should never be taken lightly. However, these are considered as passing by events by some to the extent of ridiculing the suffering person. This worsens the situation and might to lead to extreme consequences for the suffering person. This is especially true for young people and teenagers.

Mental health issues are hard to diagnose. Many a times the affected individual does not realize the state of his well-being. A lot of university students tend to undergo mental health issues at various stages of their education[3]. This becomes even more critical as they approach the end of their studies and look forward to future prospects. Market financial situations, especially due to the effect of natural situations like Covid-19 [4], make it worse for many of the students. The pressure of job or entrepreneurship is yet another critical factor that affects the mental well-being of the students. This research is primary motivated to study the state of mental well-being in university students leveraging upon various personality traits and field of specialization. Based on these features, this research develops prediction models to determine potential mental health issues among the university students.

II. Related Work

College students were found vulnerable to anxiety due to challenges in college life. This was found from a sample of 917 students which focused on various stress factors and screening tools for anxiety. Among the various machine learning algorithms, the highest accuracy was reported by the neural network at 74% [5].

Reference [6] presents prediction of mental wellness using the data mining approach in working people. The authors use Decision Tree, Naive Bayes and Random Forest, with decision tree performing at a highest accuracy of 82.2%. The researchers stress that evaluation of mental wellness is extremely critical to understand and suggest therapies for patients with a deviated mental behaviour.

Self-consciousness about higher BMI can lead to mental issues. The relationship of BMI and mental well-being has been studied in [7] . Higher BMI leads to low self-confidence, and affects the mental health of a person. The authors suggest this can be used as an effective measure by physiatrists to detect mental illness in people who are having higher BMI.

K-Nearest Neighbor was presented as an effective modelling technique for classification of mental well-being of employees in [8]. The result indicates highest accuracy of 0.85 at a k value of 17. Various machine learning models were used for prediction of mental issues in work place settings in [9]. Decision trees revealed gender, family history and workplace medical support as prominent features influencing stress. Boosting algorithms reported the highest accuracy for prediction.

Reference [10] discussed a mental illness detection method using social media. This paper used data collected from Twitter using Twitter API and analysed it to detect a mental illness. This paper explains the system called MIDAS, where it functions as mental illness detection based on the tweets. Predictions are made using Random Forest providing a precision of 96%. Similarly, random forest, support vector machines, neural networks and XGBoost were used to predict mental health problems in adolescents [11]. Model performance was tested using AUC score. The best model was Random Forests with AUC score of 73.90% followed by SVM with 73.60%, Neural Network with 70.50%, Logistic Regression with 70.00% and XGBoost with 69.20%.

A study to determine the correlation between reading habits and depressive tendency of university students based on the data set from university library records and mental health questionnaires results was reported in [12]. This paper compares different text categorization algorithms including kNN, SVM and naive Bayesian classifier on accuracy and time-consuming under different sample sizes. They have constructed a book classifier using a naive Bayesian classification algorithm based on a polynomial model, reporting an accuracy of 0.82. A psychological prediction model is built using linear regression and logistic regression, where logistic regression outperforms its counterpart under various error conditions.

Reference [13] monitored ten students with wearable sensors to measure the stress levels experienced by students during exams. Features of the electrocardiogram and electro dermal activity signals were used as input to various classification methods such as support vector machine, linear discriminant analysis and K-nearest neighbor among others. Results indicate recognition accuracy between 86-91% for the three states - relaxed, written exam, and oral exam.

A study conducted by [14] predicted depression in university undergraduates for the purpose of recommendation to a psychiatrist. The study aims on gaining insights about cause of depression in undergraduate university students of Bangladesh. The data for this research was collected by a survey designed after consultation with psychologists, counsellors and professors. Random Forest was found to be the best algorithm, closely followed by Support Vector Machine with similar accuracy and F-measure of around 75% and 60% respectively but Random Forest presented a better precision, recall and lower false negatives.

III. Experimental Setup

This part of the paper explains the details of steps taken to find out the answer of the research questions previously mentioned. Before starting, a general picture of the methods taken is portrayed in the figure below.

A. Dataset

The dataset used in this research is was acquired from Kaggle titled "Entrepreneurial Competency in University Students"[15]. The dataset contains data of 219 university students from India. The dataset contains related information from the students with a total of 16 variables. The examples of variables are student's course, gender, presence of mental illness, perseverance rating, competitiveness rating, self-reliance rating, self-confidence rating, etc. The data collected is essentially intended to design predictive models for entrepreneurship among students. However, this research uses the variables to study mental well-being and design prediction models for the same.

B. Data Pre-processing

Python was generally used to perform pre-processing of the data before using it for prediction. Other libraries involved Pandas, Numpy, Sickit-learn, etc. - commonly referred as the data science stack. The various steps in the pre-processing stage include:

1) *Data Cleaning*: This step involved checking any inconsistencies in data such as null values and outliers, etc. Irrelevant columns for analysis and prediction such as columns requiring natural language processing were discarded from the dataset. Other columns that might not be important to the research based on the problem domain were discarded from the dataset.

2) *Data Transformation*: This step involved checking of distinct values in each column to detect terms that are same in meaning but different in spelling. If this happens, the value of the cell needs to be converted to one form. This problem often happens in categorical variable columns.

After transforming the data into a uniform value, the data is converted into numerical format for efficient use of prediction algorithms. Categorical data that use string format need to be transformed into numbers according to their category. This was achieved by using the Scikit-learn library, using LabelEncoder. The function helps to encode the string values in a categorical column into a numerical value that represents the category. The overall process of data pre-processing is illustrated in the figure below.

C. Feature Selection

Feature selection is a crucial step before training a prediction model. High dimensionality in a dataset can cause reduction in performance. High complexity takes a long time to train the model. This necessitates removing of any unnecessary features from the list of variables. Choosing the right algorithm to perform feature selection is necessary. The algorithm that needs to be used is dependent on the dataset independent variables and the target variable. Since the target of this dataset is considered a categorical variable and all the features involved are also categorical variables, Chi Square Feature Selection method was deemed suitable.

The Chi Square test calculates the correlation of two events. The high Chi Square value indicates that the feature is highly dependent with the target variable. The chi square function was also used from the sickit-learn pre-processing library. The function accepts the independent variables and the target variable as parameters and return two arrays; the Chi Square values and the P-value. This work used Chi Square values to determine the correlation of the features with the target variable.

D. Models

The research used three algorithms for the prediction process – KNN, SVM and Decision Tree. A performance comparison of the prediction models is presented in Table 1. The aim is to find the most suitable model to be used for this research problem. The models are:

1) K-Nearest Neighbor (KNN): KNN is a supervised machine learning model. The algorithm makes assumptions about the output based on the proximity of the data to the other data points no matter whether it is numerical variables or categorical variables.

In this approach, the parameters used in KNN algorithm are Euclidean distance as the metric and number of nearest neighbors of 2 (k=2). Euclidean distance is the most widely used distance metric in KNN classifications. Hence this metric was chosen instead of the others. As for the number of neighbors, it was tested manually by changing the value of k (from 1 to 5). From the observations, changing the value of k had slightly changed the prediction result. Based on the results, the best value of k was chosen to obtain the best result.

2) Support Vector Machine: SVM is a supervised machine learning model which is capable of performing regression *and* classification problems. SVM works by drawing a line which can be linear or other forms where one point that falls at the side of the line will be labelled as one class while the points that fall at the other side of the line will be labelled as the alternative class. The position of data points will be based on the value of the features of the points.

In this approach, the parameter used in the SVM algorithm is the linear kernel. There are other types of kernels such as Gaussian Radial Basis Function (RBF), Polynomial Kernel and Sigmoid Kernel. The reason the linear kernel was chosen is because changing the kernel does not change the prediction result. Other parameters such as C Parameter and gamma were set as default as changing the values also did not change the prediction result. C Parameter allows the algorithm to decide how much it wants to penalize misclassified points while gamma is a parameter for non-linear hyperplanes and its value will determine the complexity to exactly fit the training data (higher value, more complexity).

3) Decision Tree: Decision Tree consists of nodes and edges where the internal nodes represent the features and the edges will represent the decision rule and each leaf node represents the outcome. Decision Tree is a supervised machine learning algorithm that is able to solve both classification and regression problems.

In this approach, the model will receive and fit the parameters consisting of a train dataset that has been split in the training dataset phase. Next, when making predictions the library will accept the independent test parameters and the result will be compared with the target test parameters. Decision Tree uses all feature space in a dataset.

E. Evaluation

The performance of the classification algorithms is presented in Table 1. using common evaluation metrics such as accuracy, recall, precision, F1-score and AUC.

IV. Results

This section is divided into two parts – the first section deals with some exploratory analysis about the mental health of university students. The second section presents comparative performance of the prediction models used in this research.

A. Exploratory Analysis:

The exploratory analysis presents the following conclusions: 1) 64 out of 219 students mounting to 29.22% of the student's report issues related to mental well-being. 2) Business students reported the highest percentage of 40.62%, suffering mental disorders. 3) Based on gender, 28.4% of male students suffered mental disorder while female students reported a higher percentage of 31.6%.
4) Figure 1. presents the correlation of every feature with the target variable (presence of mental disorder) using the Chi Square value. High Chi value indicates that the feature is highly correlated with the target variable.

B. Predictive Performance:

The predictive performance has been presented in Table 1, Fig 1 and Fig 2.

TABLE I
PREDICTIVE PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score	AUC
DT	0.64	0.61	0.62	0.61	0.65
KNN	0.59	0.45	0.56	0.49	0.61
SVM	0.44	0.31	0.25	0.26	0.0

V. Discussion

Based on the result of exploratory analysis, the percentage of business students that suffer from mental disorders is higher than arts, music or design and engineering students. This is consistent with other faculties as well. Female students are more likely to have mental disorders compared to the male counterparts. We decided to use percentage to compare the results because the number of students for each course is different which makes it more accurate to use percentage instead of using frequency.

Chi Square analysis reveals that the major/course is a strong factor pertaining to the student's mental well-being, followed by individual project. These are followed by self-reliance, strong need to achieve, competitiveness and The remaining features such as age, city, gender, perseverance, physical health, influence and self-confidence seem to be less relevant to the target variable. From this result the model development process is using those six most relevant features as the features for the model training desire to take initiative which can be considered as moderately affecting students' mental health.

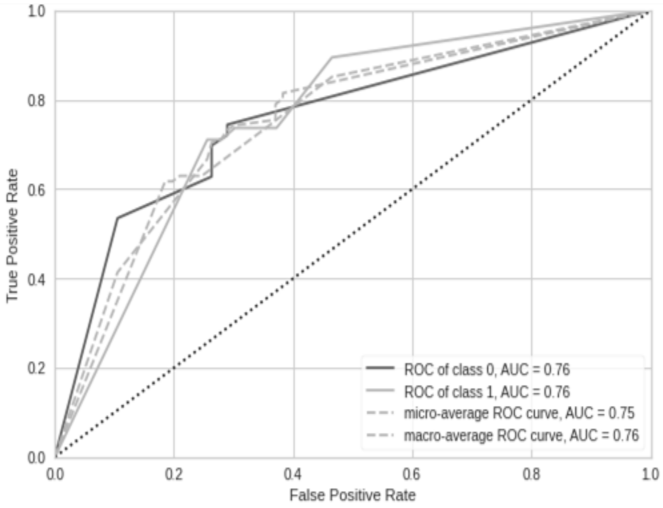

Fig. 1 ROC Curve for Decision Tree Classifier

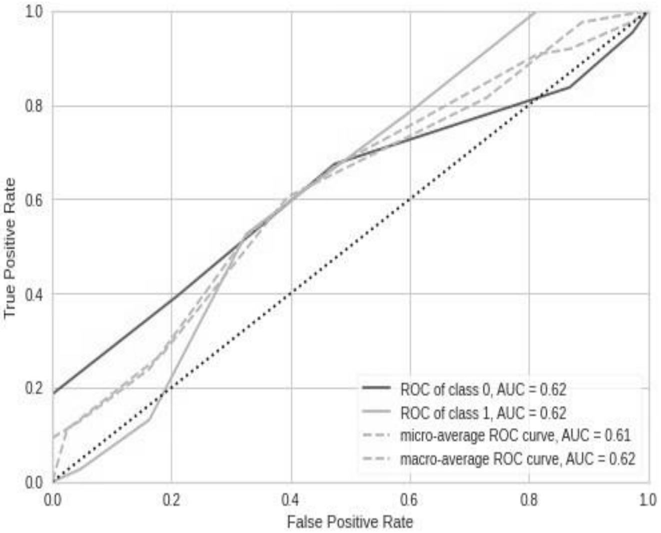

Fig. 2 ROC Curve for K-Nearest Neighbors

The predictive performance reveals that Decision Trees outperform KNN and SVM classifier for the problem at hand. Decision Tree score highest in each category with an accuracy of 64% and harmonic mean of 61%. KNN closely follows the Decision Tree for the AUC score with a difference of only 0.1. However, it lags behind Decision Tree in other metrics by a margin of 5% for accuracy and 12% for the F1-Score. SVM heavily underperforms compared to both Decision Tree and KNN. It provides as 20% lesser accuracy and 35% lesser F1-score compared to the Decision Tree predictive model.

VI. Conclusion

This paper investigated the mental disorders among university students across various majors. 30 out of 100 students reported mental disorders which is in line with the literature that university students are at higher chances of depression and anxiety as reported by [1]. Business students reported higher percentage of mental issues among different majors. Based on gender, females tend suffer more mental issues than male students. The faculty appears to have a strong effect on the mental well-being of a student.

The paper explored three predictive modelling techniques – Decision Trees, KNN and SVM for prediction of mental disorders based on the features in the dataset. The performance metrics deemed Decision Trees as the most suitable algorithm for this prediction problem in the current dataset. The data size is considerably small consisting 219 rows affecting the performance of the predictive techniques. Other models should be explored in future with more data and further data pre-processing techniques to achieve predictive performance.

References

[1] E. Chesney, G. M. Goodwin, and S. Fazel, "Risks of all-cause and suicide mortality in mental disorders: a meta-review," *World Psychiatry*, vol. 13, no. 2, pp. 153–160, Jun. 2014, doi: 10.1002/WPS.20128.

[2] D. Banerjee, J. R. Kosagisharaf, and T. S. Sathyanarayana Rao, "'The dual pandemic' of suicide and COVID-19: A biopsychosocial narrative of risks and prevention," *Psychiatry Res.*, vol. 295, p. 113577, Jan. 2021, doi: 10.1016/J.PSYCHRES.2020.113577.

[3] D. K. Cheung, D. K. Y. Tam, M. H. Tsang, D. L. W. Zhang, and D. S. W. Lit, "Depression, anxiety and stress in different subgroups of first-year university students from 4-year cohort data," *J. Affect. Disord.*, vol. 274, pp. 305–314, Sep. 2020, doi: 10.1016/J.JAD.2020.05.041.

[4] M. Shafi, J. Liu, and W. Ren, "Impact of COVID-19 pandemic on micro, small, and medium-sized Enterprises operating in Pakistan," *Res. Glob.*, vol. 2, p. 100018, Dec. 2020, doi: 10.1016/J.RESGLO.2020.100018.

[5] H. Alharthi, "Predicting the level of generalized anxiety disorder of the coronavirus pandemic among college age students using artificial intelligence technology," in *Proceedings - 2020 19th Distributed Computing and Applications for Business Engineering and Science, DCABES 2020*, Oct. 2020, pp. 218–221, doi: 10.1109/DCABES50732.2020.00064.

[6] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Classification Algorithms based Mental Health Prediction using Data Mining," Jul. 2020, pp. 1174–1178, doi: 10.1109/icces48766.2020.9137856.

[7] R. Reya Pillai, S. Saravanan, and G. Krishna Shyam, "The BMI and mental illness nexus: A machine learning approach," in *Proceedings of the International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2020*, Oct. 2020, pp. 526–531, doi: 10.1109/ICSTCEE49637.2020.9277446.

[8] H. Elmunsyah, R. Mu'awanah, T. Widiyaningtyas, I. A. E. Zaeni, and F. A. Dwiyanto, "Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor Algorithm," in *ICEEIE 2019 - International Conference on Electrical, Electronics and Information Engineering: Emerging Innovative Technology for Sustainable Future*, Oct. 2019, pp. 211–215, doi: 10.1109/ICEEIE47180.2019.8981418.

[9] U. S. Reddy, A. V. Thota, and A. Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees," Dec. 2018, doi: 10.1109/ICCIC.2018.8782395.

[10] E. Saravia, C. H. Chang, R. J. De Lorenzo, and Y. S. Chen, "MIDAS: Mental illness detection and analysis via social media," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, Nov. 2016, pp. 1418–1421, doi: 10.1109/ASONAM.2016.7752434.

[11] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," *PLoS One*, vol. 15, no. 4, p. e0230389, 2020, doi: 10.1371/journal.pone.0230389.

[12] Y. Hou, J. Xu, Y. Huang, and X. Ma, "A big data application to predict depression in the university based on the reading habits," in *2016 3rd International Conference on Systems and Informatics, ICSAI 2016*, Jan. 2017, pp. 1085–1089, doi: 10.1109/ICSAI.2016.7811112.

[13] A. Hasanbasic, M. Spahic, D. Bosnjic, H. H. Adzic, V. Mesic, and O. Jahic, "Recognition of stress levels among students with wearable sensors," May 2019, doi: 10.1109/INFOTEH.2019.8717754.

[14] A. A. Choudhury, M. R. H. Khan, N. Z. Nahim, S. R. Tulon, S. Islam, and A. Chakrabarty, "Predicting Depression in Bangladeshi Undergraduates using Machine Learning," in *Proceedings of 2019 IEEE Region 10 Symposium, TENSYMP 2019*, Jun. 2019, pp. 789–794, doi: 10.1109/TENSYMP46218.2019.8971369.

[15] "Entrepreneurial Competency in University Students | Kaggle." https://www.kaggle.com/namanmanchanda/entrepreneurial-competency-in-university-students/code (accessed Jun. 20, 2021).

Table I

Summary of Related Works

Study	Research Problem/Applications	Techniques Used	Results	Suggested Future Work
[5]	Predicting the level of generalized anxiety disorder of the coronavirus pandemic among college age students using artificial intelligence technology	7 supervised machine learning models are being	The best performance based for AUC is AdaBoost (0.943) followed by neural networks (0.936). Highest accuracy and F1 were for neural network (0.754)	None
[6]	Classification Algorithms based Mental Health Prediction using Data Mining	Decision Tree, Random Forest, Naive Bayes	Decision Tree has the highest accuracy out of 3 algorithms with 82.2%. For Random Forest and Naive Bayes, both have accuracy of 79.3% and 78.7% respectively.	Create a system to predict specific mental illnesses that the person suffers.
[7]	The BMI and Mental Illness Nexus: A Machine Learning Approach	Linear Regression	The accuracy of the model is 0.69. This shows there is a strong relationship between BMI and Mental Illness.	The relationship can help psychiatrists easily find the causes for mental disorder

[8]	Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor Algorithm	K-Nearest Neighbors	Test with variation value of K, the highest accuracy is 0.8515 with value of K = 17. As for the category of requiring mental health treatment the model correctly classifies 32 data out of 38.	None
[9]	Machine Learning Techniques for Stress Prediction in Working Employees	Logistic Regression, KNN Classifier, Decision Tree, Random Forest, Boosting	Implementation of ensemble learning(boosting) has given the highest accuracy which is 75.13%	Naive Bayes, Convoluted Neural Networks can be used
[10]	MIDAS: Mental illness detection and analysis via social media	Twitter API, TF-IDF, Random Forest	Used a 10-fold cross validation to evaluate our models. Applying only the TF-IDF features, precision of 96% for both the BP and BPD models was achived	None
[11]	Predicting mental health problems in adolescence using machine learning techniques	Random Forests, Support Vector Machines, Neural Network, XGBoost, Logistic Regression	Model performance was tested using AUC score. The best model is Random Forests with AUC score of 73.90% followed by SVM with 73.60%, Neural Network with 70.50%, Logistic Regression with 70.00% and XGBoost with 69.20%.	None
[12]	Prediction of Mental Health Problems among Higher Education Student Using Machine Learning	Linear Regression Logistic Regression, KNN, Support Vector Machines, Naive Bayes	The prediction accuracy of the two regression methods are similar. But in the smaller relative error (0.2 ~ 0.15) range, the prediction accuracy of the logistic regression model varies from 0.779 to 0.838, showing a better predictive effect. For classification algorithms, in terms of accuracy, the naive Bayes classification based on polynomial model and SVM classification have the highest accuracy, and the accuracy rate is about 0.82 when the data is 5500.	Expand the sample size to improve the accuracy of the classifier.
[13]	Recognition of stress levels among students with wearable sensors	Support Vector Machines, Linear Discriminant Analysis (LDA),	This study uses the MATLAB machine learning toolbox for the classification. Based on the	None

		Ensemble, KNN, Decision Tree	results, SVM obtained the best accuracy score with 91% followed by LDA with 90.7%, Ensemble with 89.5%, KNN with 87.6% and Decision Tree with 86.1%.	
[14]	Predicting Depression in Bangladeshi Undergraduates using Machine Learning	KNN, Random Forests, Support Vendor Machine	The model performance was tested using accuracy, precision, recall, f-measure and AUC score to determine the best model. The best model is Random Forest with 75% accuracy, 70% precision, 53% recall and 60% f-measure. In terms of AUC score, SVM gained the best score with 80.20%. There is not much difference between Random Forest and SVM but the KNN model is definitely not suitable to be used in this research.	Use a larger data set and running more algorithms to see if they can get better accuracy and lower false negatives. Moreover, they are also working on finding out the optimal features and using them to predict depression more accurately.