Prediction of Student Depression State of Mind Using Machine Learning Technique

Harshitha S[1] and Hemanth Kumar[2]
Student, MCA[1]
Associate Professor, Department of MCA[2]
Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India
h607041@gmail.com and hemanthkumar@jnnce.ac.in

Abstract*: one of the current major issues for people in the modern world is depressive disorders, the health issue is what could negatively influence people. Many students nowadays are suffering from depression. Struggling students are for one cannot see or comprehended their health problems. In this work, prediction of student depression is conducted using the method known as the Linear Regression (LR), under the domain of supervised Machine Learning (ML) techniques. Information includes social contacts, academic achievement and other types of data. It checks whether a student is depressed or not. This approach mainly applies accuracy of the predicted values using r-squared (r2) and root mean squared error (rmse).*

Keywords: Machine Learning Techniques; Prediction Model; Linear Regression(LR); Student Depression

I. INTRODUCTION

Depression among students is a mental health concern that can affect every group of people. This problem can occur to thestudents due to academic stress, social interactions, or personal problems. Depression is becoming a common thing in society. Identifying depressed students through questionnaires and clinical assessments takes a long time and fails to identify all the students who require assistance. To address this challenge, ML contributes active approach of addressing students mental health.

This work forecast depression state of mind in students. The technique used here is LR. By observing the various factors the models can be developed on prediction, that can identify the risk of the students as early stage. By noticing the various research and identifying the areas for improvement, this work strengthens a systems may help students with depressive related illnesses.

II. LITERATURE SURVEY

Troussas et. al. in [1] Has developed the ML framework to predict the sentimental analysis through facebook data using Naive Bayesian (NB) algorithm and obtained the accuracy of 77%. Reshma and Kinariwala et. al. in [2] Applied ML algorithms to analyze the people mind set. Algorithms like Support Vector Machine (SVM), NB classifier are used. And the accuracy of NB is 67%. Hussain et. al. in [3] Built a model to identify students outcomes through Artificial neural network (ANN), Logistic Regression also known as Logit Model (LM), NB, SVM, Decision Tree (DT) and resulting 75% accuracy for ANN. Ahuja et. al. in [4] Predicted the mental stress by collecting 206 students data. The ML algorithm usedin this work is SVM, Random Forest (RF), NB and K-nearest neighbor (KNN). Accuracy for RF is 83.33%, while accuracy of KNN is 55.55%. Aditya Vivek thota et. al. in [5] worked on employess stress prediction. They implemented LM, KNN classifier, DT, RF, boosting and bagging algorithms. And found that boosting had achieved highest accuracy. Hanai et. al. in [6] Performs an interaction between patient and agent by audio, video sequence to identify the patient status. By implementing F1-score algorithm they got 67% accuracy. S Samanvitha, et. al. in [7] Built the model using textdata and fetch data from different social media. The model is tested with algorithms like LM, NB, RF and SVM classifier. NB classifier gives the best results. Anju Prabha et. al. in [8] Used mechanism to identify depression among people in COVID19 situation by applying ML algorithms such as SVM, Gradient Boosting (GB) to calculate the precision of data and found that GB algorithm gave the highest exactness for the dataset. Md. Mehedi Hassan et. al.in [9] Have developed prediction models by classifying the dataset related to depression. Various

DOI: 10.48175/IJARSCT-19214

ML algorithms are employed, such as LM, KNN, SVM, and NB for building and classifying models. And got the best accuracy for K-NN which is 79%. The other algorithms such as Logistic Regression, SVM, and NB showed reliability of 77%. Shivangi Yadav et. al. in [10] Conducted ML algorithms such as KNN, DT, LR, RF Classifier, Bagging, Boosting and Stacking. Boosting algorithm give the highest result of 81.75% and the accuracy of RF is 81.22%. Ahnaf Atef Choudhury et. al. in [11] In their approach proposed predicting depression in university undergraduates and recommend them to the psychiatrist. They found RF is effective algorithm compared to SVM with accuracy around 73%. Anu Priya et. al. in [12] Proposed ML model to check the different level of stress and anxiety by adopting algorithms such as SVM, NB, RF and KNN. They also calculated different comparison factors for choosing the best algorithm and found out that NB algorithm gives the best results. Hritik Nandanwar et. al. in [13] Designed a model for depression prediction. The dataset used by them was collection of tweets from Twitter. They compared the performance of variety of ML models with labelled Twitter dataset. Different evaluation metrics like f1 score, recall and precision, Adaptive Boosting (AdaBoost) are deployed to compare the performance. AdaBoost classifier give the best results. Ananna Saha et. al. in [14] Proposed a ML model for depressed person. Implemented algorithms are RF, NB, DT and SVM, boosting technique such as LM, Bagging. Among all they got 60.54% for RF algorithm.

III. METHODOLOGY

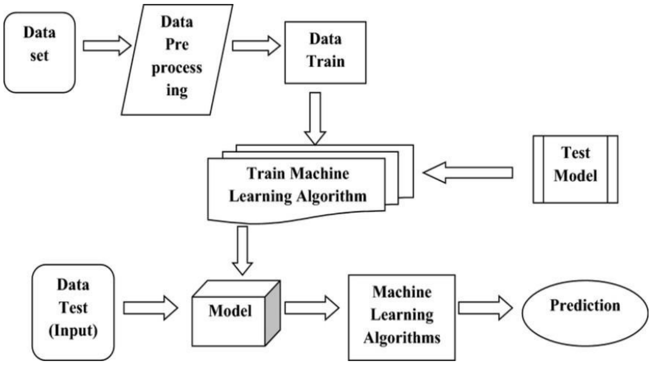


Fig. 1: Machine Learning Model

Fig. 1 illustrates the following:

- **Dataset**: Data is collected for training and testing.
- **Data pre-processing:** pre-processing is used to clean the data. For preparing a raw data, it removes the duplicatevalues.
- **Train data**: Model will be trained by using the given data.
- **ML Algorithms:** Used to learn the patterns between given variables and target variables.
- **Test data:** Examine the effectiveness of final model using provided data as output.
- **Prediction**: Result gives a developed ML model capable of processing fresh data inputs.

Linear Regression

Fig. 2 represents LR model, a quantitative technique used to predict the relationship between dependent variable and the independent variables. This method is utilized to obtain the projection of the line in model. It works based on outcomes which regulate predicted results are proportionate to the numerical values. LR algorithm is best algorithm because it showsthe clear coefficients for both dependent variable and the independent variable.

DOI: 10.48175/IJARSCT-19214

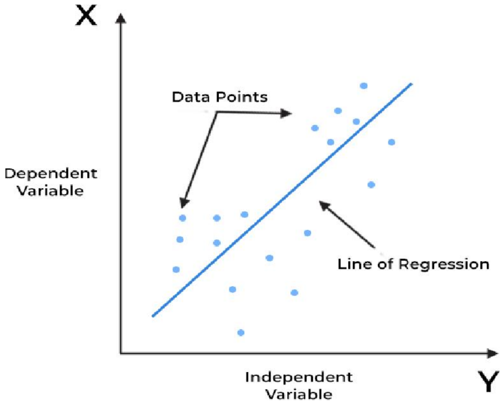
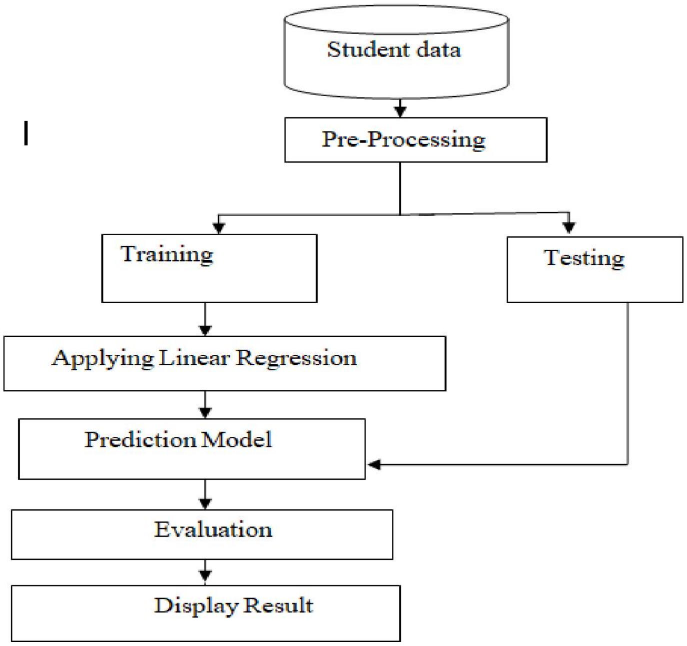


Fig. 2: Linear Regression Model

IV. PROPOSED MODEL



Flow Chart of Proposed Model

Data description

In data collection the relevant information are collected for the analysis. In this work the student activities and social interactions data are used. The data is loaded from a .csv (comma-separated value) file.

Data Preprocessing

This phase is crucial for cleaning the raw data into a perfect format. In data pre-processing, the irrelevant data is removed. The categorical features are coverted into numerical values through label encoding.

Training and Testing

Dataset is divided into two parts: training and testing. This represents how well the model performs on unseen data. A segment of 80% data is leveraged for training the model and 20% is allocating to testing. This approach ensures that the model's predictive power is assessed accurately and helps in preventing overfitting.

Applying Regression

LR is applied to visualize the relation between the variables. The model tries to give the optimal line that describes the relationship between the independent variables and the dependent variable.

V. RESULTS

The model evaluation of the LR algorithm contains a metrics, which is used to measure model performance.

A. Root Mean Squared Error (RMSE)

The RMSE is variability in prediction errors. It measures the data points of the regression line. This criterias are commonly used in regression analysis to verify results. The formula is:

$$RMSE = \sqrt{}(f\text{-}o)^2$$

Where f = forecasts (expected values or unknown results)

o = observed values (known results).

RMSE measures the average of the prediction errors, providing a clear metric of how accurately the model predicts levels of student depression. A lower RMSE indicates that the predicted depression scores are closer to the actual scores, it means the model is performing well.

R-Squared (R2)

R-squared is a statistical measure in a regression model that determines the how much a variance of the dependent variablethat can be explained by independent variable.

When anticipating student depression, R² tells how well the factors that have included in the model explain the variability in depression scores. For example, if R² is 0.8, it means 80 percent of variability in depression levels able to be accounted by the predictors in model.

$$R2 = 1.\ \frac{\sum(si-ti)^2}{\sum(si-t)^2}.$$

Where; (si−ti) is sum of squared regression

$(si-t)^2$ is total sum of squares.

Heat Map

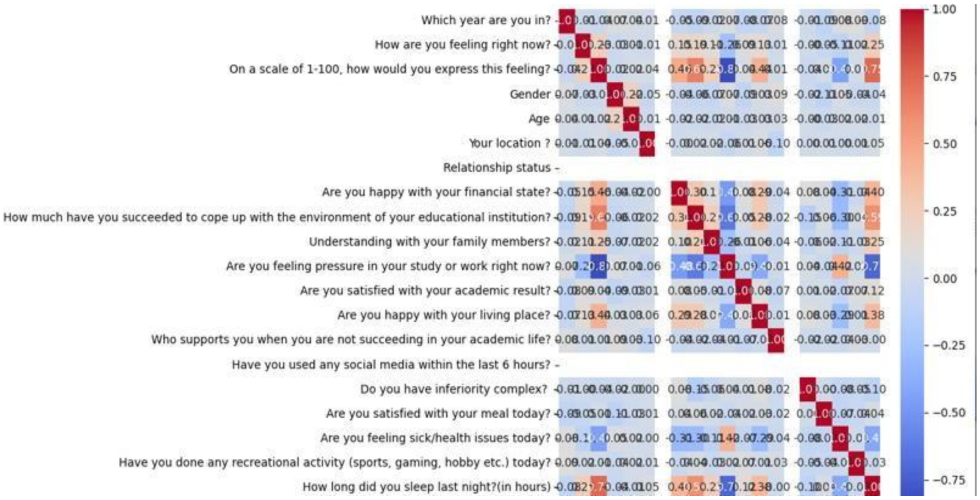


Fig. 4: Heat map of overall parameters

In Fig. 4, heatmap represent a color-coded map, shows various elements such as students emotional condition are interconnected in dark colors mean things are less connected while light colors show strong connections, this helps spot a problem which may impact how well a model predicts student feelings by showing closely different factors are linked together.

Visualization of Predicted vs Actual Values

Fig. 5 shows scatter plot of actual vs. predicted values. It is useful visualization for analyzing effectiveness of a regression model. X-axis are the actual depression scores. They represent the true state of mind of the students as recorded in the data. Y-axis are the predicted depression scores generated by the Linear Regression model. They represent the model's estimation of the students state of mind derived from input features. Each blue dot represents an individual data point where the x-coordinate is the actual depression score, and the y-coordinate is the predicted depression score of the same instance. The red dashed line represents a perfect prediction line where the predicted values exactly match the actual values

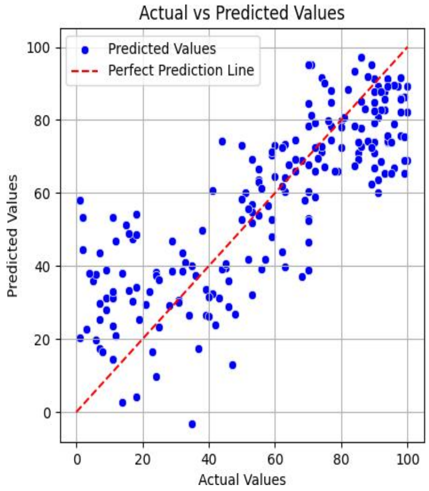


Fig. 5: Linear Plot Actual vs. Predicted values

Residual plot

The residual is a data point, employed to estimate differentiate among recorded value and prediction value of the regression model shown in Fig. 7. The positive residual indicates the actual value is higher than predicted value and negative residual indicates the actual value is lower than predicted value.

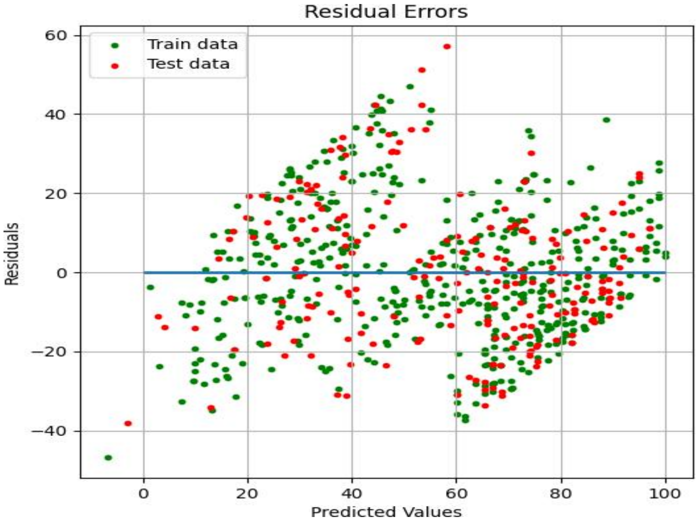


Fig. 6: Residual plot

DOI: 10.48175/IJARSCT-19214

VI. CONCLUSION

The Linear Regression model is used in this work to assess the performance level of two metrics. The R2 score is 0.66, indicating the variance of target variable. The RMSE of 17.66 indicates average prediction error. Further it categorizes the individual students based on the threshold score to find the student depression. Visualization includes scatter plots and heat maps provide a data relationships and model performance. Applying a ML algorithms can enhance the predictive accuracy results

REFERENCES

[1]. C. Troussas, K.J. Espinosa, M. Virvou. "Affect Recognition through Facebook for Effective Group Profiling Towards Personalized Instruction". Informatics in Education, 2016, Vol. 15, No. 1, 147–161.

[2]. S. K. Reshma Radheshamjee Baheti, "Detection and Analysis of Stress using Machine Learning Techniques," IJEAT, 2019, vol. 9, no. 1, pp. 2249 –8958.

[3]. Hussain, Jamil. "SNS Based Predictive Model for Depression". International Conference on Smart Homes and Health Telematics, vol. 9102, pp. 132-142, 2015.

[4]. Ahuja, Ravinder and Banga, Alisha, "Mental stress detection in university students using machine learning algorithms", Procedia in Computer Science, vol. 152, 2019.

[5]. Reddy, U. S., Thota, A. V and Dharun, "Machine learning techniques for stress prediction in working employees". International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4, 2018.

[6]. Al Hanai, Tuka, Mohammad M. Ghassemi, and James R. Glass, "Depression Detection with text and Sequence Modeling of Interviews." In Interspeech, pp . 1716-1720, 2018.

[7]. Samanvitha, S., Bindiya, A. R., Sudhanva, S., & Mahanand, B. S, "Naive Bayesian classifier for depression detection using text data". International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 418 - 421, 2021.

[8]. Prabha, A., Yadav, J., Rani, A., & Singh, V, A pilot study for "Depression detection during COVID -19 using Stroop test". International Conference on Signal Processing and Integrated Networks , pp. 45-58, October 2021.

[9]. Hassan, M. M., Khan, M. A. R., Islam, and Rabbi, M. M. F, "Depression detection system with statistical analysis and data mining approaches". International Conference on Science and Contemporary Technologies, pp. 1-6 2021.

[10]. Yadav, S., Kaim, T., & Gupta, S, "Predicting depression from routine survey data using Machine Learning". International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), pp. 163-168, 2020.

[11]. Choudhury, A. A., Khan, M. R. H, Nahim, N. Z., Talon, S. R., Islam, S., & Chakrabarty, A, "Predicting depression in Bangladeshi undergraduates using Machine Learning". IEEE Region 10 Symposium, pp. 789-794, 2019.

[12]. Priya, A., Garg, S., & Tigga, N. P, "Predicting stress in modern life using ML algorithms". Procedia Computer Science, vol. 167, pp. 1258-1267, 2020.

[13]. Nandanwar, H and Nallamolu, "Depression prediction on Twitter using Machine Learning algorithms". In Proceedings of the 2nd Global Conference for Advancement in Technology (GCAT), pp. 45-58, November 2021.

[14]. Ananna Saha, Ahmed Al Marouf, Rafayet Hossain, "Sentiment Analysis from Depression-Related User Generated Contents from social media". International Conference on Computer and Communication Engineering (ICCCE), pp. 259-264 2021.