

Depression Analysis Based on Machine Learning and Model Visualization

Xurui Zhang

Department of Applied Science, Macao Polytechnic University, Macao Special Administrative Region, China

p2320284@mpu.edu.mo

Abstract. Depression is among the most prevalent mental health issues in the world. Traditional diagnostic methods rely on clinical interviews and scoring scales, which are subjective and biased. In this paper, the adopted model is based on a massive dataset with more than 27,901 student samples to conduct experiments, where this study compares three machine learning algorithms, including Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN). The methods on SVM, RF, and K-NN preprocessed, i. e., excluding outliers through the Inter Quartile Range (IQR) method, feature encoding, and Z-score normalization procedure. The experimental results shows that all models achieved good performance, with SVM having the highest accuracy followed by RF then K-NN. This study also analyzed the confusion matrix in detail. It indicates that SVM performs well in positive class recognition, while random forest excels in negative class recognition. The research results indicate that machine learning models can capture the nonlinear relationships in depression data. They can also handle complex interactions between different factors. This is something traditional linear regression methods cannot do well. This study offers important new insight for automatic depression screening. It also contributes to the development of objective diagnostic tools in mental health care.

Keywords: Depression analysis, machine learning, mental health diagnosis.

1. Introduction

In addition to the competition and increasing competitive pressure, the rate of modern life is increasing more quickly and quickly, particularly in the workplace and research fields. Whether it is the daily work pressure or the fierce competition in study, it will make people feel tremendous mental stress. This kind of competitive anxiety often brings deep unease and self-doubt, gradually evolving into depressive moods. In today's society, depression has become a common disease [1]. A common mental condition called depression, also referred to as depressive disorder, mostly manifests as a high degree of low mood and a decreased level of interest. Depression is one of the major causes of disability worldwide, characterized by a high incidence rate, a high recurrence rate, but a low acceptance rate for treatment. From 2001 to 2020, the global prevalence of self-reported depressive symptom points increased by 34%, the prevalence of Major Depressive Disorder (MDD) was 8%, and the prevalence of elevated depressive symptoms among adolescents rose by 13% [2]. Furthermore, compared to male adolescents [3], there was a greater incidence of elevated depressed symptoms among female adolescents. In the world of psychiatric disease, the most common condition is depression [4].

At present, the diagnosis of depression mainly relies on clinical interviews and self-rating/other-rating scales such as PHQ-9 [5] and HAMD [6], etc. This conventional diagnostic approach carries a degree of subjectivity and bias. Many depression studies rely on common statistical methods such as linear regression to model and analyze the relationships between variables [7]. Although traditional linear regression methods are easy to understand and explain, they have limitations in depression research: they assume a linear relationship between the independent variable and the dependent variable, but depression and its influencing factors often exhibit nonlinear and complex interactions, making it difficult for linear models to fully capture these complexities [8]. Meanwhile, depression data may have excessive dispersion and many zero values, making it difficult for ordinary linear regression to fit, and the model may have biases [9]. Furthermore, mental health data is diverse and

heterogeneous. A single linear model is difficult to accurately reflect the characteristics of different subgroups and is sensitive to outliers and distribution assumptions. Inaccurate the data will reduce the reliability of the results, which means that the data does not meet the homogeneity of variance and normalcy [9]. All of these factors could result in a decrease in the performance of model fitting and predictions.

This study's objective was to conduct a thorough dataset-based data mining and analysis of research on depression. Using data visualization techniques and focusing on adolescent populations as the experimental group. Machine learning models were employed to automatically extract and learn patterns from diverse patient data, establishing classification models to facilitate the screening and diagnosis of depression. The use of machine learning models can decrease subjective errors, increase diagnosis efficiency and accuracy, and give objective supplementary diagnostic basis for clinical practice as compared to traditional diagnostic methods. The accuracy and robustness of depression analysis and diagnosis are improved by handling high-dimensional heterogeneous data and outliers with ease and by flexibly capturing nonlinear linkages and complicated interactions. By adaptably capturing nonlinear relationships and complex interactions, handling high-dimensional heterogeneous data and outliers, and improving the accuracy and robustness of depression analysis and diagnosis, this overcomes the drawbacks of conventional linear regression models in depression research and offers compelling support for precision medicine and intervention.

2. Method

2.1. Data description and preprocessing

In this project, this study uses the student depression dataset provided by a dataset on Kaggle [10].

The size of the original dataset is 27,901 rows \times 18 columns. The total number of missing values is 0. The distribution of depression labels was as follows: 1 (suffering from depression): 16,336 people (58.5%), 0 (No depression): 11,565 people (41.5%). The number of features is 9, such as 'id', 'Age', 'Academic Pressure', and 'Depression'. The number of classification features is 9, such as 'Gender' (2 unique values), 'Sleep Duration' (5 unique values, 'Have you ever had suicidal thoughts?' (2 unique values), etc.

The preprocessing is consisted of four parts. Since the dataset has no missing values and missing target variables, this study omitted the steps of data cleaning and handling missing values. First, this study conducted abnormal values detection and handling using the IQR method. This paper chose to replace outliers with boundary values. Among them, 12 outliers were handled in the feature 'Age', 3 outliers in the feature 'Work Pressure', 9 outliers in the feature 'CGPA', and 8 outliers in the feature 'Job Satisfaction'. Second, the classification features were encoded and divided into corresponding categories. The dataset was separated into two sections: a test set (5,581 samples) and a training set (22,320 samples). Ultimately, this paper applied StandardScaler for Z-score normalization.

2.2. Machine learning models

Three machine learning models — Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN) — were employed in this project to assess the outcome. Choose the model with the highest accuracy by comparing its performance.

2.2.1 Random forest

The random forest method improves the accuracy and stability of predictions by constructing multiple decision trees and allowing them to 'vote' [11]. Its basic principle is to use Bootstrap sampling to selectively extract multiple sub-datasets from the original data, train a decision tree for each sub-dataset, randomly select some features at each node split, and finally integrate the prediction results of all trees through majority voting (classification) or mean (regression). The overfitting risk of a single decision tree is successfully decreased by this randomness and integration approach. In the experimental model, this paper sets the parameters 'n estimators'=100, 'random state'=42, 'max

depth'=10, 'min samples split'=5, and 'min samples leaf'=2. Under this parameter configuration, during the training phase, multiple sub-datasets were constructed through Bootstrap repeated sampling. Each sub-dataset underwent random feature selection, and a decision tree was built. In the prediction phase, all trees were independently predicted, and the result was obtained through voting (classification) or taking the average (regression). Meanwhile, the significance of each feature in explaining the model's decision-making process and the extent to which each feature contributes to the prediction outcomes may also be determined using this model.

2.2.2 Support vector machine

Support vector machines' basic idea is to find the best decision boundary (hyperplane) that maximizes the margin (distance) between categories while also ensuring accurate classification, allowing only the most important boundary samples (support vectors) to affect the classification surface [12]. Its basic principle is to find the hyperplane that maximizes the interval by solving a constrained quadratic optimization problem. The original problem is converted into a dual problem for resolution through the application of the Lagrange multiplier technique. When the data is linearly inseparable, a kernel function (such as the RBF kernel) is introduced to map the data to a high-dimensional space to achieve nonlinear classification. Meanwhile, a soft boundary and a penalty parameter C are introduced to handle noise and outliers.

This paper established the support vector machine's primary parameters in this experimental model: To guarantee the reproducibility of the results, the gamma parameter utilized the auto-scaling mode (gamma='scale'), the regularization parameter C was set to 1.0, the random seed was set to 42, and the kernel function uses the radial basis function (kernel='rbf').

The entire workflow of a support vector machine included the following steps: Firstly, data preprocessing and feature scaling were carried out to ensure that features of all dimensions were on the same scale; Then selected the appropriate kernel function and parameter configuration; Next, constructed and solved the Lagrange dual optimization problem, and found the optimal solution through quadratic programming. During the optimization process, identified the key support vectors and calculated the relevant parameters of the decision function; Ultimately, the decision function (1) is used to classify and predict the new samples, where $K(x_i, x)$ is the kernel function, b is the bias term, y_i is the training sample label, and α_i is the Lagrange multiplier.

$$f(x)=\text{sign}(\sum\alpha_i y_i K(x_i, x)+b) \quad (1)$$

2.2.3 K-nearest neighbor

The core idea of K-NN is that adjacent samples in the feature space are likely to belong to the same category, so the category of a new sample can be predicted based on the labels of the nearest K neighbors [13]. This is an inert learning algorithm that does not require a training process. It directly stores all training samples. The prediction process involves calculating the distance (typically the Euclidean distance) between the new sample and all training samples, choosing the K samples with the closest distances as neighbors, and determining the prediction result by either taking the average (regression) or majority voting (classification). To increase the influence of closer neighbors, distance weights can be applied.

In this experiment, the parameters of the K-nearest neighbor algorithm were set as follows: Select 5 nearest neighbors ($n_neighbors=5$), and made predictions using distance weighting (weights='distance'), with the distance metric being the minkowski distance (metric='minkowski').

The system first stored all the training data as reference samples. When a new sample to be classified was received, the distance between this sample and all the training samples was calculated. Then, the samples were sorted by distance size and the K nearest neighbor samples were selected. Based on the label information of these K neighbors, weighted voting or taking the average was conducted to determine the prediction result. Finally, outputted the prediction results of classification or regression.

3. Results and Discussion

3.1. Model evaluation results

Table 1. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.8387	0.8382	0.8387	0.8380
SVM	0.8418	0.8413	0.8418	0.8409
K-NN	0.8219	0.8212	0.8219	0.8207

The performance evaluation of the three machine learning models is summarized in Table 1, which presents the accuracy, precision, recall, and F1-score metrics for each algorithm. All three models obtained similar performance levels for all the evaluation measures, as shown in the experimental results. With an accuracy of 0.8418, precision of 0.8413, recall of 0.8418, and F1-score of 0.8409, SVM shows the highest performance. With an accuracy of 0.8387, precision of 0.8382, recall of 0.8387, and F1-score of 0.8380, Random Forest was the second-best performance. Among the three models, K-NN performed the lowest, with an accuracy of 0.8219, precision of 0.8212, recall of 0.8219, and F1-score of 0.8207, respectively.

The confusion matrices for all three models are presented in Fig. 1, revealing the detailed classification outcomes for each algorithm. Random Forest generated 2880 true positive (TP) cases, 1801 true negative (TN) cases, 512 false positive (FP) cases, and 388 false negative (FN) cases. SVM produced 2899 TP cases, 1799 TN cases, 514 FP cases, and 369 FN cases. K-NN resulted in 2859 TP cases, 1728 TN cases, 585 FP cases, and 409 FN cases. The total number of test samples across all experiments was consistent at 5581 instances, ensuring fair comparison among the models.

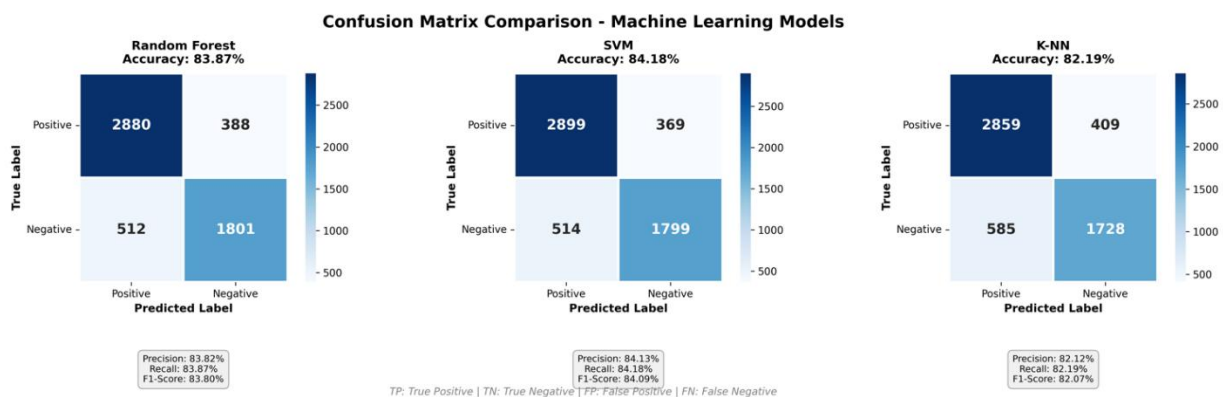


Figure 1. Confusion Matrices Comparison for Student Depression Prediction Models (Picture credit: Original)

3.2. Discussion

The experimental results reveal several important insights into the comparative performance of the three machine learning algorithms. The performance metrics of all models are very close, indicating that the dataset features are highly suitable for multiple classification methods, and no single algorithm shows an overwhelming advantage. However, upon a more careful examination of the results, subtle but meaningful differences will be discovered.

The outstanding performance of SVM can be attributed to its remarkable ability in positive class recognition, as evidenced by its highest true positive count (2899) and lowest false negative rate (369). This feature is particularly valuable in classification tasks because the lack of positive instances can have significant consequences. The model's robust decision boundary optimization, which is indicative of the mathematical underpinnings of support vector machines in marginal maximization, allows it to decrease false negatives while retaining a respectable level of accuracy.

Random forest demonstrated different intensity profiles, especially excelled in negative class recognition, with the highest true negative count (1801). This indicates that the integration method

effectively captures the features of negative samples, possibly because different decision trees provide complementary perspectives in terms of feature importance. A somewhat lower true positive count, however, suggests a more cautious classification approach as compared to SVM, which could be advantageous in situations when the cost of false positives is especially significant.

The relatively weak performance of K-NN reveals specific limitations when dealing with this particular dataset. This model tends to overpredict, as evidenced by the highest false positive count (585) and the lowest true negative count (1728), suggesting that the local community structure may not be able to fully capture the potential class boundaries. This behavior is a characteristic of K-NN when dealing with datasets where there are similar feature patterns among different categories, resulting in a fuzzy neighborhood composition and causing the prediction to be biased towards positive classes.

The differences between accuracy and recall in each model are very small, indicating a performance balance between the two classes, which suggests that there are no obvious class imbalance issues in the dataset. This balance is crucial for practical applications because both positive and negative predictions are equally important. The consistency of this pattern across all three models further validates the robustness of the experimental setup and the quality of the dataset preparation.

From the perspective of error analysis, the similar false positive rate between SVM (514) and Random Forest (512) indicates considerable specificity, while the significantly lower false negative rate of SVM has a decisive advantage in overall performance. This model indicates that the optimization objective of SVM more effectively balances the trade-off between the sensitivity and specificity of this specific classification problem.

These findings have practical significance for model selection in similar classification tasks. When the maximum recall rate is of critical importance, support vector machines become the preferred choice. Random forests offer an advantage when conservative predictions with high specificity are expected, while K-NN may require additional preprocessing or parameter tuning to achieve competitive performance. The smaller performance gap also indicates that the integrated approach combining these methods can potentially produce better results by leveraging the unique advantages of each algorithm to compensate for their respective weaknesses.

4. Conclusion

This study successfully demonstrated the application of machine learning algorithms for depression analysis and prediction. It compares 3 different machine learning approaches to reach some important conclusions on their respective applications. This study demonstrates that machine learning approaches offer promising potential for improving depression screening and diagnosis. Support Vector Machine emerged as the optimal choice. Its superior performance can be attributed to its mathematical foundation in margin maximization and its exceptional ability to minimize false negatives, which is particularly crucial in clinical settings where missing positive cases can have serious consequences. Nevertheless, the current studies could still be upgraded by adding features like deep learning architectures, ensemble methods integrating multiple algorithms and fine-tuning the feature engineering methods. The subsequent direction for future work would be focusing on finding methods, using such techniques as neural networks and recurrent architecture, which could fully explore out the whole set of information, which leads to detecting more complex patterns.

References

- [1] Moussavi S, Chatterji S, Verdes E, et al. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet*, 2007, 370: 851-858.
- [2] Andrade L, Caraveo-anduaga J J, Berglund P, et al. The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *International Journal of Methods in Psychiatric Research*, 2003, 12: 3-21.
- [3] Thapar A, et al. Depression in young people. *The Lancet*, 2022, 400 (10352): 617-631.

- [4] Penninx B W, Milaneschi Y, Lamers F, et al. Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile. *BMC Medicine*, 2013, 11: 129.
- [5] Kroenke K, Spitzer R L, Williams J B W. The PHQ-9. *Journal of General Internal Medicine*, 2001, 16: 606-613.
- [6] McIntyre R S, Konarski J Z, Mancini D A, et al. Measuring the severity of depression and remission in primary care: validation of the HAM-D-7 scale. *CMAJ*, 2005, 173 (11): 1327-1334.
- [7] Kim S W, Chang M C. The usefulness of machine learning analysis for predicting the presence of depression with the results of the Korea National Health and Nutrition Examination Survey. *Annals of Palliative Medicine*, 2023, 12 (4): 748-756.
- [8] Twenge J M, Hamilton J L. Linear correlation is insufficient as the sole measure of associations: The case of technology use and mental health. *Acta Psychologica*, 2022, 229: 103696.
- [9] Xu T, Zhu G, Han S. Study of depression influencing factors with zero-inflated regression models in a large-scale population survey. *BMJ Open*, 2017, 7: e016471.
- [10] Shamim A. Student Depression Dataset. Kaggle, 2024.
- [11] Salman H A, Kalakech A, Steiti A. Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 2024: 69-79.
- [12] Suthaharan S. Support Vector Machine. *Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems*, 2016, 36.
- [13] Zhang Z. Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 2016, 4 (11): 218.