ORIGINAL ARTICLE

PREDICTION OF DEPRESSION USING MACHINE LEARNING TOOLS TAKING CONSIDERATION OF OVERSAMPLING

Md. Murad Hossain*[1,2], Md. Asadullah[2], Mohammad Amzad Hossain[3] and Muhammad Saad Amin[4]

[1]Modeling and Data science, University of Turin, Via Verdi,8-10124 Turin, Italy.
[2]Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj 8100, Bangladesh.
[3]Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali 3814, Bangladesh.
[4]Department of Computer Science, University of Turin, Via Verdi,8-10124 Turin, Italy.

***Corresponding author: Md. Murad Hossain.**
Email: md.hossain50@edu.unito.it

ABSTRACT

Depression is a psychiatric condition characterized by a persistent sense of sadness and dullness. It is also known as a severe burdensome problem or clinical sorrow, and it impacts how a person feels, thinks, and behaves and triggers a slew of emotional and physical issues. Various components are liable for this issue, and many related sicknesses are expanding because of this infection. It is not just at risk for well-being perils, yet also produces perilous social offense, like self-destruction and family misuse. In this study, we used machine learning methods such as Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB). We also used accuracy, precision, recall, and F1-score to survey the exhibition assessment of arrangement results. These machine learning algorithms developed and analyzed Confusion matrices through data augmentation to assess the classification performance. This study used machine learning technologies to predict depression and revealed the significance of the trait. Then we have tried to utilize an oversampling technique that shows the distinction in model execution. Indeed, we wanted to see how well the recommended machine learning algorithms performed before and after rebalancing standardized data. In our suggested framework, the RF classifier performed better with 89% accuracy and 90% precision than other models.

KEYWORDS: *Depression, Machine Learning, Classification, Accuracy, oversampling.*

INTRODUCTION

People are, naturally, turning out to be goal-oriented these days and look for each reasonable chance to develop professionally. Nervousness, sadness, stress, disappointment, and dissatisfaction have become so ordinary that individuals presently trust them to be vital in daily life[1]. explains that depression is a common psychiatric disease. According to the World Health Organization, it afflicted more than 300 million individuals globally. The seriousness of the epidemic has caused many health professionals to concentrate their studies on it[2]. use six predictive models and two missing value imputation methods for constructing an inner depression foresight model for elderly individuals. Several machine-learning techniques, including logistic regression, a ridge estimator, random forest, and even two faulty data inference approaches, were utilized to examine the possibility of using machine-learning methods for a widely available dataset. But they have not considered any tuning method for enhancing the performance of the model[3]. the entire exploration sees machine learning algorithms for predicting depression by systematically defining relevant data properties uses three extraction approaches and one composite methodology, along with stat tech analysis tools, selects subsets of characteristics to eliminate redundant attributes[4]. to find key biomarkers linked to depression they used a three-step methodology that included several accusations, an LR model with machine learning, and an enhanced regression model. They examined machine learning techniques that automatically select significant data qualities for stress detection using these algorithms[5]. look at some of the most basic classification algorithms for brain scan diagnosis and detection. Ultimately, problems, prospective paths, and possible drawbacks address relevant to major depression disorder biomarker detection[6]. aim to create classification techniques for identifying the risk of postpartum depression within a week after delivery, allowing early detection, and creating a mobile health app for the Android platform. They focused on the best model for both young moms and physicians who want to keep track of their patient's test results[7]. utilize Machine learning algorithms to conjecture tension, gloom, stress, and severity levels of anxiety. Information acquired from working ruined people from different foundations, and five separate ml calculations extended their appearance on five degrees of solidarity[8].

proposed a generic application for forecasting postpartum depression risk based on data from e-health reports. A cycle of data extraction, unravelling, and information science use to choose a modest number from the electric security dataset to submitted unwavering quality and future place of time hazard expectations[9]. analyze the cross-sectional findings using the 2014 Cognitive Risk Factor Monitoring Method, uses the dataset to predict type 2 diabetes through the SVM classification scheme, DT, LR, RF, neural network (NN), and Gaussian Naïve Bayes optimization techniques. They also used univariate analysis and multivariate calibrated logistic regression analysis to evaluate the connections of adverse outcomes with type 2 diabetes[10]. use SVM, DT, NB classification techniques, and K-nearest neighbours are among the information retrieval protocols mentioned. The authors look at the effects of the above machine learning algorithms on groups and recommend future research[11]. identify a feature extraction approach. They created a pre-processing data system for detecting depression types. Also used Several statistical and machine learning approaches test through the ten-fold validation set framework[12]. evaluated significant zones in the dataset that majorly affected ordering a patient's psychological steadiness. They applied correlation analysis and selected machine learning-based strategies[13]. demonstrated how to use non - linearity Electroencephalogram (EEG) signal analysis to distinguish between depression patients and control. The models used to characterize the groups are KNN, discriminant classification, and logistic regression[14]. aimed to conduct a mental illness analysis on Facebook data obtained from a community internet site. The author proposes machine learning as an effective and helpful method to analyze the influence of emotion classification[15]. evaluate training machine learning strategies for detecting stress behaviours with data modifications, state-of-the-art ensemble learning. The writers also tried to assign class labels in a concise manner.

A couple of the papers use quantifiable research, such as fundamental structures, dispersions, and regression models, to describe wretchedness. Several papers explain the bio-maker using various approaches for authenticating and validating data. In numerous papers, it has been discussed how covid-19 increased depression when society was on lockdown. However, just a few studies used machine learning methods to augment the oversampling model. There is no such study that has attempted to identify and measure pressure using machine learning technology. In this study, we propose an oversampling strategy and employed state-of-the-art machine learning algorithms to predict depression using the important factor that causes depression. We tried to demonstrate improved performance utilizing oversampling techniques compared to the available machine learning tools. We have assessed the usefulness of numerous analysis methods throughout this research from the perspectives of description, validation, and correctness. Few authors find any effects of depression, while most authors discuss issues associated to depression. Nobody, however, specifies how to manage imbalanced data when using a machine learning algorithm to forecast target variables. In this study, we attempted to forecast depression and increase model precision by running imbalanced data.

METHODS

Data Description

We follow secondary data collection method. The link to the data is https://www.kaggle.com/diegobabativa/depression. There are a few different variables (details in **Table 1**) in this dataset such as sex, age, Married, Number_children, education_level, total_members, gained_asset, durable_asset, save_asset, living_expenses, other_expenses, incoming_salary, incoming_own_farm, incoming_business, incoming_no_business, incoming_agricultural, farm_expenses, labor_primary, lasting_investment, no_lasting_investmen, target.We utilized the R package version 4.03 for data management and analysis.

Preprocessing (Data Normalization)

For minimizing execution time and improve results, the data need to filter in the first step. We normalize the data for this reason so that the characteristics continue to follow: We used min-max function scaling (normalization) for all the aspects in this study. It is a method of re-scaling and moving aptitudes so that they end up right in the middle.

$$x_{normalized} = (x - x_{minimum}) / (x_{maximum} - x_{minimum})$$

Feature Importance Plot

The feature value specifies that the features in the data set are more practical or significant than others. Employing feature extraction might allow you to understand better the problem you have solved and improve your model in some situations. The process of assigning a value to input data based on its effectiveness in anticipating a target variable is known as feature value.

Machine Learning Technique

Random Forest (Rf)

A RF is a classification and prediction technique based on data mining and machine learning. Ensemble learning is a method for solving complicated problems that incorporates several classifiers. Many decision trees are used in a random forest method. The random forest algorithm used bagging or bootstrap aggregation to generate the 'tree.' Bagging is a term that refers to the grouping of machine learning approaches to improve their accuracy. The

(random forest) algorithm calculates the outcome based on tree-based predictions. It forecasts by averaging or combining the output of different trees. As the number of trees increases, the accuracy of the output improves [16]

Table 1: Arrangement of inquiries in the research project on depression.

S/N	Feature	Description
1	sex	Gender of the respondent
2	Age	The age of the respondent
3	Married	The respondent's Marital Status
4	Number_children	Number of kids of the Participant's
5	education_level	Educational attainment
6	total_members	Total family members of the Respondent
7	gained_asset	Acquired assets of the Respondent
8	durable_asset	Sustainable resource of the Respondent
9	save_asset	Saving assets of the Respondent
10	living_expenses	Living costs of the Respondent
11	other_expenses	Other expenditure of the Respondent
12	incoming_salary	Incoming income of the Participant's
13	incoming_own_farm	The Participant's own incoming farm
14	incoming_business	The Participant's incoming company
15	incoming_no_business	Incoming intensive agricultural fees of the Respondent
16.	incoming_agricultural	Income comes from agricultural sector
17	farm_expenses	Total expenditure in perspective of farm
18	labor_primary	Status of the labor needed in primary stages
19	lasting_investment	Total amount of lasting investment
20	no_lasting_investmen	Total amount of no lasting investment
21	target	[Zero: No depressed] or [One: depressed] (Binary for target class)

LOGISTIC REGRESSION (LR)

When the response variable is categorical, LR is the best regression strategy to use (binary). The approach for determining the risk factor is usually based on the assumption that the independent variables are normally distributed with equal variances. However, in most real-world circumstances, some of the variables are qualitative or measured on nominal or ordinal scales, which violates the normalcy assumption. The logistic regression model is then used, which does not require any distributional assumptions[17]

.

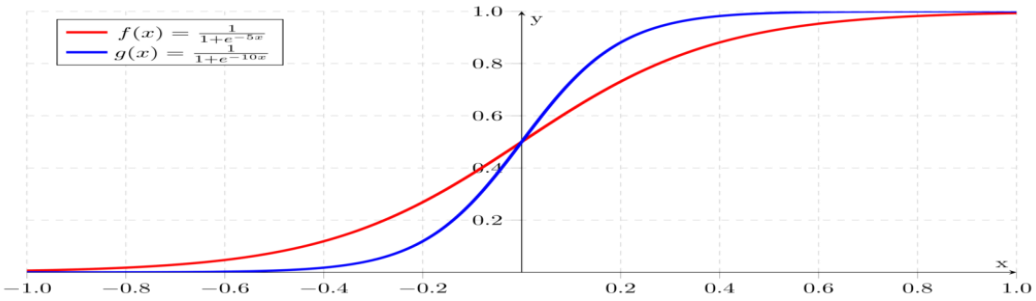

Figure-1: Graphical view of Logistic function [18].

NAÏVE BAYES (NB)

For very large volumes of data, the Naive Bayes (NB) classifier framework is simple to build. It's a mathematical model based on the Bayes' rule and premised on separate determinants. Of basic terms, an NB learning approach based on a certain characteristic in a class has no bearing on any other functionality. The calculation of class conditional density[19].is a major flaw in the naive Bayes technique. Depending on the data points, the conditional class density is usually determined. As a result, for unknown classification issues, we may be able to determine the conditional class density from unknown data objects designated by probability distributions.

Evaluation Criteria

Confusion Matrix:

The efficiency of a classification model is evaluated using a $n \times n$ matrix, where n is the number of core points. The matrices produce the overall output scores based on the classifier's predictions. It provides a clear picture of where our classification method is working and what inaccuracies it generates. We would use 2 × 2 matrices with four values for a binary classifier, as seen **Table 2**

Table 2: pattern of confusion matrix

		Actual Value	
Predicted Value		TP	FP
		FN	TN

The true positive value, true negative value, false positive value, and false negative value are used to calculate the accuracy, precision, recall, specificity, and F1-score[20].

True Positive (Tp)

The real observation suggests that depression exists, and the machine learning algorithm diagnoses depression from the given data (i.e., the detection result is true positive).

True Negative (Tn)

The real observation suggests that depression exists, however the ML system is unable to detect depression based on the data provided (i.e., the detection result is a true negative).

False Positive (Fp)

The real observation suggests that no depression exists, and the ML algorithm indicates that no depression is recognized from the given data.

FALSE NEGATIVE(FN): the real observation reveals that there is no depression, however the ML system detects depression from the available data (i.e., the detection result is a false negative).

Accuracy

Accuracy, precision, recall, and f1-measure are four critical variables for evaluating categorization outcomes. Accuracy is one of the most important categorization grading criteria[19]. which is expressed as the following:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FN} + \text{TP} + \text{FP}}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

What percentage of all positive instances is meant to be positive? The denominator is the model prediction that renders as positive from the entire dataset. Consider it a test to see how accurate the model is when it claims to be correct [19].

Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Good instances are a percentage of overall positive instances; throughout this case, the number of positive occurrences in the data is the denominator. Assume we are attempting to determine then how many accurate ones the system skipped while they were available. As a result, the formula for measuring recall/sensitivity is as follows [19].

F1-SCORE

$$F1 - Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

It is the amount of recall and precision in a harmonic form. This is where both factors come into play; the higher the F1 score, the better. The technique achieves well in the F1 ranking if the expected positive is actual positive (precision) and the model does not lose out on positives when predicting negatives (recall). The F1 score, which is a number between 0 and 1, represents the mean value. The F1 ranking, which is as follows, can be used to assess accuracy[21].

Support
Support is the number of correct representations of the entity in the provided dataset. Uneven training data support could suggest serious problems with the classifier's claimed scores, necessitating stratified sampling or reorganization[22].

Proposed Architecture
The proposed scheme depicts in Figure-2 as a pictorial display.

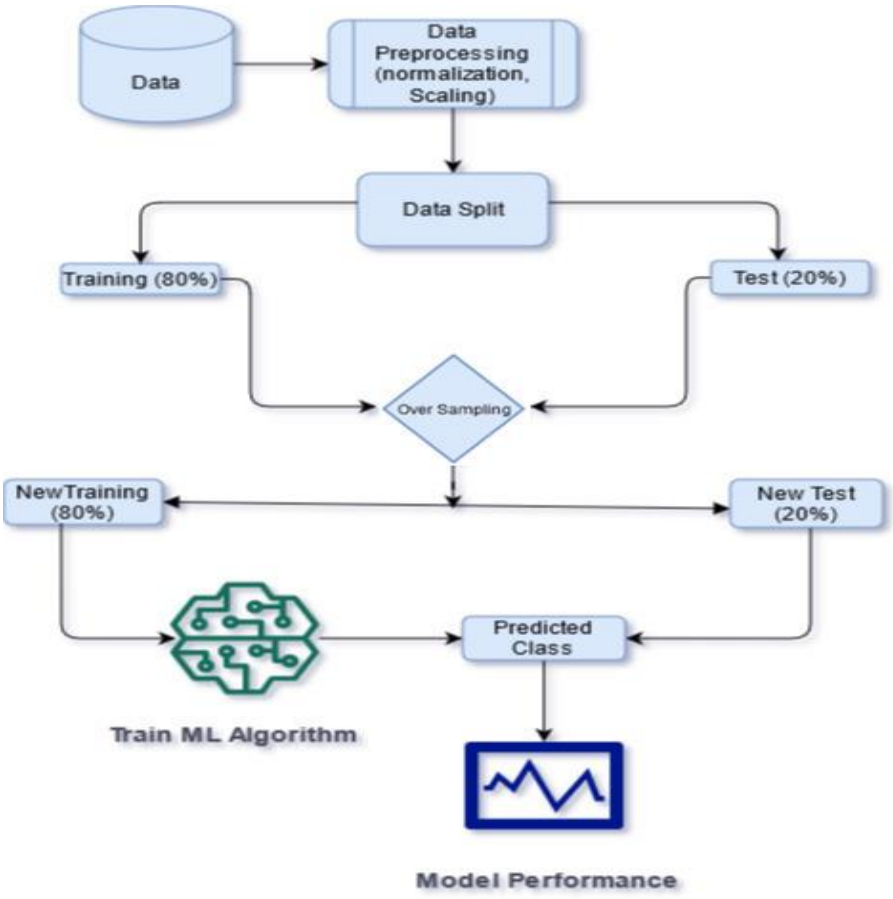

Figure-2: Proposed Model Architecture

RESULTS

Feature Importance Plot
In **Figure-2**, the feature "Age", "lasting_investment",and "no_lasting_investment" were the dataset's top three essential features. Where "incoming_salary," "sex," and "incoming_business" are the less critical features in our dataset.
Now we will examine four assessment methods for our dataset. **Table 3** and **Figure 4** show that with an imbalanced dataset, the random forest provided the highest accuracy 83%, where naïve Bayes provides the lowest accuracy 81%. It means that for the imbalance dataset, our three-machine learning algorithm provides the most insufficient accuracy. If we also observe that other criterion, in all of the cases highest for random forest algorithm.
Figure 5 corresponds to the confusion matrix for NB, LR, and RF algorithms' unbalanced data. These uncertainty matrices' false-negative and true negative values are higher than the true positive and false-positive values, as seen in **Figure 5.**

To deal with the unbalanced data in this dataset, we used oversampling considering K-fold(10-fold) validation techniques. Until then, after enforcing oversampling, there were substantial differences in machine learning model efficiency. **Table-4** and **Figure-6** provide a detailed breakdown of the overview. We connected our three models to existing methods and using oversampling for imbalanced data resulted in significantly better results—our proposed method basis on a random forest model. Our designed methodology with a random forest model surpassed most current models with the maximum accuracy of 89 percent and precision of 90 percent. Here also naïve Bayes provides the lowest accuracy and precision. If we look at other measurement criteria, random forests meet all measurement criteria and outperform.

Figure 7 displays the confusion matrix for the NB, LR, and RF algorithms after data balancing and

normalization, correspondingly. These graphs illustrate that these confusion metric's true positive and true negative value increased compared to confusion matrix **Figure 5** of before normalization. It indicates that the overall performance increased after balancing with normalization. In addition, when compared to the unbalanced data set, the false positive and false negative values in most algorithms dropped. As a result, the value of accuracy, precision, recall, and the f-1-score is frequently increased once we reconcile our data.

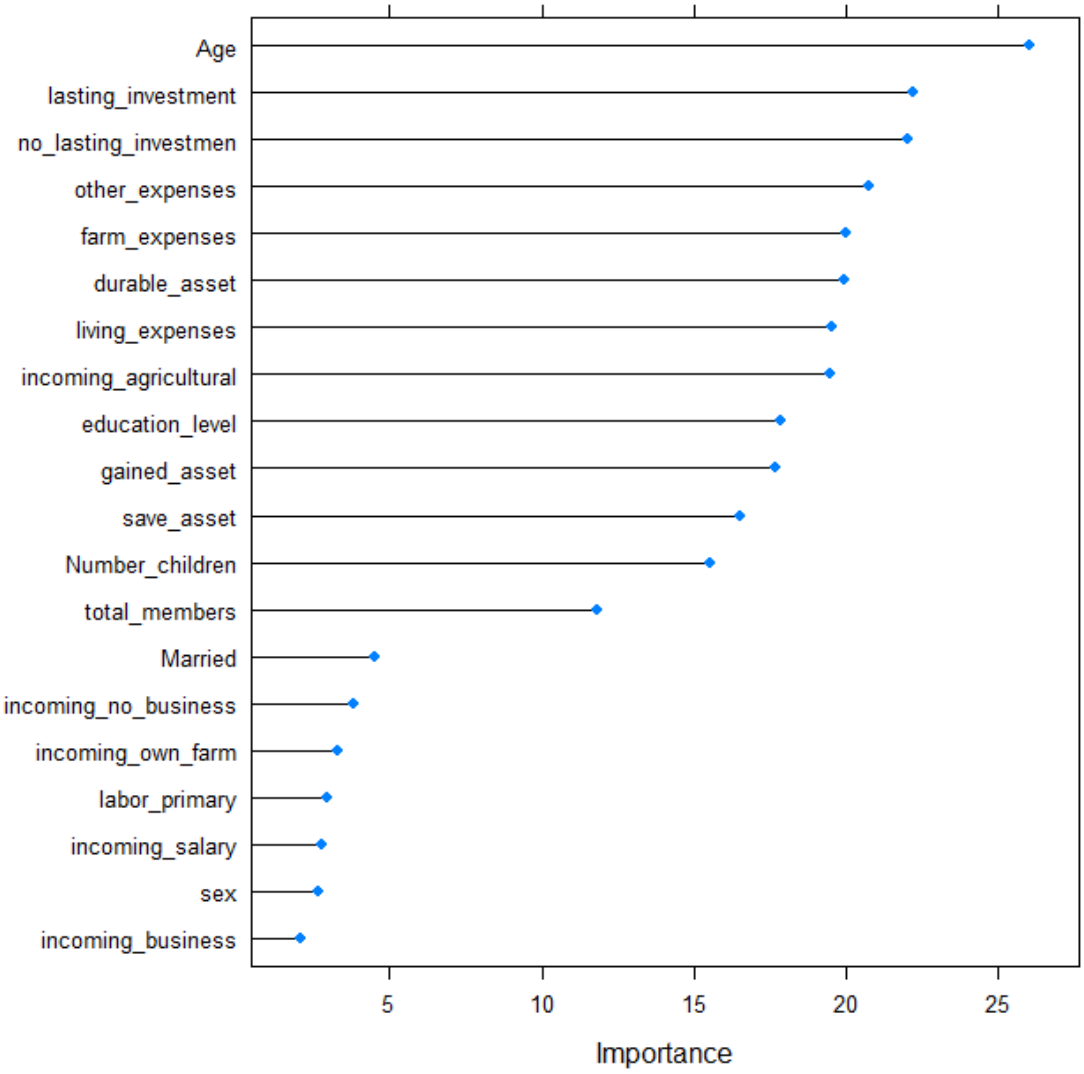

Figure 3: Feature importance Score

Table 3: Percentage of classification results with imbalance data

Methods	Accuracy	Precision	Recall	F1
Rf	0.83	0.9	0.3	0.45
Lr	0.82	0.8	0.26	0.39
Nb	0.81	0.57	0.19	0.28

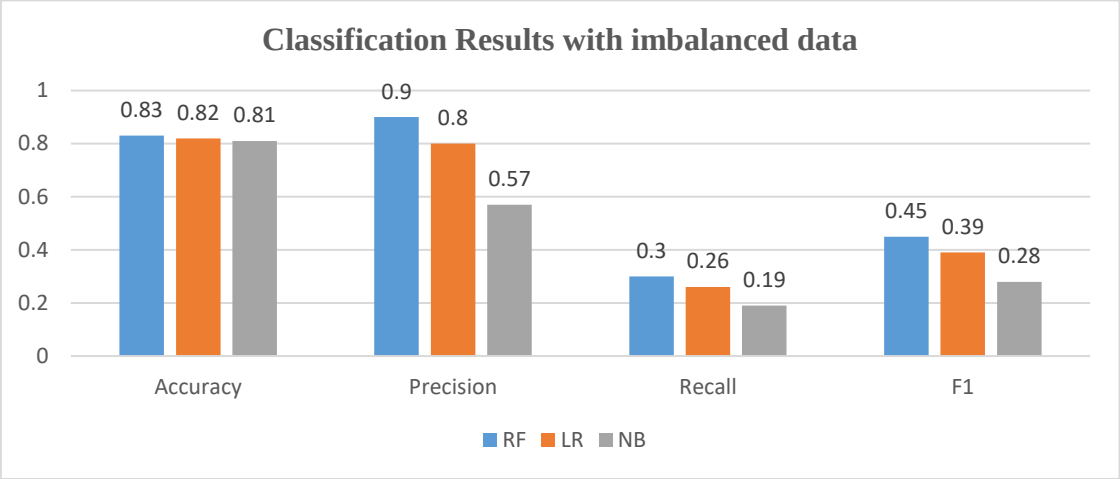

Figure-4: Percentage of classifications containing data on imbalance

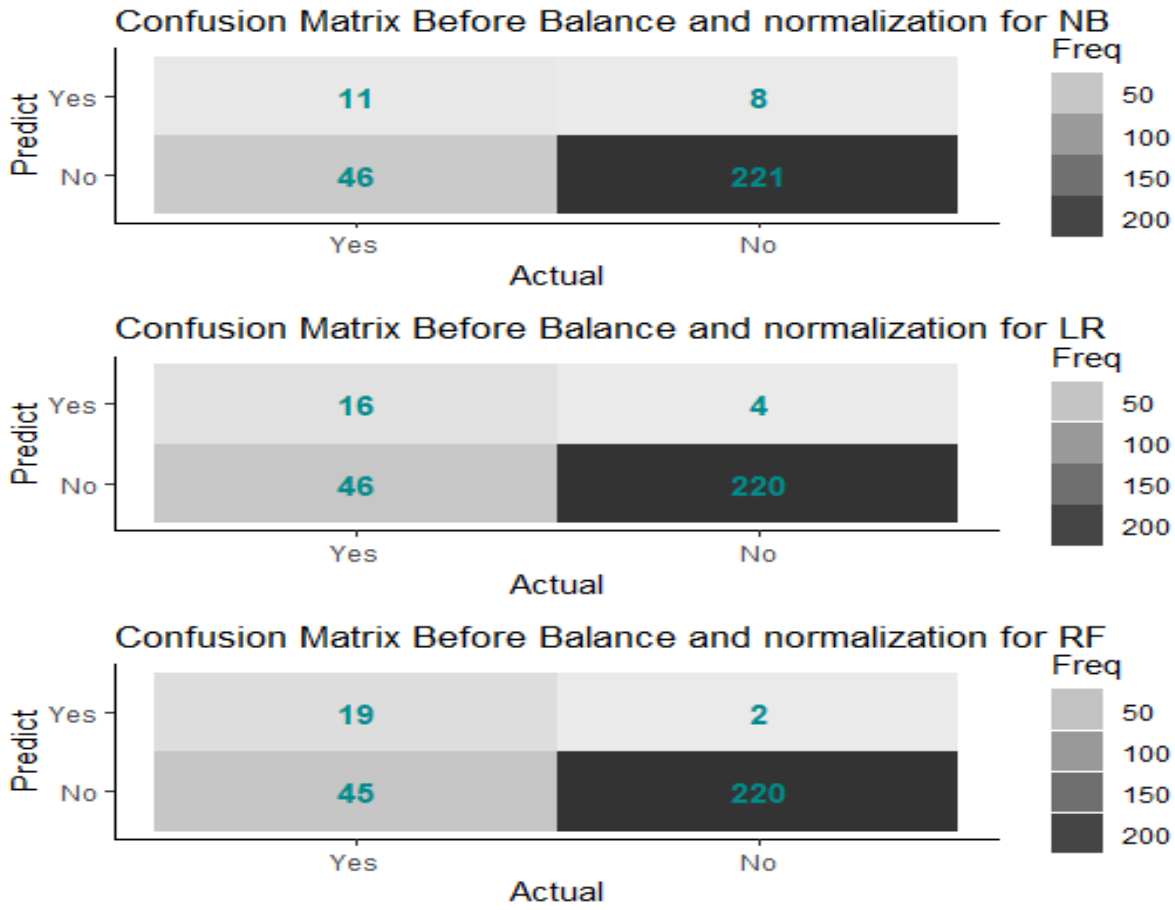

Figure-5: Confusion matrix for NB, LR, and RF algorithms before data balancing and standardization

Table-4: Percentage of classification results with balance data

Methods	Accuracy	Precision	Recall	F1
Rf	0.89	0.9	0.48	0.63
Lr	0.87	0.86	0.42	0.56
Nb	0.86	0.8	0.36	0.5

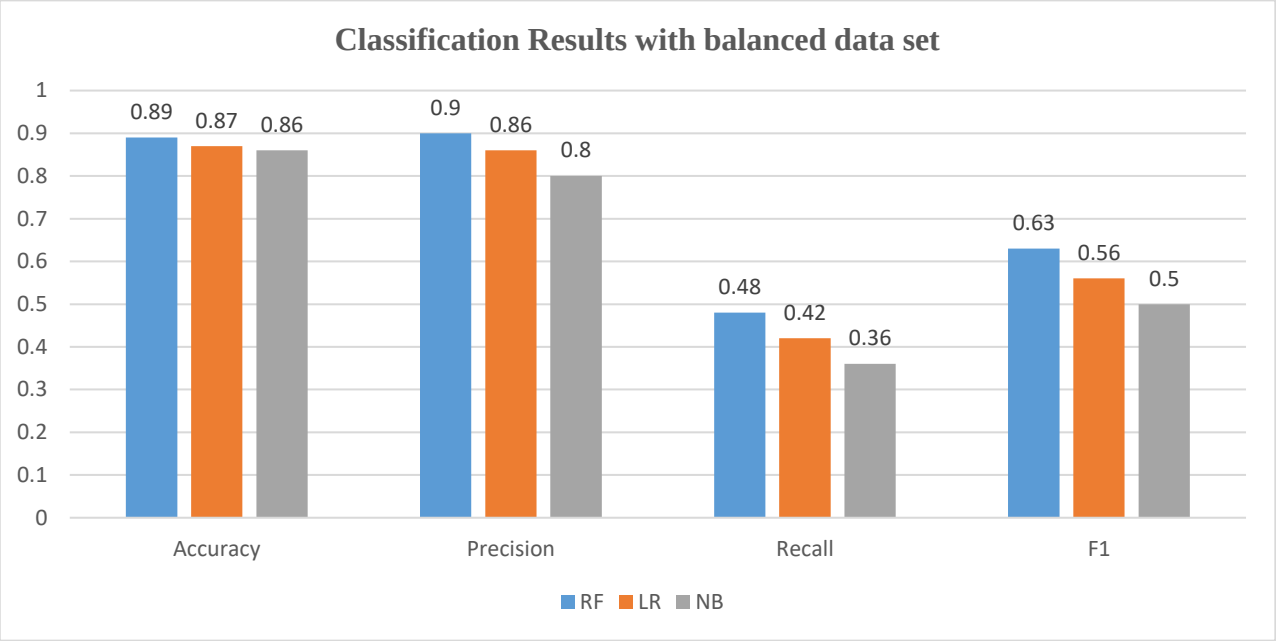

Figure-6: The proportion of categorization outcomes with balanced data (Using oversampling)

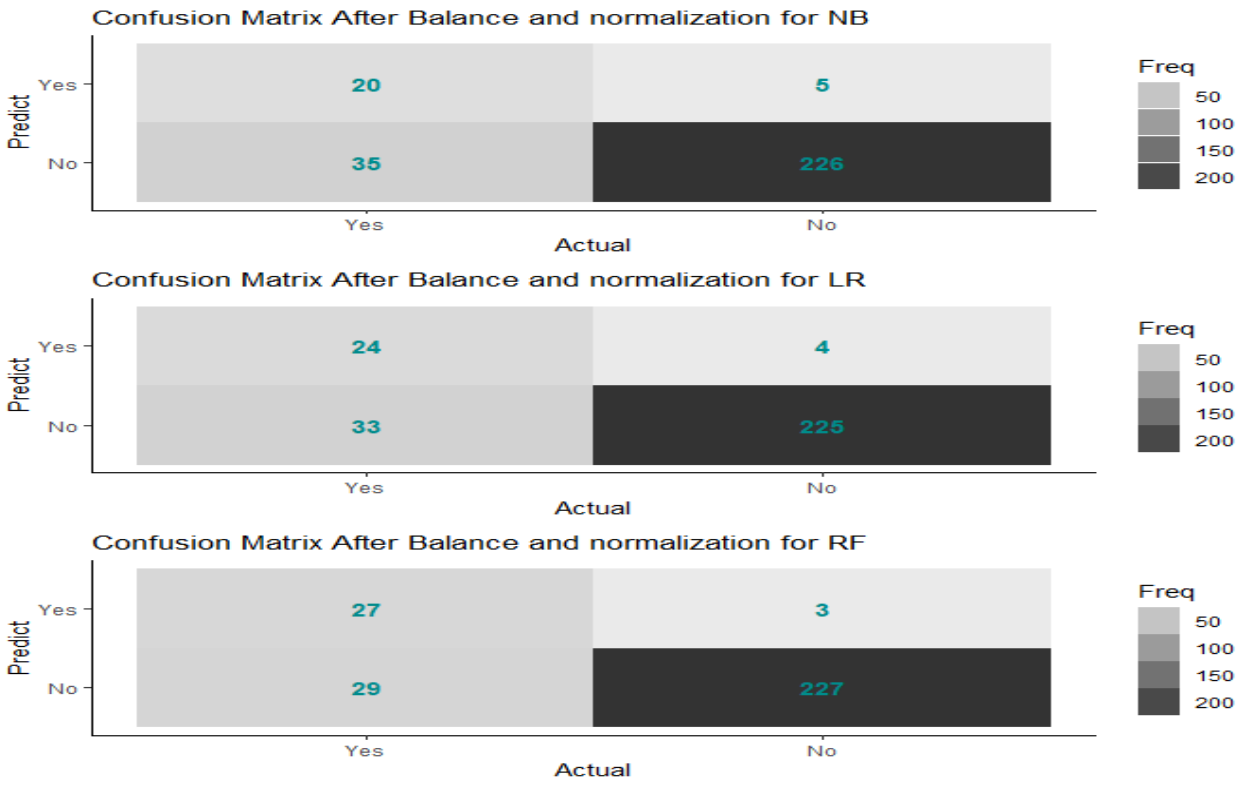

Figure 7: Confusion matrix for the NB, LR, and RF algorithms after data balancing with normalization

DISCUSSION

The features "Age," "lasting investment," and "no lasting investment" were the dataset's top three crucial features, according to our feature importance score in **Figure 3**. Whereas in our dataset, "incoming salary," "sex," and "incoming business" are the least important factors. The overview is decomposed in depth in **Table 4** and **Figure 6**. Our proposed strategy, which is based on a random forest model, produced noticeably superior results when we used oversampling for data that were imbalanced when we coupled our three models to current methods. Our developed methodology, which used a random forest model, outperformed most of the models at the time with a maximum accuracy and precision of 89 and 90 percent, respectively. Naive Bayes offers the lowest accuracy and precision in this situation as well. Other measurement criteria show that random forests outperform and satisfy all of them.

Following data balance and normalization, **Figure 7** shows the confusion matrix for the NB, LR, and RF algorithms, respectively. Various graphs show that, in comparison to the confusion matrix **Figure 5** of before normalization, the real positive and true negative values of these confusion metrics rose. It shows that after balancing and normalization, overall performance improved. Additionally, the false positive and false negative values in most algorithms decreased as compared to the unbalanced data set. As a result, once we reconcile our data, the value of accuracy, precision, recall, and the f-1-score is typically raised. Comparing overall machine learning tools for the accuracy, precision and recall where random forest shows highest values.

CONCLUSION

The major goal of this study was to use the K-fold validation (10-fold) technique to evaluate the performance of three distinct machine learning categorization models. We used the Over-sampling technique to improve model efficiency within that study, we also noticed a considerable improvement in a variety of structural performance measures. Rather than machine learning classifiers, random forest classifiers can usually do well. Random forest classifier achieved maximum accuracy of 89 percent and precision of 90 percent after using oversampling, while nave Bayes achieved the lowest accuracy and precision for balanced dataset. Oversampling techniques and machine learning technologies can be coupled to increase the efficiency of potential study findings. The highest accuracy and precision for imbalanced data in random forests are 83 and 90 percent, respectively, according to experimental results. For imbalance data, Naive Bayes has an accuracy of 81 percent, whereas balance data has an accuracy of 86 percent. Most of the time, accuracy and precision have increased because of using oversampling to balance data. In short, the suggested oversampling strategies with a balanced data set improve our model's overall performance.

ACKNOWLEDGEMENTS

This research was supported in part by the Dept. of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, 8100, Bangladesh, and in part by Dept. of Modeling and Data science, University of Turin, Italy.

Conflicts of Interest
The authors declare no conflict of interest.

REFERENCES

1. Sau, A., Bhakta, I. (2017)"Predicting anxiety and depression in elderly patients using machine learning technology. "*Healthcare Technology Letters* 4 (6)**:** 238-43.

2. J. Choi, J. Choi, and H. T. Jung, “Applying Machine-Learning Techniques to Build Self-reported Depression Prediction Models,” *CIN - Comput. Informatics Nurs.*, vol. 36, no. 7, pp. 317–321, 2018, doi: 10.1097/CIN.0000000000000463.

3. K. Kipli, A. Z. Kouzani, and I. R. A. Hamid, “Investigating Machine Learning Techniques for Detection of Depression Using Structural MRI Volumetric Features,” *Int. J. Biosci. Biochem. Bioinforma.*, vol. 3, no. 5, pp. 444–448, 2013, doi: 10.7763/ijbbb.2013.v3.252.

4. [J. F. Dipnall *et al.*, “Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression,” *PLoS One*, vol. 11, no. 2, pp. 1–23, 2016, doi: 10.1371/journal.pone.0148195.

5. F. Hasanzadeh, M. Mohebbi, and R. Rostami, “Prediction of rTMS treatment response in major depressive disorder using machine learning techniques and nonlinear features of EEG signal,” *J. Affect. Disord.*, vol. 256, no. May, pp. 132–142, 2019, doi: 10.1016/j.jad.2019.05.070.

6. S. Jiménez-Serrano, S. Tortajada, and J. M. García-Gómez, “A mobile health application to predict postpartum depression based on machine learning,” *Telemed. e-Health*, vol. 21, no. 7, pp. 567–574, 2015, doi: 10.1089/tmj.2014.0113.

7. A. Priya, S. Garg, and N. P. Tigga, “Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 1258–1267, 2020, doi: 10.1016/j.procs.2020.03.442.

8. Y. Zhang, S. Wang, A. Hermann, R. Joly, and J. Pathak, “Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women,” *J. Affect. Disord.*, vol. 279, no. September 2020, pp. 1–8, 2021, doi: 10.1016/j.jad.2020.09.113.

9. Z. Xie, O. Nikolayeva, J. Luo, and D. Li, “Building risk prediction models for type 2 diabetes using machine learning techniques,” *Prev. Chronic Dis.*, vol. 16, no. 9, pp. 1–9, 2019, doi: 10.5888/pcd16.190109.

10. M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," *J. Med. Syst.*, vol. 42, no. 5, 2018, doi: 10.1007/s10916-018-0934-5.

11. Iliou *et al.*, "ILIOU machine learning preprocessing method for depression type prediction," *Evol. Syst.*, vol. 10, no. 1, pp. 29–39, 2019, doi: 10.1007/s12530-017-9205-9.

12. S.Kumar and I. Chong, "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states," *Int. J. Environ. Res. Public Health*, vol. 15, no. 12, 2018, doi: 10.3390/ijerph15122907.

13. B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Comput. Methods Programs Biomed.*, vol. 109, no. 3, pp. 339–345, 2013, doi: 10.1016/j.cmpb.2012.10.008.

14. M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–12, 2018, doi: 10.1007/s13755-018-0046-0.

15. R. M. Khalil and A. Al-Jumaily, "Machine learning based prediction of depression among type 2 diabetic patients," *Proc. 2017 12th Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2017*, vol. 2018-Janua, pp. 1–5, 2017, doi: 10.1109/ISKE.2017.8258766.

16. Khosrowabadi, Reza, et al. "A Brain-Computer Interface for classifying EEG correlates of chronic mental stress." *The 2011 international joint conference on neural networks*. IEEE, 2011.

17. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression* (p. 536). New York: Springer-Verlag.

18. https://towardsai.net/p/machine-learning/logistic-regression-with-mathematics.

19. Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." *Proceedings of the 23rd international conference on Machine learning*. 2006.

20. Beauxis-Aussalet, Emma, and Lynda Hardman. "Simplifying the visualization of confusion matrix." *26th Benelux Conference on Artificial Intelligence (BNAIC)*. 2014.

21. Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC genomics* 21.1 (2020): 1-1

22. Gunn, Steve R. "Support vector machines for classification and regression." *ISIS technical report* 14.1 (1998): 5-16.