ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

"Evaluating the Efficacy of Various Machine Learning Models for Diabetes Prediction"

[1] Yarasu Madhavilatha, [2] Dr.J.Vanitha Vani, [3] Sitaram.CH

Abstract: Numerous machine learning algorithms find application across various domains, including disease prediction and diagnosis, recommendation systems, stock price forecasting, and object recognition. This study assesses the performance of three machine learning models, specifically Gaussian Naïve Bayes (GNB), Linear Support Vector Machine (LSVM), and Random Forest (RF), utilizing predictive performance accuracy, Area under Curve (AUC) score, and precision as evaluation metrics. The experimental results revealed that Linear Support Vector Machine (LSVM) excelled in diabetes prediction with an accuracy of 78.39%. Gaussian Naïve Bayes outperformed Random Forest, achieving an accuracy of 74.15%, while Random Forest (RF) exhibited the lowest performance, yielding an average accuracy of 72.72% for diabetes prediction.

Index Terms: Diabetes, forecasting diabetes, Linear Support Vector Machine (LSVM), Gaussian Naïve Bayes, Random Forest, machine learning, efficacy of machine learning techniques

1 INTRODUCTION

Diabetes is a condition characterized by abnormal fluctuations in glucose levels within the body [1]. This occurs due to an insufficient production of insulin, the hormone responsible for regulating blood sugar levels. The absence of adequate insulin production results in the body's inability to maintain proper sugar levels. Diabetes is a leading cause of mortality worldwide, underscoring the critical importance of timely symptom prediction for reducing its associated death toll.

Presently, extensive diabetes-related data is accessible from various repositories like Kaggle, MNIST, and UCI. This wealth of data serves as a valuable resource for the development of machine learning models aimed at predicting diabetes. These learning models are employed as analytical tools to draw meaningful insights from this extensive dataset.

The objective of this study is twofold: first, to propose a machine learning model that can assist physicians in their decision-making processes using machine learning algorithms, and second, to assess the performance of these models in predicting diabetes using the Kaggle diabetes data repository. In the diagnosis of diabetes, physicians rely on a range of features to identify the condition. These features include age, body mass index, blood pressure, insulin levels, and plasma glucose concentration (PGC). Physicians employ these features to categorize the disease's likelihood, subsequently conducting laboratory tests for confirmation.As the number of diabetes cases continues to rise, healthcare professionals may find themselves overwhelmed by treatment demands. Machine learning models can provide an innovative solution to the classification problem, aiding in the early identification of diabetes symptoms. Machine learning, especially in the realms of disease diagnosis and object recognition, has gained significant attention for its potential to assist physicians in the identification of various medical conditions [2].Given the significance of early diabetes diagnosis, various machine learning models, namely Linear Support Vector Machine (LSVM), Gaussian Naive Bayes, and Random Forest, can be employed to address the classification challenge associated with diabetes. These models facilitate the early detection of diabetes, enabling timely intervention to mitigate diabetes-related complications. Researchers have proposed multiple machine learning models for disease classification, yet the quest for a flawless model remains elusive. This is attributed to the differing accuracies and training complexities of these models. The Linear Support Vector Machine (LSVM), Gaussian Naïve Bayes, and Random Forest models are favored in disease classification due to their simplicity in both training and disease categorization.

[1] Associate Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions,
[2] Associate Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions,
[3] Assistant Professor, Department of CSE, Rise Krishna Sai Gandhi Group of Institutions

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

This paper seeks to address the following research inquiries:

1. Can we develop a machine learning model using LSVM, Gaussian Naive Bayes, and Random Forest classification algorithms that yields an acceptable level of accuracy in predicting diabetes?
2. What is the performance of LSVM, Gaussian Naive Bayes (GNB), and Random Forest (RF) classification algorithms in diabetes prediction?
3. How can the accuracy of diabetes prediction be enhanced by optimizing LSVM, Gaussian Naive Bayes, and Random Forest algorithms?

2 RELATED WORK

This section delves into existing research on diabetes diagnosis employing machine learning models. Despite numerous investigations in this area, the problem remains. Linear Support Vector Machine (LSVM) is a popular choice for classification tasks [3], largely due to its simplicity in prediction. In a study outlined in [4], various machine learning models were proposed for diabetes prediction. The authors conducted a comparative analysis involving different models, including decision tree classification, K-Nearest Neighbor (KNN), LSVM, and Naive Bayes. The evaluation metrics used to assess the performance of these classification models in predicting diabetes encompassed accuracy, recall, and precision.

The research findings demonstrated that LSVM outperformed other models when classifying the diabetes dataset collected from a medical center in Bangladesh. Based on this study, we have made our decision.To employ the LSVM classifier in constructing a machine learning system for diabetes diagnosis, another study comparing the performance of various machine learning models [5], including Random Forest and Support Vector Machine (SVM), revealed that SVM exhibited superior accuracy in classifying diabetes when contrasted with Random Forest, particularly on the PIMA data repository. Additionally, in [6], the RB-Bayes algorithm was utilized for diabetes prediction. The PIMA Indian dataset was employed for training and testing the algorithm, achieving a 72.9% accuracy rate. SVM yielded a 70.90% accuracy, Naïve Bayes registered 67.71%, and the decision tree achieved 68.18%.

Another examination focusing on the performance assessment of classification algorithms in diabetes prediction [7] showed that the accuracy of Support Vector Machine was 67.79%, whereas K-Nearest Neighbor reached 74.89%. This analysis involved dividing the PIMA Indian data repository into training (70%) and testing (30%) sets, using 538 samples for training and 230 for testing out of a total of 768 samples.

In recent years, LSVM has gained substantial prominence in the field of diabetes disease classification [8-9] due to its remarkable performance [8]. One key characteristic of LSVM is its capability to handle non-linear classification with superior accuracy. Consequently, LSVM was the model chosen for this study. LSVM, a supervised learning algorithm, plays a vital role in disease diagnosis, encompassing prediction and regression tasks [10]. In one instance, authors harnessed the UCI data repository for diabetes disease classification, achieving an average accuracy of 75.5% using the LSVM model. LSVM has also demonstrated effectiveness in multiclass and multidimensional data classification [10], with precision and accuracy serving as pivotal metrics, among others, for gauging its efficiency.

A comparative analysis [11] of LSVM, Decision Tree, and Naive Bayes classification models indicated that LSVM excels in diabetes dataset classification. Existing literature [12-13] contains several comparative studies exploring machine learning models, such as LSVM, Naïve Bayes, and decision tree classification, primarily emphasizing accuracy as the evaluation metric. However, in this particular study, three models, LSVM, GNB, and RF, were employed, and a range of metrics, including confusion matrix, recall-precision analysis, and AUC score, were used to evaluate the models' performance in diabetes disease classification.

Machine learning algorithms play a pivotal role in automating diabetes prediction through various learning models. For instance, in [14], a random forest algorithm was leveraged for diabetes prediction using the UCI diabetes data repository. Early diabetes prediction is vital for enhancing survival rates and delivering timely treatment to diabetes patients [15]. Authors devised an early prediction model for diabetes by employing artificial neural network (ANN), RF, and K-means clustering algorithms. The analysis of the study demonstrated that these algorithms performed with varying levels of accuracy, with the best accuracy reaching 74.7%.

In [16], a diabetes risk prediction model was introduced through the utilization of the random forest algorithm. Authors scrutinized the predictive performance of this proposed random forest-based diabetes prediction model.prediction model and the result shows accuracy variation based on the data scaling. Another study [17], applied Gaussian naïve Bayes and proposed a machine learning model for diabetes prediction. The predictive accuracy of the proposed Gaussian naïve based diabetes prediction model is 73.33%. In machine learning, this can be an acceptable level of accuracy but, still it can be improved by increasing the training set and by applying feature selection to get more accurate result. In [18], K-Means clustering which is unsupervised algorithm is applied to propose a diabetes prediction model. The aim of the study is to explore the factors resulting in diabetes.

3 RESEARCH METHOD

In this research, the kaggle data repository is used to build a machine learning model using the Linear Support Vector Machine (LSVM), Gaussian Naïve Bayes (GNB)

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 9, Issuse.4 Dce 2018

and Random Forest (RF) algorithms. In the kaggle repository, there are 768 diabetes and non-diabetes features. The 75% of the data repository is used for training and 25% is used in testing the algorithms. Nearly, 576 samples of diabetic and non-diabetic data is used in the training and 192 samples are used in testing. The python programming language is used to build the machine learning models. This language has scientific learning kit, under which the Linear Support Vector Machine (LSVM), GNB and RF library is integrated.

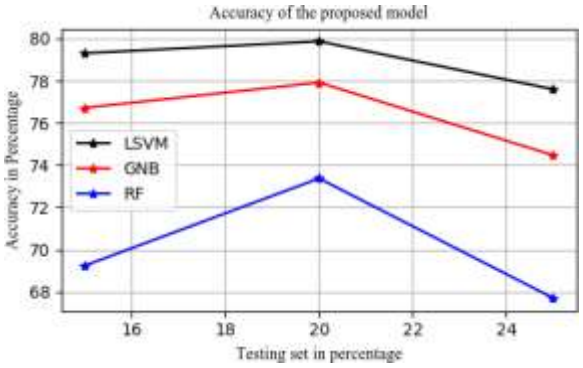

Fig.1. *diabetes prediction using machine learning*

3.1. FEATURE EXTRACTION

Feature extraction is an important step in developing machine learning models in that extracting features optimizes the accuracy of the predictive models and reduces the computational time. In this research, we have employed Pearson's correlation analysis to extract the features having strong correlation with the diabetes dataset class label feature. As shown in figure 2, the class label is strongly correlated to PGC, which implies that the use of this feature in the training is vital to the predictive accuracy of the proposed model.

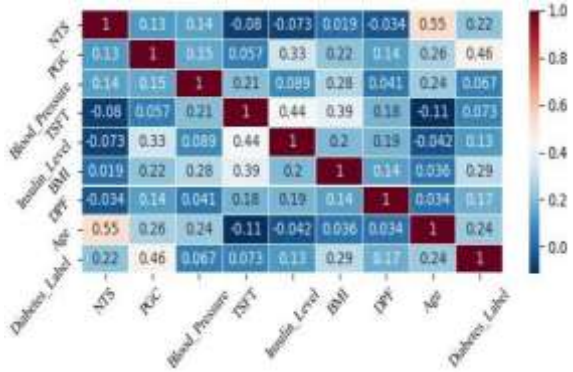

Fig.1*. Pearson's correlation of diabetes features*

4 RESULTS AND DISCUSSIONS

In this section, the results of the research are explained. The performance of LSVM, Gaussian Naïve Bayes and Random Forest models in classification of diabetes is evaluated and analyzed using the performance metrics such as Accuracy, AUC curve and precision in the classification. And the analysis results are discussed in sections 4.1, 4.2 and 4.3 respectively.

4.1. ACCURACY OF THE MODELS

The accuracy is an important metric in evaluation of any machine learning model. To evaluate the performance of the LSVM, Gaussian Naïve Bayes and Random Forest learning models, the accuracy of the models is tested once the models are built and trained on the kaggle dataset. As shown in figure 3, the test size has effect on the accuracy of the proposed model, all the proposed models performed well when 20% of the dataset is used in testing and 80 % of the dataset is used in training. The graph also shows that, LSVM has better performance as compared with the other algorithms on diabetes prediction.

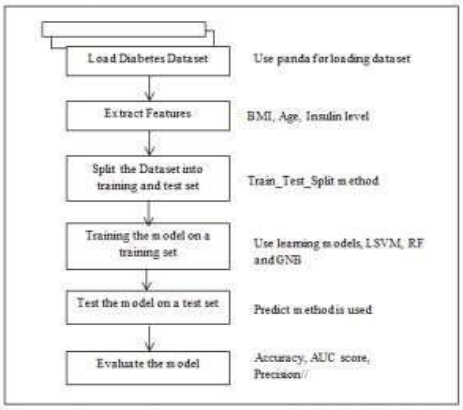

Fig.3*. Accuracy of the models*

As shown in figure 3, the average accuracy for the Linear Support Vector Machine (LSVM) is better than the Naïve Bayes and Random Forest models on the diabetes classification. The LSVM has an average accuracy of 78.39%, the GNB has an average accuracy of 74.15% and the RF has the least accuracy 72.72% on the random tests conducted on the models in the classification of the diabetes.

4.2. RECEIVER OPERATING CHARACTERISTICS ANALYSIS

The receiver operating characteristic (ROC) is another important metric used in the evaluation of the performance of amachine learning model. The ROC curve of the LSVM, Gaussian Naïve Bayes and Random Forest learning models is shown in figure 3. The ROC is shows the false positive rate (FPR) against true positive rate (TPR) of the classifiers. The TRPR, true positive rate is the recall of the classifiers, whereas the FPR, false positive rate is the fraction of false positive out of all negatives, means the rate at which the classifiers predicted diabetes negatives as diabetes positive.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 9, Issue.4 Dce 2018

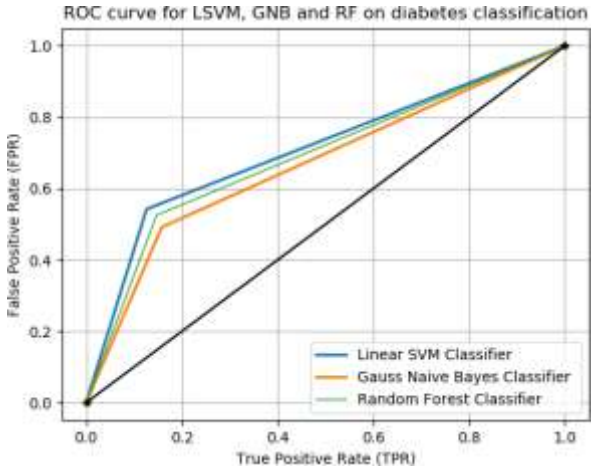

Fig.4. *ROC graph of LSVM, GNB and RF models*

4.3. PRECISION ANALYSIS

The precision is an important metric in the evaluation of the performances of machine learning models. The precision value indicates that how often the machine learning model is correct in predicting diabetic patient as diabetes positive. The precision of LSVM, GNB and RF models on the random tests on the debates classification is illustrated in table 1.

TABLE 1: *PERCSION OF THE MODELS*

Experimental test	*Accuracy of the models in %*		
on model	*GNB*	*LSVM*	*RF*
1	*70.19*	*69.79*	*67.5*
2	*68.49*	*69.09*	*60.44*
3	*67.34*	*66.66*	*62.35*
4	*79.25*	*69.13*	*66.58*
5	*61.33*	*71.55*	*65.29*

5 CONCLUSION

In this study, we have analyzed the predictive performance of three machine learning algorithms namely LSVM, GNB and RF on diabetes prediction using the kaggle diabetes data repository. Feature extraction was conducted by employing the Pearson's correlation to find the relationship between the class label and other features of the diabetes disease. The feature extraction helped us to identify the relevant features and improve the accuracy of the models by using only relevant features in the training. The experimental result on the predictive performance analysis of the three algorithms shows that, LSVM performed well on the prediction of diabetes with highest accuracy compared to GNB and RF.

6 REFERENCES

[1] Ahmed Hamza Osman, Hani Moetque Aljahdali.–Diabetes Disease Diagnosis Method based

n Feature Extraction using K-SVM,‖ International Journal of Advanced Computer Science and Applications, Vol. 8, No. 1, 2017.

[2] Tsehay Admassu Assegie, Pramod Sekharan Nair, Handwritten digits recognition with decision tree classification: a machine learning approach, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, October 2019, pp. 4446~4451.

[3] Vishakha Vinod Chaudhari, Prof. Pankaj Salunkhe, Diabetic Retinopathy Classification using SVM Classifier, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 6, Issue 7, July 2017.

[4] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus, International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, IEEE, 2019.

[5] Sinan Adnan, Diwan Alalwan, Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset, Indonesian Journal of Electrical Engineering and Computer Science Vol. 14, No. 1, April 2019, pp.88~95.

[6] Rajni Amandeep, RB-bayes algorithm for the prediction of diabetic in -PIMA Indian dataset‖, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 6, December 2019, pp. 4866~4872.

[7] Ratna Patil, Sharavari Tamane, A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes, International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 5, October 2018, pp. 3966~3975.

[8] Davar Giveki, Hamid Salimi, GholamReza Bahmanyar, Younes Khademian, Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search.

[9] Classification of Diabetes Disease Using Support Vector Machine, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

[10] S Amarappa, Dr. S V Sathyanarayana, Data

classification using Support vector Machine (SVM), a simplified approach, International Journal of Electronics and Computer Science Engineering.

[11] Chitra Arjun, Mr.Anto S, Diagnosis of Diabetes Using Support Vector Machine and Ensemble Learning Approach, International Journal of Engineering and Applied Sciences (IJEAS) ISSN: 2394-3661, Volume- 2, Issue-11, November 2015.

[12] Shital Tambade, Madan Somvanshi, Pranjali Chavan, Swati Shinde, SVM based Diabetic Classification and Hospital Recommendation, International Journal of Computer Applications (0975 – 8887) Volume 167 – No.1, June 2017.

[13] Ihsan Salman Jasim, Adil Deniz Duru, Khalid Shaker, Baraa M. Abed, Hadeel M. Saleh ,Evaluation and Measuring Classifiers of Diabetes Diseases, 978-1- 5386-1949-0/17, IEEE, 2017.K.Vijiya, Kumar, IEEE, Proceeding of International Conference on Systems Computation Automation and Networking 2019.

[14] Talha Mahboob Alama,, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb , Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc , Muhammad Awais Malikb, Muhammad Mehdi Razab , Salman Ibrarb, Zunish Abbas, A model for early prediction of diabetes, Informatics in Medicine Unlocked, 2019.

[15] Weifeng Xu, Jianxin Zhang, Qiang Zhang, Xiaopeng Wei , Risk prediction of type II diabetes based on random forest model, 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics, IEEE, 2017.

[16] Messan Komi, J un Li Y ongxin Zhai, Xianguo Zhang,, Application of Data Mining Methods in Diabetes Prediction, 2nd International Conference on Image, Vision and Computing, IEEE, 2017.

[17] Gagandeep Singh, Gurpreet Singh, Diabetes classification using K-Means, APEEJAY journal of computer science and application, 2019.