Analysis and Prediction of Diabetes Mellitus using Machine Learning: A Study on Diabetic Dataset

T. Muni Dharani

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract*—** *Diabetes mellitus is a chronic metabolic disorder affecting millions of people worldwide. The increasing prevalence of diabetes poses significant challenges to healthcare systems and requires effective early detection and management strategies. This research paper explores the application of machine learning techniques for analyzing and predicting diabetes based on a comprehensive diabetic dataset. The dataset consists of various clinical and demographic features of patients, making it an ideal resource for building predictive models. Through the study, we aim to identify key factors contributing to diabetes and develop accurate models for early diagnosis. The dataset used in this study is sourced from the UCI Machine Learning Repository. Two machine learning algorithms, namely), Multilayer Perceptron (MLP) and Naïve Bayes classifiers, are employed to analyze the dataset and determine the most effective performance and accuracy. Among these classifiers, the MLP algorithm demonstrates the highest performance with an accuracy of 85.50%.*

I. INTRODUCTION

Diabetes mellitus is a complex and progressive disease characterized by chronic hyperglycemia, resulting from the body's inability to produce or effectively utilize insulin. In this research, we leverage a comprehensive diabetic dataset to explore the use of machine learning algorithms for diabetes prediction and risk assessment. In recent literature, various AI algorithms have been employed for the detection of diabetic. Diabetes mellitus is the leading cause of blindness among a significant age group in Western countries and its prevalence is also increasing in developing nations. Individuals with diabetes are at a significantly higher risk of developing blindness compared to those without diabetes. Moderate diabetic and clinically significant macular edema can result in severe vision loss. Early detection through regular screening is crucial as it can be effectively treated in its initial stages. However, the cost and manual effort involved in screening are significant, making automated screening highly desirable. In diabetic, the blood vessels that nourish the retina start leaking fluid and blood, leading to characteristic visual features such as microaneurysms, hemorrhages, hard exudates, cotton wool spots, and vein occlusion [9].

II. SUPERVISED LEARNING

Supervised learning is a subfield of artificial intelligence (AI) and machine learning that involves training a model using labeled data. It is called "supervised" because the training data provides explicit supervision or guidance to the model in the form of input-output pairs [2][10].

In supervised learning, the goal is to learn a mapping function that can accurately predict the output or label for new, unseen input data. The training data consists of examples where both the input (features) and the desired output (labels) are known. The model learns from these examples to generalize and make predictions on new, unseen data.

III. METHODOLOGY

In this way, the paper proposed Multilayer Perceptron (MLP) and Naïve Bayes calculations for productively finding the arrangement errands of the diabetic information.

3.1 Multilayer Perceptron (MLP)

A MLP is a legend among the most generally saw Frontal cortex Affiliation plan that has been utilized for different applications. The MLP coordinate is overall created utilizing various fixations or supervising units, and it is sorted out into an improvement of something like two layers [7]. The focal layer (or the most decreased layer) is named as a data layer where it gets the outside data while the last layer (or the most overwhelming layer) is a yield layer where the reaction for the issue is gotten. The mystery layer is the all around enthralling layer in the information layer and the yield layer, and may outline with some spot just about one layers. The plan of MLP could be conveyed as a nonlinear improvement issue. The target of MLP learning is to track down the best loads that limit the parcel between the data and the yield. The most transcendent preparing assessment utilized in NN is Back causing (BP), and it has been utilized in supervising different issues in model validation and depiction. This assessment

relies on a couple of limits, for example, novel covered focus fixations at the concealed layers learning rate, energy rate, demand work and how much expecting to occur. Additionally, these limits could change the show on the getting from stunning to remarkable precision [8].

3.2 Naïve Bayes

Naive Bayes classification is a popular machine learning algorithm that is based on Bayes' theorem with an assumption of independence between the features. It is a simple yet effective probabilistic model used for classification tasks. The algorithm is called "naive" because it assumes that the presence or absence of a particular feature is unrelated to the presence or absence of other features. In other words, it assumes that all features are independent of each other, which is not always true in real-world scenarios. Despite this simplifying assumption, Naive Bayes often performs well in practice and can provide reliable results [7]. The Naive Bayes algorithm works by calculating the probabilities of a sample belonging to each possible class based on the observed feature values. It then assigns the sample to the class with the highest probability. The calculation of these probabilities involves estimating the likelihood of each feature given each class and the prior probability of each class. The algorithm is particularly useful when working with high-dimensional datasets and when the assumption of feature independence is reasonable. It is known for its computational efficiency and is often used in text classification, spam filtering, sentiment analysis, and other similar tasks[5][6].One key advantage of Naive Bayes is its ability to handle both numerical and categorical data. It can handle continuous features by assuming a specific distribution, such as Gaussian (for continuous variables) or multinomial (for discrete variables).

IV. EXPERIMENTAL RESULTS

The experiments were conducted using the Python programming language, utilizing the powerful Scikit-learn library for data representation, manipulation, and analysis. For this study, the Diabetes dataset from the University of California, Irvine (UCI) library of AI datasets was employed [11]. This dataset consists of 768 instances, each containing 8features, along with a binary target variable indicating the presence or absence of diabetic. It was curated by a team of researchers from the University of Debrecen, Hungary, who extracted features from the test images to predict the presence of diabetic. In this study, Two machine learning algorithms, namely Naïve Bayes and Multilayer Perceptron (MLP), were applied to the Diabetes dataset. The performance of each algorithm was evaluated using accuracy, precision, and recall as evaluation metrics. The experimental results are summarized in the table-1 and same shown in the figure-1:

Table-1
Classifier Performance

Algorithm	Accuracy	Precision	Recall
Naïve Bayes	83.47	83	83.0
Multilayer Perceptron	85.50	85.8	85.7

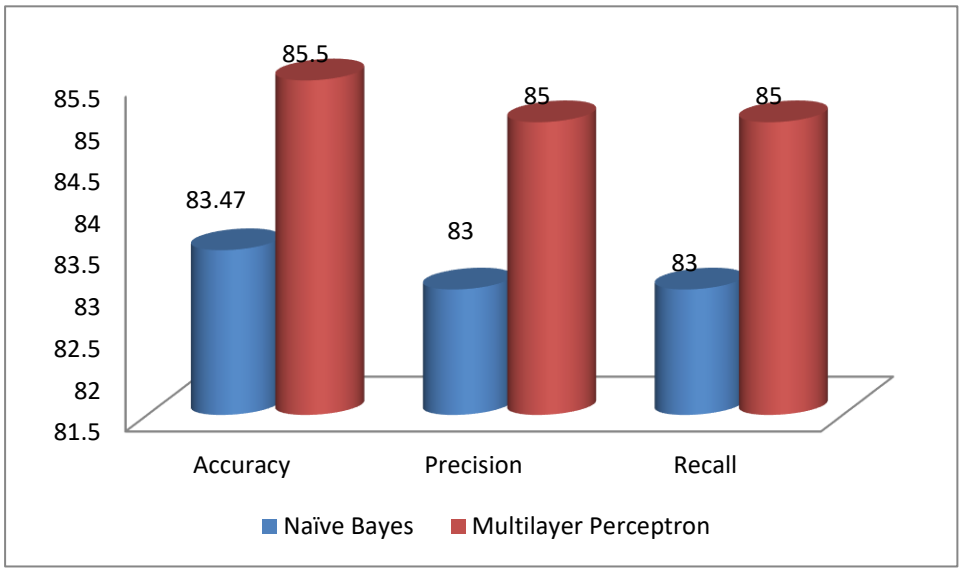

Figure-1: Classifier Results

We observe in the figure-1, Naïve Bayes demonstrated a respectable performance in predicting diabetic with an accuracy of 83.47%. The precision and recall values were also high, indicating the model's ability to correctly identify instances of diabetic. However, compared to the other algorithm, Naïve Bayes showed slightly lower accuracy.

The Multilayer Perceptron algorithm exhibited promising results in diabetic prediction. With an accuracy of 85.50%, it outperformed MLP. The precision and recall values were also high, indicating the model's ability to accurately identifying instances of diabetic. outperformed MLP in terms of accuracy and overall performance.

Overall, all two machine learning algorithms yielded promising results in predicting diabetic using the Diabetes Debrecen dataset. MLP emerged as the top-performing algorithm, followed by Naïve Bayes. These findings suggest that these algorithms have the potential to assist in the early detection and diagnosis of diabetic.

V. CONCLUSION

Based on the experimental results, it can be observed that all Two machine learning algorithms achieved high accuracy in predicting diabetic. The MLP algorithm showed the highest overall performance, with an accuracy of 85.50%. The Naïve Bayes algorithm also performed well, with an accuracy of 83.47%. Naïve Bayes demonstrated good performance, albeit slightly lower than the other algorithm, with an accuracy of 83.47%.

These results indicate that machine learning models have the potential to effectively predict diabetic using the Diabetes dataset. The high accuracy, precision, and recall values achieved by the algorithms highlight their ability to accurately classify instances of diabetic. The findings of this study contribute to the development of AI-based models for the early detection and diagnosis of diabetic, which can aid in timely interventions and prevent vision loss in patients with diabetes. Further research and refinement of these algorithms can potentially improve their performance and expand their applications in the field of diabetic prediction.

REFERENCES

[1] Agurto, Carla, et al. (2010) "Multiscale AM-FM methods for diabetic retinopathy lesion detection." IEEE transactions on medical imaging 29(2): 502-512.

[2] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[3] Ferris, F. L. (1993). How effective are treatments for diabetic retinopathy?. Jama, 269(10), 1290-1291

[4] Fong, D. S., Aiello, L., Gardner, T. W., King, G. L., Blankenship, G., Cavallerano, J. D., & Klein, R. (2004). Retinopathy in diabetes. Diabetes care, 27(suppl 1), s84-s87.

[5] G. Ravi Kumar, S. Rahamat Basha, Surya Bhupal Rao, "A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, ISSN:0973-8975.

[6] G. Ravi Kumar, K. Tirupathaiah and B. Krishna Reddy, "Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques", International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019

[7] J.Han and M.Kamber,"Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nded.San Mateo, CA; Morgan Kaufmann, 2006.

[8] N.Michael, "Artificial Intelligence – A Guide to Intelligent Systems", 2nd Edition, Addison Wesley 2005

[9] Ong, Gek L., et al. (2004) "Screening for sight-threatening diabetic retinopathy: comparison of fundus photography with automated color contrast threshold test." American journal of ophthalmology 137(3): 445-452.

[10] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[11] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/