IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction Using Different Machine Learning Techniques

Anam Khan (Assistant Professor)
Information Technology
Galgotias College of Engineering and Technology
Greater Noida, India

Rahul Kumar Sachdeva (Student)
Information Technology
Galgotias College of Engineering and Technology
Greater Noida, India

Rahul Lodhi (Student)
Information Technology
Galgotias College of Engineering and Technology
Greater Noida, India

Prakhar Kumar Singh (Student)
Information Technology
Galgotias College of Engineering and Technology
Greater Noida, India

***Abstract*—Diabetes mellitus, particularly type-2 diabetes, represents a substantial portion of global diabetes cases, exerting significant pressure on healthcare systems worldwide[1]. This metabolic disorder, marked by inadequate insulin production or response leading to heightened blood sugar levels, is linked with numerous health complications, including heart and kidney diseases. Conventional diagnosis involves frequent visits to diagnostic centers, consuming both time and financial resources. However, the advent of machine learning technologies offers a promising solution to this challenge. By leveraging advanced data processing techniques, machine learning models can predict the onset of diabetes, enabling early intervention and improved patient outcomes. This research aims to support physicians in the timely identification and effective diagnosis of type 2 diabetes. Supervised machine learning techniques were executed to "Pima dataset", utilizing six predictors to develop predictive models. The study employs classification algorithms such as SVM, KNN, Naive Bayes, Gradient Boosting Classifier, Logistic Regression, and Random Forest. Results indicate promising accuracy levels across the models, with Support Vector Machine achieving 76%, KNN 80%, Naive Bayes 76%, Gradient Boosting Classifier 85%, Logistic Regression 80%, and Random Forest 96%. These outcomes underscore the efficacy of machine learning approaches in diabetes prediction, offering a valuable tool for healthcare professionals to enhance diagnosis and patient care. This study advances the creation of accurate and effective type 2 diabetes diagnosis tools by utilizing machine learning's predictive capabilities. The findings highlight the potential of machine learning algorithms to analyze large volumes of diabetes-related data, enabling proactive healthcare interventions and ultimately improving patient outcomes. Moreover, the study underscores the importance of ongoing research and confirmation efforts to guarantee the dependability and effectiveness of machine learning in clinical settings.**

Keywords*— *Machine Learning, SVM, KNN, Naive Bayes,
Gradient Boosting Classifier, *Random Forest Algorithm.*

I. INTRODUCTION

Diabetes, a chronic metabolic non-communicable disease (NCD), poses a significant global health challenge, with an estimated 415 million cases worldwide, projected to rise to 642 million by 2040.[7] It is characterized by abnormally high blood glucose levels, primarily caused by insulin dysfunction. While the human body requires glucose for energy, inefficient insulin production or utilization leads to hyperglycemia, the hallmark of diabetes. Type 2 diabetes, the most prevalent form, often stems from a combination of unhealthy lifestyle habits and insufficient physical activity. Consequently, glucose remains in the bloodstream, contributing to various systemic complications affecting the kidneys, eyes, neurological system, and arteries. Hyperglycemia, a key feature of diabetes, can result from insulin deficiency, as observed in type 1 diabetes, where pancreatic beta cells fall short in producing adequate

insulin. Type 2 diabetes, on the other hand, involves insulin resistance, where the body cannot efficiently utilize the insulin it produces. It is essential to comprehend the complex nature of diabetes and the underlying mechanisms in order to create preventative and management plans that work. This abstract provides a concise overview of diabetes, highlighting its global prevalence, etiology, and the distinction between type 1 and type 2 diabetes.

1.1 TYPES OF DIABETES

[2] Type 1 diabetes occurs due to pancreas failure in producing an adequate amount of insulin. Formerly known as "insulindependent diabetes mellitus" (IDDM) or "juvenile diabetes," its cause remains unidentified. Typically diagnosed in individuals under the age of twenty, those with type 1 diabetes must manage the condition throughout their lives through insulin injections. Doctors often recommend regular exercise and a healthy lifestyle for effective management.

Type 2 diabetes originates from insulin resistance, where cells do not efficiently respond to insulin. Referred to as "noninsulin-dependent diabetes mellitus," this type is commonly associated with excessive weight. The prevalence of type 2 diabetes is expected to increase by 2025. Diabetes rates are 3% lower in rural areas compared to urban areas.

The coexistence of diabetes mellitus, obesity, and hypertension is observed, with research indicating that maintaining normal blood pressure contributes to overall health.

Gestational diabetes, classified as Type 3, occurs in a pregnant woman when she develops elevated blood sugar levels without a previous history of diabetes. Studies reveal that 19% of pregnant women experience gestational diabetes. There is an increased risk of developing gestational diabetes in older age during pregnancy.

II. LITERATURE REVIEW

Detecting diabetes early is essential for swiftly intervening and managing the condition to prevent unforeseen outcomes. This study explores the use of ML classification methods to develop models for early identification of diabetes development. Shafi et al. [9] emphasize the significance of accurate prediction frameworks in assessing diabetes risk. They utilize three ML algorithms which are Decision Trees (DT), SVM, and Naive Bayes Classifier (NBC)—to analyze the UCI repository's PID dataset. The experimental results reveal NBC's adequacy with 74% accuracy, followed by DT with 72%, and SVM with 63%. The authors suggest the potential extension of this framework and ML methodologies for diagnosing other diseases.

Saravananathan and Velmurugan [6] evaluate ‘J48’, ‘CART’, ‘SVM’, and ‘k-Nearest Neighbors’ algorithms on a medical dataset. Their comparison were in light of specificity, precision, sensitivity, error rate and accuracy demonstrates J48's superiority with 67.2% accuracy, subsequently k-NN (53.4%), CART (62.3%), and SVM (65%).

Nai-Arun and Moungmai [5] proposed a web app using disease classifiers and real-world data from 30,122 individuals. Thirteen classification models, including Decision Trees, Neural Networks (NN), Logistic Regression (LR), Naive Bayes, and Random Forest, are examined to identify a predictive model. The RFC method emerges as the most robust, outperforming others in accuracy and Receiver Operating Characteristic (ROC) curve analysis. Its superior performance is attributed to its capability to consider a wide range of variables, enhancing precision in diabetes risk prediction.

Overall, these studies underscore the significance of ML in diabetes risk assessment and diagnosis. By leveraging diverse classification algorithms and datasets, researchers aim to develop accurate and reliable predictive models. The findings highlight the potential of ML techniques to enhance early detection and management of diabetes, paving the way for improved healthcare outcomes. Further research may focus on refining existing models, exploring additional ML algorithms, and extending applications to broader healthcare contexts.

III. PROPOSED METHODOLOGY

The study looks for a model that can more accurately predict diabetes. We tried a various classifications to forecast the presence of diabetes. We go over the phase in brief in the following-

Description of the dataset: The data originates from “Pima Indian Diabetes Dataset”, acquired from “UCI repository”.

Table 1: Description of the dataset

S.No.	Attributes
I	Pregnancy
II	Glucose
III	Blood Pressure
IV	Skin Thickness
V	Insulin
VI	BMI (Body Mass Index)
VII	Diabetes Pedigree Function
VIII	Age

Data Preprocessing: In healthcare datasets, the class variable indicates diabetic outcomes (0 for negative, 1 for positive). Data preprocessing is vital to address missing values and impurities, ensuring the accuracy and effectiveness of machine learning techniques. This process enhances data quality, leading to successful predictions. By preprocessing the data, researchers optimize the dataset for machine learning analysis, thereby improving the accuracy and reliability of predictive models for diabetic outcomes.

[3]

Fig. 1: Proposed Model

Data Selection: The study begins with data selection from the UCI repository, addressing missing values, inconsistencies, and erroneous information.

Data Preparation: Subsequently, databases in Excel and text formats are split into training and testing datasets, with 70% and 30% allocation respectively.

Machine Learning: The techniques of Machine learning such as Naïve Bayes, SVM, and KNN are then employed for prediction, constituting a crucial stage in achieving research objectives. This methodical approach ensures data integrity and enables accurate predictions through diverse machine learning techniques:-

1) **Support Vector Machine-** SVM is an "Supervised Machine Learning" algorithm, commonly abbreviated as SVM. [8] The most used categorization method is SVM. Two classes are divided by a hyperplane made by SVM. In high-dimensional space, it can produce a single hyperplane or a set of hyperplanes. Regression and classification are further uses for this hyperplane. In addition to classifying entities that lack data support, SVM distinguishes between instances within particular classes. The closest training point for each class is reached through the use of a hyperplane for separation.

Algorithm-

- Identify the hyperplane that optimally separates the classes.
- Calculate the margin, which the distance measured between each data point and the hyperplanes.
- A low distance between classes increases the likelihood of misclassification, and vice versa.
- Opt for the class with the highest margin, where the margin is computed as the sum of distances to positive and negative points.

2) **K-Nearest Neighbor-** 'KNN' is another type of "Supervised Machine Learning" algorithm, which assists in addressing both regression as well as classification tasks. KNN presumes that similar objects are situated nearby each other. Data points which are alike are located adjacent to themselves. KNN aids in categorizing novel work in light of similarity metrics. All of the data is documented by the "KNN Algorithm", which then categorizes them on how alike they are accordingly. Uses a tree structure to compute distance between points. This algorithm identifies the closest neighbors of a new data point in the training data set for generating a prediction for it. The value of K, which stands for "number of nearby neighbors," is always positive. The neighbor's value is picked from a set of classes. Euclidean distance is the primary measure used to define "closeness". The "Euclidean Distance" between two points 'P' and 'Q' i.e. P (p1, p2, …. pn) and Q (q1, q2,..qn) is determined by this subsequent formula:-

$$d(P,Q) = \sum_{i=1}^{n} (P_i - Q_i)^2$$

Algorithm-

- Utilize a sample dataset, such as "Pima dataset", consisting of rows and columns.
- Prepare a test dataset containing attributes and rows.
- Calculate the Euclidean distance using the appropriate formula.

$$EuclideanDistance = \sqrt{\sum_{i=1}^{y} \sum_{j=1}^{m} \sum_{l=1}^{n-1} (R_{(j,l)} - P_{(i,l)})^2}$$

- Determine a "random value K", representing number of closest neighbors.
- Utilize the "minimum distance" and "Euclidean distance" to determine nth column for each.

Obtain corresponding output values for the determined columns. If values are identical, then the patient has diabetes, otherwise not.

3) **Naive Bayes Classifier-** The likelihood that an event will occur depends on past knowledge of potential eventrelated circumstances, as determined by Naive Bayes.[10] The most straightforward and quick classification algorithm, Naive Bayes, works well with large data blocks. The NB classifier is used in many different applications, including recommender systems, text categorization, sentiment analysis, and spam filtering. The probability of the unknown classes is predicted using the Bayes theorem. The Naive Bayes algorithm is simple to understand and apply. This is why sparse data sets have the potential to outperform more complex models.

P(h|e) = (P(h|e) * P(h)) / P(e) where,

- "(P(h|e))" signifies 'posterior probability', representing the probability of 'h'(hypothesis) given 'e'11(event).
- "(P(e|h))" signifies likelihood, indicating the probability 'e'(event) given that 'h'(hypothesis) is true.
- "(P(h))" represents 'prior probability', denoting the probability of 'h'(hypothesis) being true.
- "(P(e))" signifies probability of 'e'(event).

occurring.

4) **Gradient Boosting-** Gradient Boosting is a classification technique and the most potent ensemble method for prediction. To create powerful learner models for prediction, it combines week learners collectively. The Decision Tree model is employed. It is a very popular and effective method for classifying complex data sets. The performance of the gradient boosting model gets better with each iteration.

Algorithm-

- Begin with sample desired values, denoted as 'P'.
- Calculate the error present in the desired values.
- Adjust and update the weights to minimize the error, denoted as 'M'.
- Update the target values using the formula P[x] = p[x] + alpha * M[x].
- Analyze and compute the performance of model learners using a loss function F.
- Perform the aforementioned measures until the P(desired result) is achieved.

5) **Logistic Regression-** The probability in logistic regression establishes if a particular data entry belongs to the class indicated by the number [4] ('Brownlee', 2016c). The data is modeled using 'sigmoid function' in logistic regression in the following ways:

$$P(X) = \frac{1}{1+e^{-y}}$$

In this case, 'y' represents the real numerical value, 'e' is base of the natural logarithms, and 'P(X)' is probability that X lies between 0 and 1.

6) **Random Forest–** For problems requiring regression and classification, this kind of collective learning approach is used. On comparing with other models, the accuracy it offers is higher. Huge datasets can be managed with ease through this method. "Leo Bremen" is the one who created Random Forest. It seems to be highly liked method for group learning. On decreasing variance, Random Forest Enhances Decision Tree Performance. To operate, this algorithm constructs numerous decision trees in the training phase. It subsequently produces a classification based on either the average prediction (for regression) from the individual trees or the consensus classification of all the trees combined.

Algorithm-

- Begin by selecting "K" features from the total "M" features, where K is significantly smaller than M.
- Identify the best split point within the selected "R" features for each node.
- Based on the optimal split, break the node into subnodes.
- Repeat steps 1 to 3 until the desired number of nodes, denoted as "l," has been achieved.
- Construct a forest by iteratively repeating steps 1 to 4 for a certain amount of times, creating "z" total trees.

IV. RESULTS AND DISCUSSION

Different actions were taken in this work. The suggested method is implemented in Python and makes use of various ensemble and classification techniques. These techniques are commonplace machine learning techniques meant to extract maximum accuracy from the data. We can observe from this work that the random forest classifier performs better than the others. All things considered, we have achieved high performance accuracy in prediction by utilizing the best machine learning techniques. The outcome of these machine learning techniques is displayed in graphs. The "Receiver Operating Characteristic Area Under the Curve" (ROC AUC)[Figure 2] is a performance measure used to evaluate the capacity of a classification model to distinguish between different classes. The graph demonstrates the relationship between the "True Positive Rate" (TPR) and "False Positive Rate" (FPR). 'TPR', or True Positive Rate, is the ratio of correctly categorized positive cases. On the other hand, FPR, or "False Positive Rate", symbolizes ratio of negative instances that are erroneously classified as positive. The AUC values vary between 0 and 1, where a value of '1' represents 'immaculate classifier' and 0.5 suggests 'random guessing'. The Accuracy Score[Figure 1] measures the ratio of accurately categorized cases by the model.

Upon analyzing the plotted data, it is evident that Logistic Regression and Random Forest display the most superior ROC AUC and Accuracy Scores, but Naive Bayes exhibits the poorest performance. The x-axis displays a selection of methods, including 'SVM', 'Decision Tree', 'Gradient

Boosting classifiers', 'Random Forest', 'KNN' and 'Logistic Regression'.

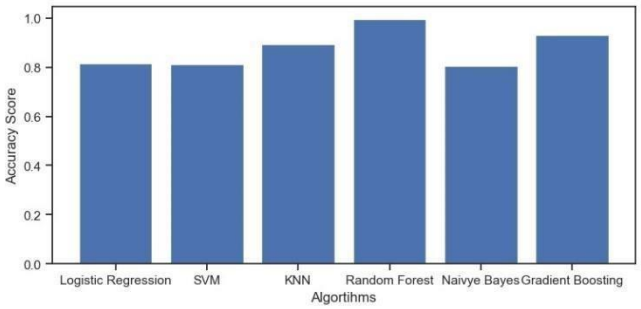

Figure 1: Comparative Analysis of Machine Learning Algorithms (Accuracy)

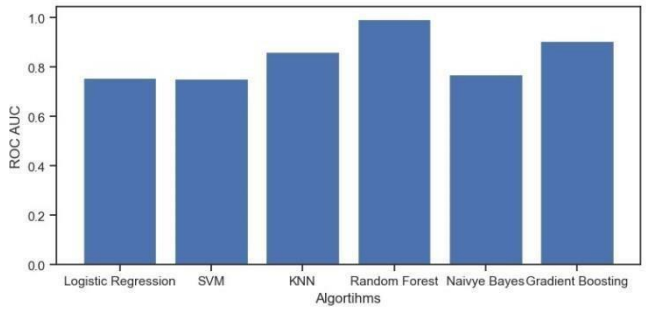

Figure 2: Comparative Analysis of Machine Learning Algorithms (ROC AUC)

V. CONCLUSION

In conclusion, the project successfully achieved its primary objective of implementing machine learning techniques to predict diabetes as well as conducting analysis of the functionality. The suggested framework integrates ensemble learning and classification techniques, such as 'SVM', 'Decision Tree', 'Gradient Boosting classifiers', 'Random Forest', 'KNN' and 'Logistic Regression'. The experiment outcomes offer valuable insights for medical professionals, enabling them to make early predictions and informed decisions for diabetes treatment, ultimately contributing to saving lives.

VI. FUTURE WORKS

Although various datasets were utilized for multiple experiments, there remains ample scope for further research and development by incorporating a diverse range of deep learning techniques. Future endeavors will involve exploring larger and more comprehensive datasets containing additional attributes to enhance prediction accuracy. Additionally, plans are in place to deploy the web app Amazon Web Services as well as other cloud services, enabling free access for actual users to evaluate its effectiveness.

VII. REFERENCES

1. WHO/IDF 2006 (2007, Jan.). Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia, World Health Organisation [Online]. Available: http://www.who.int/diabetes/publications/Definition%20and%20diagnosis%20of%20diabetes_new.pdf
4.Brownlee, J. (2016c). Logistic regression for machine learning.https://www.geeksforgeeks.org/understandinglogist ic-regression/ Accessed: 2021-03-20.

5. Nai-Arun N, Moungmai R. Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci.

2. International Diabetes Federation, Diabetes Atlas, 3rd ed. Brussels, Belgium:International Diabetes Federation.

3. 'Classification of Diabetes Patients Using Kernel-Based Support Vector Machines, 'authored by G.A. Pethunachiyar and presented at the 2020 International Conference on Computer Communication and Informatics (ICCCI)

2015:69:132–42.

6. Saravananathan K, Velmurugan T. Analyzing diabetic data using classification algorithms in data mining. Indian J Sci Technol. 2016:9:1–6.

7. K. Anandha Kumar, "A survey on diabetes mellitus prediction using machine learning techniques," International Journal of Applied Engineering Research, vol. 11, 2022.

8. S. V. K. R. Rajeswari and P. Vijayakumar, "Prediction of diabetes mellitus using machine learning algorithm," Annals of the Romanian Society for Cell Biology, vol. 25, pp. 5655– 5662, 2021.

9. Shafi S, Ansari GA. Early prediction of diabetes disease &classification of algorithms using machine learning approach. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021) Available from:SSRN 3852590 (2021).

10. Sadhu A, Jadli A. Early-stage diabetes risk prediction: A comparative analysis of classification algorithms. Int Adv Res J Sci Eng Technol (IARJSET) 2021;8:193–201.