Enhancing PID Mellitus Classification through Decision Trees and Random Forest Algorithms

J Chaitanya[1], Anthappagudem Samatha[2], V. Shirisha[3] and Remalli Rohan[4]
[1]*Lecturer in Computer Science, Telangana Social Welfare Residential Degree College for Women, Jagathgirigutta, Hyderabad, India.*
[2]*Assistant professor, Department of Computer Science and Engineering (AIML), Gokaraju Rangaraju Institute of Engineering and technology, Hyderabad, India.*
[3]*Junior Lecture, Department of Computer Science and Engineering, Govt. Junior college, Zaheerabad, Hyderabad, India.*
[4]*Computer Science Educator, Researcher, Hyderabad, India.*

***Abstract*—Diabetes mellitus, or just diabetes, is becoming more widespread. Diabetes presents considerable hurdles in prognosis because it cannot be cured but can be efficiently controlled with early detection. Manual assessments for early detection might be problematic since they rely on healthcare professionals' observations, which may miss important trends. As a result, automated computer-based analysis is critical for the early detection of Pima Indian Diabetes (PID). This study seeks to provide an automated system for detecting and classifying PID. The dataset, supplied by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), is intended to predict diabetes risk using a variety of diagnostic indicators. The dataset specifically contains Pima Indian women aged 21 and up, drawn from a larger population. After gathering the dataset, it was pre-processed to identify and remove null values. Next, exploratory data analysis (EDA) was used to investigate the correlations between the attributes. Feature selection was then performed to prepare the data for machine learning classifiers. Finally, eight input variables were found, with one target variable employed by twelve machine learning methods. Among these, the Decision Tree (DT) and Random Forest (RF) classifiers were the best in detecting PID diabetes. This study emphasises the importance of automated procedures in diabetes diagnosis and management, as they provide a vital tool for early intervention.**

***Index Terms*—Diabetes, Decision tree classifier, EDA, Pima India Dataset, Random Forest classifier.**

I. INTRODUCTION

Diabetes is a major global health issue, with a fast increasing prevalence. In 2016, diabetes was the seventh biggest cause of early death, with the World Health Organisation (WHO) reporting 1.6 million fatalities each year [1]. By the end of 2014, the number of persons with diabetes had increased from 108 million (4.2%) to 422 million (8.5%) [2]. Current estimates imply that there are approximately 500 million diabetics worldwide, with projections indicating growth of 25% and 51% by 2030 and 2045, respectively [3].

On World Diabetes Day 2018, the WHO emphasised the critical need to address this issue, stating that one in every three persons is overweight. Between 1980 and 2014, the prevalence of diabetes among adults rose from 4.7% to 8.5%, primarily in developing countries [4]. By 2017, there were 451 million diabetics, with the figure anticipated to rise to 693 million by 2045 [5]. Diabetes cannot be cured, but effective management and prevention are feasible with early diagnosis.

1.1 Pima Indian Diabetes (PID)

Non-insulin-dependent diabetes mellitus is especially common in some ethnic groups, most notably numerous American Indian tribes. Pima Indians in Arizona have the greatest known prevalence of this illness. This essay investigates the epidemiology of diabetes in this community. Since 1965, the Gila River Indian Community, which includes the closely related Tohono O'odham and Pima Indians, has taken part in a longitudinal diabetes study. The majority of the data for this analysis, including prevalence, incidence, risk factors, and aetiology, came from biannual examinations that included oral glucose tolerance tests and assessments of diabetes-related comorbidities.

II. RELATED WORK

Understanding diabetes risk factors and prevalence is critical for successful public health planning. Sourav Kumar Bhoi et al. [6] The Pima Indian Diabetes dataset was analysed using supervised learning methods, and Logistic Regression proved to be the most effective. The suggested IoMT application achieved a sensitivity rate of 88.43% to 89.92% using the J48 decision tree model. Huma Naz and Sachin Ahuja [7] achieved an amazing 98.07% accuracy in diabetes diagnosis using deep learning, hence improving autonomous diagnostic systems. Nahla H. Barakat et al. [8] developed a smart SVM model, emphasising that early detection can avert 80% of type 2 diabetes complications. Other investigations looked into data mining and machine learning approaches. Stefan Ravizza et al. [9] evaluated sickness risk, whereas Vaishali et al. [10] employed Goldberg's Genetic Algorithm and achieved 83.04% accuracy. Kumar Das et al. [11] achieved 90% accuracy with Random Forest, demonstrating the usefulness of diverse approaches in diabetes prediction.

III. METHODOLOGY

The first stage is to prepare the dataset, followed by pre-processing tasks including imputation, standardisation, and managing missing and categorical information. Features for identifying Pima Indian Diabetes (PID) are chosen, and various machine learning techniques are used in Python Jupyter Notebook.

3.1 Pima India Diabetes Data Collection

The National Institute of Diabetes and Digestive and Kidney Diseases is the original source of this dataset [12].

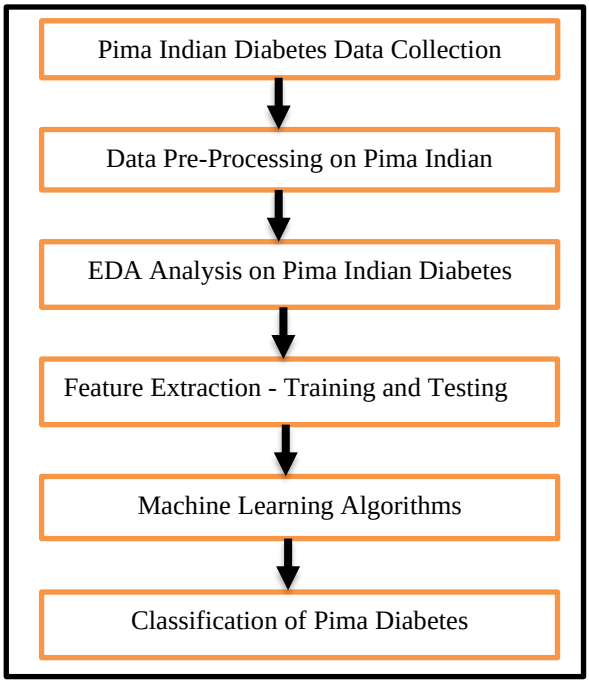

Fig.1: Block diagram of PID classifications

Because of the high prevalence of diabetes, the Pima Indian Diabetes (PID) cohort study has been running since 1965. Its goal is to assess diabetes risk based on diagnostic criteria. The dataset contains 768 instances, with 268 diagnosed with diabetes and 500 without. Pregnancies, BMI, insulin level, age, blood pressure, skin thickness, and glucose are all significant risk factors for diabetes, as shown in Table 1. Figure 1 depicts the categorization procedure.

Table 1: Overview of Pima Indian Diabetes Dataset

S.no	Feature	Description	Data Type	Range
1	Pregnancies	This attribute denotes the number of times a woman has been pregnant.	Numeric	(0,17)
2	Glucose	Plasma glucose concentration exceeded 2 hours during an oral glucose tolerance test.	Numeric	(0,199)
3	Blood Pressure	Blood pressure (mm Hg), the force that propels blood throughout the circulatory system is the blood pressure.	Numeric	(0,122)
4	Skin Thickness	This metric represents the thickness of the triceps skinfold (in millimeters).	Numeric	(0,99)
5	Insulin	The 2-hour serum insulin level (mu U/ml) is expressed by this parameter.	Numeric	(0,846)
6	BMI	Body Mass Index - This parameter describes the body mass index (weight in kg/height in m)2.	Numeric	(0, 67.1)

7	DPF	Diabetes Pedigree Function (DPF) is a function that rates the likelihood of diabetes based on family history.	Numeric	(0.078, 2.42)
8	Age	This field denotes the person's age (in years).	Numeric	(21,81)
9	Outcome	As a result, this parameter is a class variable. One indicates diabetes, while zero indicates no diabetes.	Categorical	(0,1)

3.2 Data Processing on PID Dataset

Collecting data is an important phase in the implementation process. It entails converting raw data into a clean, processed format for use in algorithms. The purpose of data preparation is to produce clear and accurate results. In Jupyter Notebook, use the isnull() command.sum() checks for null values. Figure 2 shows that the PID dataset contains no null values. Next, we look for zeros in the dataset. Five columns—glucose, blood pressure, skin thickness, insulin, and BMI—have zero values. The Age and Diabetes Pedigree Functions do not have a minimum zero value; hence they do not require replacement. Zero values will be replaced by median and mean values.

```
Duplicated rows are:
 0


 Null values per column are:
 Pregnancies                 0
Glucose                      0
BloodPressure                0
SkinThickness                0
Insulin                      0
BMI                          0
DiabetesPedigreeFunction     0
Age                          0
Outcome                      0
dtype: int64


 Zero values per column are:
 Pregnancies                111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64 ,
```

Fig.2: No null and no duplicate values in PID dataset

Preprocessing Outcome

As observed in the preprocessing result, the dataset is free of duplicated and missing values. The statistical portion of this dataset demonstrates that Pregnancies: The average number of pregnancies is roughly 3.84, with a standard deviation of 3.37. The maximum number of documented pregnancies is 17. Glucose: Glucose levels are 120.89 on average, with a standard deviation of 31.97. The minimum value is 0 which is not medically possible and indicates missing or inaccurate data. Blood Pressure: The average blood pressure is around 69.10 with a standard deviation of 19.36. A blood pressure of zero, like glucose, is not conceivable and indicates missing or inaccurate data. Skin Thickness: The average skin thickness is roughly 20.54, with a standard variation of 15.95. There are additional records with skin thickness of 0, indicating missing or erroneous data. Insulin: The average insulin level is around 79.80, with a standard deviation of 115.24. Records with an insulin level of 0 also indicate missing or inaccurate data. BMI: The average BMI is around 31.99, with a standard deviation of 7.88. A BMI of 0 is not possible and shows that there is missing or erroneous data. Diabetes Pedigree Function: The average value is roughly 0.47, with a standard deviation of 0.33. Age: The average age is roughly 33.24, with a standard variation of 11.76. Outcome: Diabetes affects 34.9% of the patients in the dataset.

3.3 Exploratory Data Analysis (EDA) on Pima Indian Diabetes

Exploratory data analysis (EDA) is a critical stage in exploring data and identifying patterns or odd values. It aids in understanding facts before drawing any judgements. Clustering and other sorts of visualizations (univariate, bivariate, and multivariate) are examples of EDA approaches. Graph creation is made easier using libraries like NumPy and Seaborn. A histogram, for example (shown in Figure 3), is used to illustrate how data values are distributed. It is useful for summarizing both discrete and continuous data effectively. EDA is critical for validating ideas and preparing for future investigation.

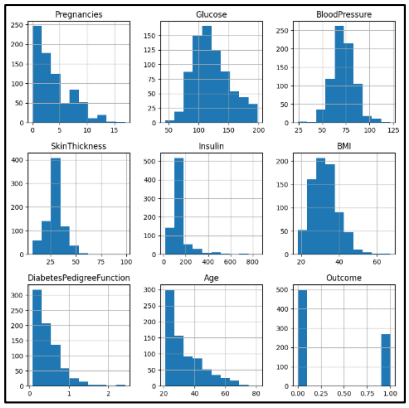

Fig.3: Histogram visualization

3.3.1 Outliers

Outliers in data can be identified using the interquartile range (IQR). They are defined as values less than Q1 - 1.5 IQR or greater than Q3 + 1.5 IQR. In a boxplot, the 'whiskers' represent the greatest and lowest values inside the range, whereas outliers appear as isolated points (as shown in Figure 4).

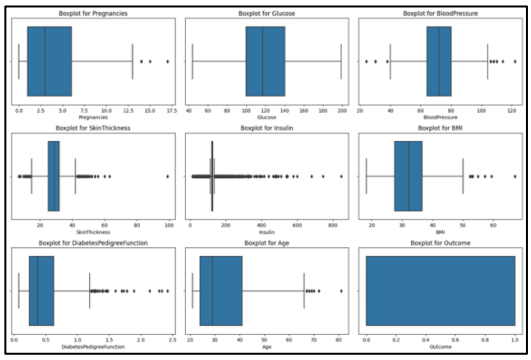

Fig.4: Boxplot of PID dataset attributes.

For example, the insulin feature contained an outlier with a maximum value of 846 U/ml, which is physiologically impossible. Outliers were found and deleted using the variable's lowest, highest, and quartile values. Outliers in DPF, age, insulin, glucose, BMI, and blood pressure are observed in figure 8, which could be attributable to other underlying variables. It is best to standardize the data to avoid the negative impact of outliers. Because the dataset isn't very huge, it's best to avoid eliminating rows that aren't absolutely necessary.

3.3.2 Correlation Matrix

The correlation coefficients range between -1 and 1. If the correlation coefficient is near to one, there is a strong positive correlation between the two variables. The variables have a high negative association when it is close to -1. Glucose, age, and BMI are all moderately associated to outcome. Pregnancies and age have a strong relationship as shown in the figure 5. Pair plots are used to comprehend pairwise relationships. This can be quite useful in understanding how the variables interact with one another and identifying any potential patterns or trends.

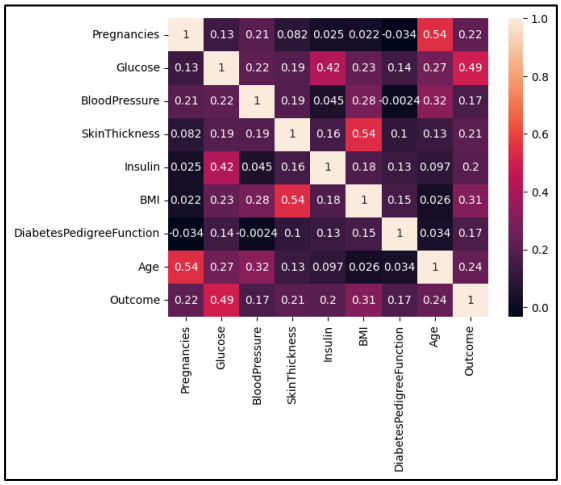

Fig.5: Correlation Matrix of PID Matrix

3.4 Feature Extraction

We split our data into two sets: a training set of 614 samples and a test set of 154 samples. Next, we'll scale the features. Scaling is not required for many algorithms (such as decision trees and random forests), but it does benefit others, such as logistic regression and SVM. We'll use scikit-learn's StandardScaler to standardise the features, ensuring they have a mean of 0 and a variance of one. This standardisation technique applies to all predictor variables, including glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree, pregnancy, and age. Following scaling, we will train and assess multiple models using a predefined function to streamline our process. The scikit-learn function Train_Test_Split() divides the data into 80% training and 20% testing.

3.5 Machine Learning Algorithms

After completing the pre-processing stage and dividing the training/testing sets, we fitted multiple machine learning models. This section delves into the many approaches used to separate persons with and without diabetes. Machine learning is used to create models that map the available inputs (independent variables or features) to the desired output (dependent variable or target). Our categorization challenge

involves categorizing Pima Indian women as "Diabetic" or "Not Diabetic." For this experiment, we used a variety of classification methods, including KNN, Decision Tree, Gaussian Naive Bayes, Linear Discriminant Analysis, Support Vector Classifier (SVC), Linear SVC, AdaBoost, Random Forest, Perceptron, Bagging, Logistic Regression, and Gradient Boosting Classifiers.

3.5.1 Decision Tree Classifier

A decision tree is a prediction model that is widely used in operations research and decision analysis. It visualises the links between the properties and values of various variables. The paradigm resembles a tree structure, with conditional statements displaying numerous conditions and their potential consequences. Each node in the decision tree represents a specific attribute, and each branch represents a possible value for that attribute. There are three types of nodes: decision, chance, and end. The path from a decision node to an end node depicts one possible outcome, with each node serving as a "test" of an attribute and each branch reflecting the test's outcome.

3.5.2 Random Forest Classifier

Decision trees are a popular machine learning method, although they are susceptible to overfitting due to their inclination to expand indefinitely, resulting in minimal bias and significant variance. To overcome this issue, random forests were created as a classification method that combines numerous decision trees. In a random forest, several decision trees are built from random subsets of the candidate variables, resulting in substantially uncorrelated models. This ensemble technique improves prediction accuracy since the different trees can adjust for each other's faults. Random forests excel in processing high-dimensional data quickly, without the requirement for dimensionality reduction or feature selection. They may also assess the value of different qualities and how they interact. Importantly, random forests preserve accuracy even with a considerable number of attributes are missing, making them a robust choice for classification tasks.

IV. EXPERIMENTAL RESULTS

In this investigation, we used twelve different machine learning approaches. Their accuracy levels are as follows: KNeighbors Classifier - 0.81, Decision Tree Classifier - 1.0, Gaussian NB - 0.75, Linear Discriminant Analysis - 0.77, SVC - 0.77, Linear SVC - 0.65, AdaBoost Classifier - 0.83, Random Forest Classifier - 1.0, Perceptron - 0.60, Bagging Classifier - 0.99, Logistic Regression - 0.77, and Gradient Boosting Classifier - 0.93. The Decision Tree and Random Forest classifiers achieved the maximum accuracy, as shown in the accompanying figure, which compares the accuracies of various machine learning models for predicting and categorizing PID diabetes.

```
==========================================
Model: DecisionTreeClassifier
Accuracy: 1.0
Confusion Matrix:
[[401   0]
 [  0 213]]
Classification Report:
              precision    recall  f1-score

           0       1.00      1.00      1.00
           1       1.00      1.00      1.00

    accuracy                           1.00
   macro avg       1.00      1.00      1.00
weighted avg       1.00      1.00      1.00
```

Fig.12: Accuracy Decision tree classifier.

```
==========================================
Model: RandomForestClassifier
Accuracy: 1.0
Confusion Matrix:
[[401   0]
 [  0 213]]
Classification Report:
              precision    recall  f1-score

           0       1.00      1.00      1.00
           1       1.00      1.00      1.00

    accuracy                           1.00
   macro avg       1.00      1.00      1.00
weighted avg       1.00      1.00      1.00
```

Fig.15: Accuracy Random Forest Classifier.

V. CONCLUSION & FUTURE WORK

According to the World Health Organization (WHO), diabetes kills 1.6 million people each year. Diagnosing diabetes is difficult, and while it cannot be cured, early detection enables efficient treatment and prevention. The goal of this project is to use intelligent approaches to automate the diagnosis and classification of Pima Indian Diabetes (PID). After gathering the dataset, it is preprocessed to remove null values, followed by exploratory data analysis (EDA) to investigate attribute correlations. Features are chosen for training and testing using different machine learning classifiers. Finally, Decision Tree and Random Forest classifiers had the highest accuracy in diagnosing PID diabetes, which improved clinical decision-making.

Future work on this project will include adjusting hyperparameters to improve performance. Embedded devices such as Field Programmable Gate Arrays (FPGAs) can help improve classification accuracy and training time for Pima Indian Diabetes (PID). Understanding diabetes risk factors helps with public health planning and implementation of preventative strategies. Furthermore, classification models can aid in the identification of molecular indicators associated with diabetes, hence encouraging additional study and the creation of new treatments and therapies.

REFERENCES

[1] "Global Report on Diabetes, 2016". Available at: https://www.who.int/

[2] "Diabetes: Asia's 'silent killer'", November 14, 2013". Available at: https://www.bbc.com/news/world-asia-24740288

[3] The Emerging Risk Factors Collaboration, ''Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies,'' Lancet, vol. 375, pp. 2215–2222, Jun. 2010.

[4] N. H. Cho, et al., ''IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,'' Diabetes Res. Clin. Pract., vol. 138, pp. 271–281, Apr. 2018.

[5] P. Saeedi, et al., ''Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition,'' Diabetes Res. Clin. Pract., vol. 157, Nov. 2019, Art. no. 107843.

[6] Sourav Kumar Bhoi, et al., "Prediction of Diabetes in Females of Pima Indian Heritage: A Complete Supervised Learning Approach", Turkish Journal of Computer and Mathematics Education Vol.12 No.10 (2021), 3074-3084.

[7] Huma Naz and Sachin Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset", Journal of Diabetes & Metabolic Disorders (2020) 19:391–403.

[8] Barakat N, Bradley AP, Barakat MNH, "Intelligible support vector machines for diagnosis of diabetes mellitus", IEEE Trans Inf Technol Biomed", 2010;14(4):1114–20.

[9] Ravizza S et al., "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data", Nature Medicine. 2019;25(1): 57–9.

[10] Vaishali et al., "Genetic algorithm-based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in Proceedings of the 2017 International Conference on Computing Networking and Informatics (ICCNI), pp. 1–5, IEEE, 2017, October.

[11] S. Kumar Das, A. Kumar Mishra, and P. Roy, "Automatic diabetes prediction using tree-based ensemble learners," International Journal of Computational Intelligence & IoT, vol. 2, no. 2, 2019.

[12] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.