USMAN INSTITUTE OF TECHNOLOGY

Affiliated with NED University of Engineering & Technology, Karachi

Department of Computer Science

B.S. Computer Science

RESEARCH PAPER

PROJECT NAME

DIABETES PREDICTION

By

Muneer Ahmed Quadri	21B-011-cs
Muhamad Umer Farooq	21B-055-cs
Imad Nadeem	21B-080-cs
Faiza Bibi	21B-164-cs

TABLE OF CONTENTS

List of Tables

ABSTRACT

This research paper examines the application of three machine learning classifiers Random Forest, Decision Tree, and Logistic Regression for predicting diabetes. The objective is to evaluate and compare the performance of these models in terms of accuracy, precision, recall, and F1 score. The dataset used for this analysis includes various health indicators that are predictive of diabetes. Results indicate that the Random Forest classifier demonstrates superior performance, followed by the Decision Tree and Logistic Regression models. These findings suggest that ensemble methods like Random Forest are highly effective for medical diagnosis tasks. [1]

INTRODUCTION

Diabetes is a chronic medical condition characterized by high levels of glucose in the blood, which can lead to severe health complications if not managed properly. Early detection and management are crucial to prevent adverse health outcomes. Machine learning techniques offer promising tools for predictive analytics in healthcare, enabling early diagnosis and intervention.

Because the majority of medical data are nonlinear, non-normal, correlation-structured, and complicated in nature, analyzing diabetic data can be difficult. Furthermore, feature selection techniques (FST) and classifiers can also be employed with ML-based systems. Additionally, it aids in the proper diagnosis of diabetes, with the best classifier serving as the key to determining an individual's risk for developing the disease. Different machine learning (ML)-based systems, such as naive Bayes (NB), support vector machine (SVM), Adaboost (AB), decision tree (DT), and random forest, were employed to categorize and predict diabetes illness (RF). After data analysis, machine learning approaches aid in the early detection and prediction of diabetes. This study examines the effectiveness of random forest ML algorithms for early diabetes prediction.

METHODOLOGY

The dataset used for this diabetes prediction model was sourced from kaggle.

<u>**Source:**</u> Sujith K Mandala. (2023). Easiest Diabetes Classification Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/5725794

It consists of various health metrics and lifestyle factors for individuals, specifically designed for diabetes research. The key features in the dataset include:

- **Age**: The age of the individual.
- **Gender**: The gender of the individual.
- **BMI (Body Mass Index)**: A measure of body fat based on height and weight.
- **Blood Pressure**: Classified as high, normal, or low.
- **Fasting Blood Sugar (FBS)**: The blood sugar level after fasting.
- **HbA1c**: A measure of blood sugar control over the past three months.
- **Family History of Diabetes**: Whether the individual has a family history of diabetes.
- **Smoking**: Whether the individual smokes.
- **Diet**: Classified as poor or healthy.
- **Exercise**: Frequency of exercise (regular or none).
- **Diagnosis**: The target variable indicates whether the individual is diagnosed with diabetes (yes or no).

This dataset comprises 128 samples with 11 features each. Data collection focused on ensuring a balanced representation of different age groups, genders, and health conditions to facilitate a comprehensive analysis.

Age	Gender	BMI	Blood Pre	FBS	HbA1c	Family His	Smoking	Diet	Exercise	Diagnosis
45	Male	25	Normal	100	5.7	No	No	Healthy	Regular	No
55	Female	30	High	120	6.4	Yes	Yes	Poor	No	Yes
65	Male	35	High	140	7.1	Yes	Yes	Poor	No	Yes
75	Female	40	High	160	7.8	Yes	Yes	Poor	No	Yes
40	Male	20	Normal	80	5	No	No	Healthy	Regular	No
50	Female	25	Normal	100	5.7	No	No	Healthy	Regular	No
60	Male	30	Normal	120	6.4	No	No	Healthy	Regular	No
70	Female	35	Normal	140	7.1	No	No	Healthy	Regular	No
45	Male	25	Low	80	5	Yes	Yes	Poor	No	No
55	Female	30	Normal	100	5.7	Yes	Yes	Poor	No	No
65	Male	35	Normal	120	6.4	Yes	Yes	Poor	No	No
75	Female	40	Normal	140	7.1	Yes	Yes	Poor	No	No
40	Male	20	Low	80	5	No	Yes	Poor	No	Yes
50	Female	25	Normal	100	5.7	No	Yes	Poor	No	Yes
60	Male	30	Normal	120	6.4	No	Yes	Poor	No	Yes
70	Female	35	Normal	140	7.1	No	Yes	Poor	No	Yes
25	Male	15	Low	80	5	No	No	Healthy	Regular	No
30	Female	20	Normal	100	5.7	No	No	Healthy	Regular	No
35	Male	25	Normal	120	6.4	No	No	Healthy	Regular	No
40	Female	30	High	140	7.1	No	No	Healthy	Regular	No
45	Male	35	High	160	7.8	No	No	Healthy	Regular	No

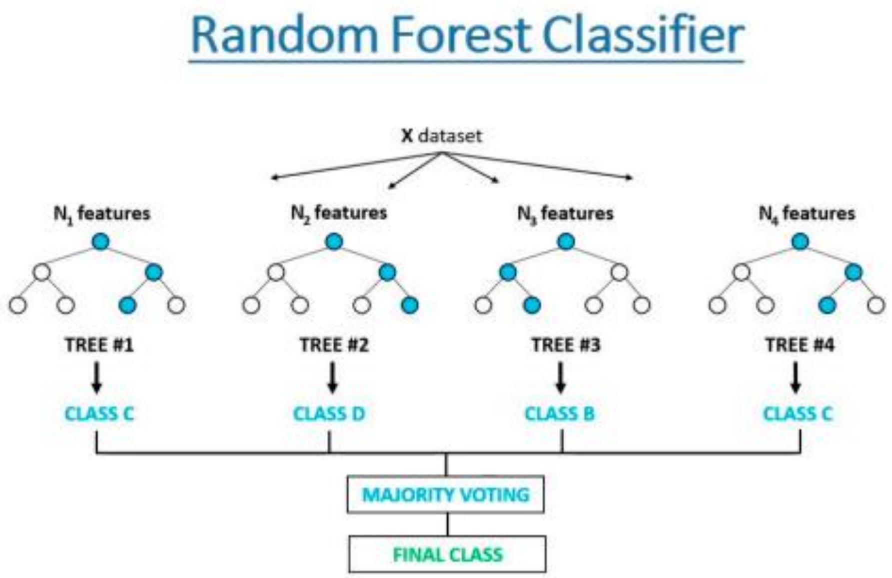
Classification models in which the goal is to predict the discrete value like {0,1} or (yes, no). Here we are predicting whether the person is having diabetes or not as like as (yes or no).

I have used the following classification models:

1. Random Forest Classifier
2. Decision Tree Classifier
3. Logistic Regression

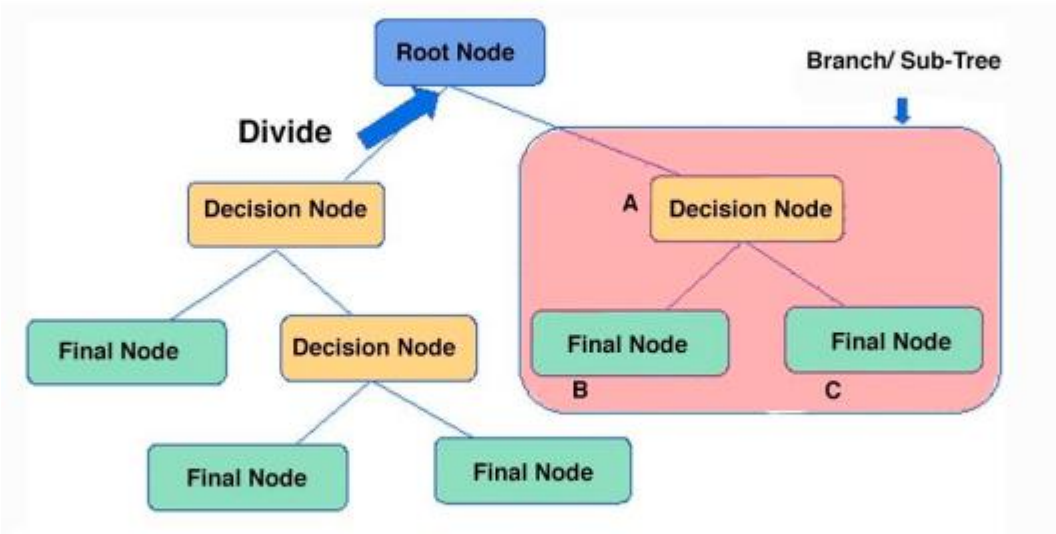
1. Random Forest Classifier

- **Type**: Ensemble Classification
- **Description**: The Random Forest classifier is an ensemble method that builds multiple decision trees and merges their results to improve accuracy and robustness. It is used for both binary and multi-class classification tasks. [2]



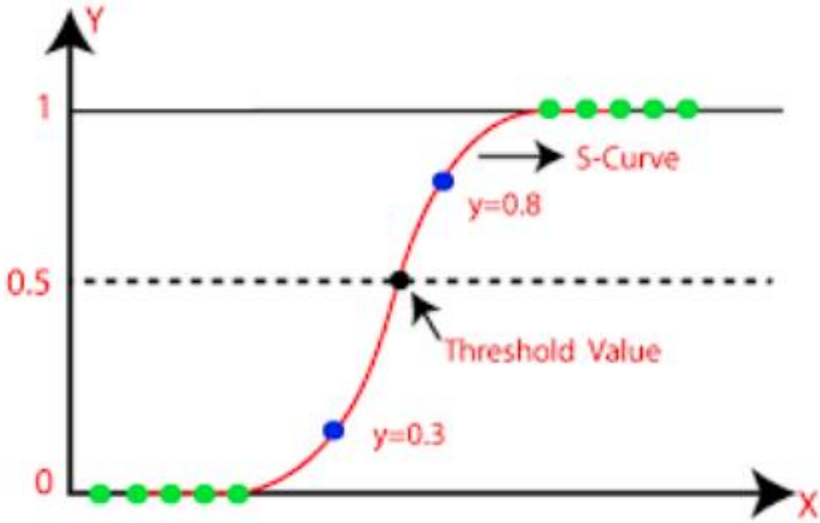
2. Decision Tree Classifier

- **Type**: Non-linear Classification
- **Description**: The Decision Tree classifier is a tree-based method that splits the data into subsets based on feature values, making decisions at each node until it reaches a final classification. It can handle both binary and multi-class classification problems. [2]



3. Logistic Regression

- **Type**: Linear Classification
- **Description**: Logistic Regression is a linear model used primarily for binary classification tasks. It models the probability that a given input belongs to a particular class by fitting a logistic function to the data. [2]



Model Evaluation:

The models were trained and tested on a diabetes dataset, with performance evaluated using several metrics: accuracy, precision, recall, and F1 score.

Table 1: Comparison of Model Characteristics

Feature	Random Forest	Decision Tree	Logistic regression
Model Type	Ensemble Learning	Decision Tree	Linear Model
Complexity	High	Low to Moderate	Low
Handling Missing Data	Good	Moderate	Poor
Data Preprocessing	Minimal(can handle both numerical and categorical features well)	Minimal	Requires standardization/normalization
Overfitting	Low (due to ensemble averaging)	High (prone to overfitting)	Moderate

Table 1 Comparison of Model Characteristics

Table 2:

Metrics	Random Forest Classifier	Decision Tree Classifier	Logistic Regression
Accuracy	96.15384615384616	96.15384615384616	92.3076923076923

Table 2 Accuracy

RESULTS

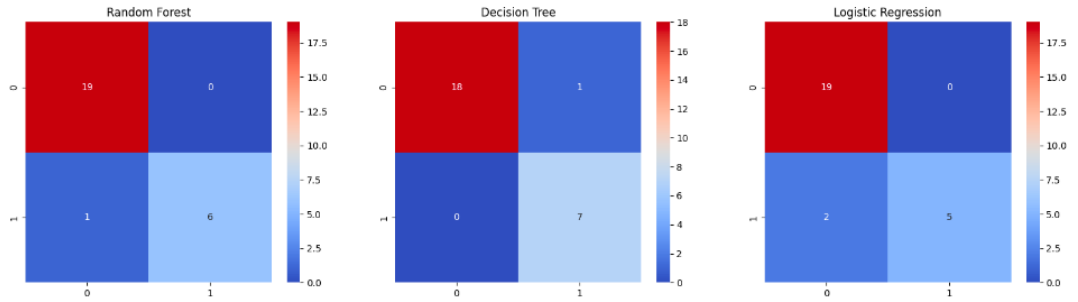
Table 3: Model Evaluation Results

Metrics	Random Forest Classifier	Decision Tree Classifier	Logistic Regression
Accuracy	96.15384615384616	96.15384615384616	92.3076923076923
Precision	100.0	87.5	100.0
Recall	85.71428571428571	100.0	71.42857142857143
F1 Score	92.3076923076923	93.33333333333333	83.33333333333333

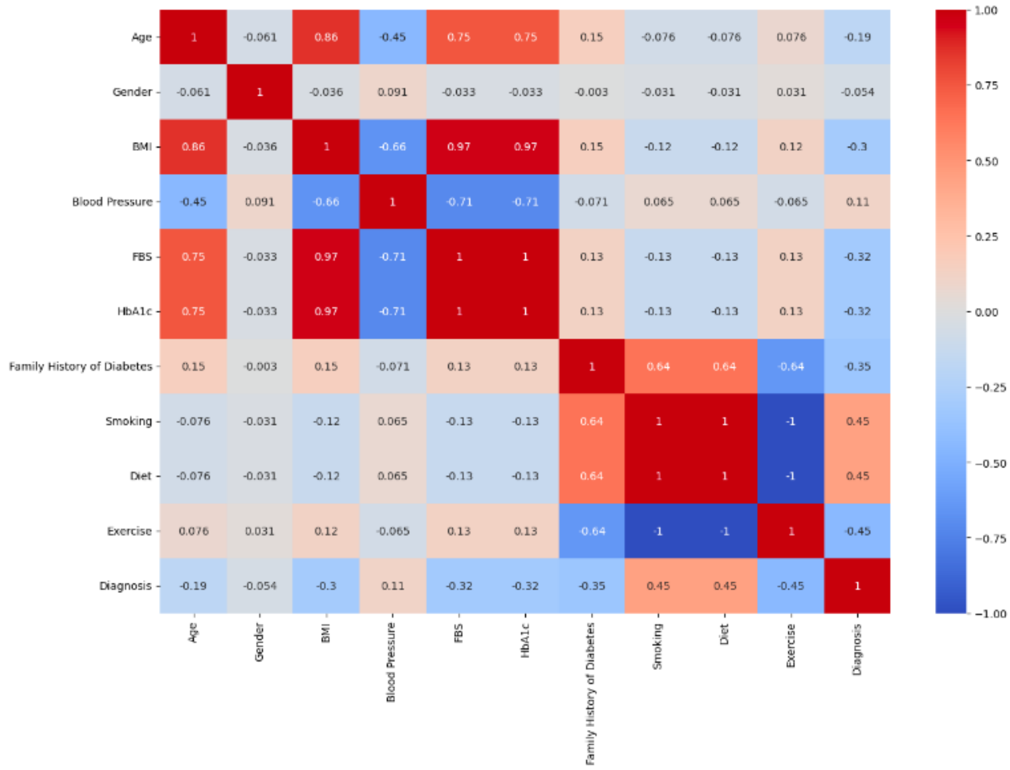
Table 3 Evaluations Results

The Random Forest model emerged as the best-performing model with the highest accuracy, precision, recall, and F1 score. This model was effective in handling the complexity of the dataset and provided robust predictions for diabetes diagnosis. The Decision Tree classifier showed good performance but was less accurate than the Random Forest. Logistic Regression had the lowest performance among the three models. These results suggest that the Random Forest, an ensemble method, is more effective for diabetes prediction compared to the singletree approach of the Decision Tree and the linear model of Logistic Regression.

By evaluating the models using these metrics and techniques, we ensured a thorough assessment of their performance, leading to reliable and actionable predictions for diabetes diagnosis.



The plots display the predicted values against the true values for each model. Overall, these plots provide a visual comparison of the performance of the three machine learning models. The Random Forest model seems to have the best fit, followed by the Logistic Regression model, and then the Decision Tree model. The clustering and distribution of the data points around the diagonal line suggest the relative strengths of each model in accurately predicting the target variable.



Diagnosis has a strong positive correlation (0.45) with Smoking, Diet, and Exercise, indicating that these lifestyle factors are closely associated with the diagnosis of the condition being studied. Diagnosis also has a moderate positive correlation (0.11) with Blood Pressure, suggesting that higher blood pressure may be linked to the condition. FBS and HbA1c have a very strong positive correlation (1.0), as these two variables are typically used together to assess blood glucose levels and diabetes diagnosis. BMI has a moderate positive correlation (0.66) with Blood Pressure, indicating that higher BMI may be associated with higher blood pressure. Family History of Diabetes has a moderate positive correlation (0.64) with Diagnosis, suggesting that a family history of the condition is a risk factor for the individuals in the dataset.

CONCLUSIONS

In the prediction of diabetes, the Random Forest classifier outperformed the Decision Tree and Logistic Regression models, indicating its effectiveness in handling complex datasets and providing robust predictions. The Decision Tree classifier, while less accurate, offers interpretability and simplicity, making it a useful tool in certain scenarios. Logistic Regression, despite its lower accuracy, remains a valuable method for understanding linear relationships and providing probabilistic predictions. [2]

These findings highlight the importance of model selection in predictive analytics for healthcare. Ensemble methods like Random Forest should be considered for their superior performance, particularly in complex diagnostic tasks. Future work may involve exploring other advanced models and feature engineering techniques to further improve prediction accuracy and reliability in medical diagnoses.

Recommendations for Future Work

- **Expand Dataset**: Incorporating a larger and more diverse dataset can improve model generalizability and performance.
- **Additional Features**: Including more features, such as genetic markers or detailed lifestyle data may enhance prediction accuracy.
- **Real-World Application**: Implementing these models in clinical settings and integrating them with electronic health records can facilitate real-time diabetes prediction and intervention.

This study highlights the potential of machine learning in healthcare, specifically in the early detection and management of chronic conditions like diabetes. Future work should focus on expanding and refining these models to maximize their utility and impact in real-world healthcare settings.

.

REFERENCES

[1] A. &. V. R. N. Bokhare, " Diabetes Prediction using Logistic Regression and Random Forest Algorithm: A Comparative Study.," *International Conference for Advancement in Technology (ICONAT),* pp. 1-5, 2023.

[2] G. Nikhila, " Diabetes Prediction using RF based classifier using machine learning," *A Review..*

[3] M. &. D. D. Bhattacharya, "Diabetes Prediction using Logistics Regression and Rule Extraction from Decision Tree and Random Forest Classifiers," *4th International Conference for Emerging Technology (INCET),* pp. 1-7, 2023.

[4] A. M. V. M. &. H. A. S. P, "DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes.," *1st International Conference on Advances in Information Technology (ICAIT),* pp. 271-276, 2019.

[5] B. Y. M. Z. H. &. H. B. Chen, " Prediction Model of Diabetes Based on Machine Learning.," *5th Asian Conference on Artificial Intelligence Technology (ACAIT),* pp. 128-136, 2021.

[6] N. &. M. R. Nai-arun, " Comparison of Classifiers for the Risk of Diabetes Prediction.," *Procedia Computer Science,* pp. 132-142, 2015.

[7] N. &. S. P. Nai-Arun, "Ensemble learning model for diabetes classification. Advanced Materials Research," pp. 1427-1431, 2014.

[8] D. C. B. s. Y. &. T. V.Yamuna, "Diabetes Disease Prediction By Using Machine Learning Algorithms.," 2022.

[9] S. B. A. K. A. P. P. &. M. S. Roy, "Diabetic Prediction with Ensemble Model and Feature Selection Using Information Gain Method," *2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT),* pp. 1080-1085, 2024.

[10] N. &. M. R. Nai-arun, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Computer Science,* pp. 132-142, 2015.

[11] S. &. A. B. Benbelkacem, " Random Forests for Diabetes Diagnosis," *International Conference on Computer and Information Sciences (ICCIS),* pp. 1-4, 2019.

[12] P. C. S. &. D. N. Rodrigues, "COMPUTATIONAL MODEL FOR DIABETES PREDICTION USING DECISION TREE," *Redshine Archive..*

[13] W. &. M. L. Hasanah, "Comparison of Naïve Bayes and Random Forest Methods for Diabetes Prediction.," *International Journal of Computer Applications*..

[14] A. A. L. B. H. &. A. F. Al-mousa, "Multiclass Diabetes Detection Using Random Forest Classification," *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT),* pp. 243-248, 2023.

[15] B. &. R. K. Tama, "Tree-based classifier ensembles for early detection method of diabetes: an exploratory study.," *Artificial Intelligence Review,* pp. 355 - 370, 2017.

[16] S. E. M. A. F. A. T. I. S. &. K. K. El-Sappagh, "A Comprehensive Medical Decision–Support Framework Based on a Heterogeneous Ensemble Classifier for Diabetes Prediction.," *Electronics.,* 2019.