The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 19(2): S.I (1), 590-596, 2024 www.thebioscan.com

Utilizing Machine Learning for Predicting Diabetes in Pregnant Women: A Comparative Analysis of Logistic Regression, Random Forest, and Naive Bayes Models

UTHAYA KUMAR.J[1], SARITHA P.S[2], RESHMA P[3], Dr.RAMASAMY.S[4]

1. Assistant Professor, Department of Computer Science Engineering, Hindusthan Institute of Technology, Coimbatore – 32.
2. Assistant Professor, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan College of Engineering, Coimbatore – 105.
3. Assistant Professor, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan College of Engineering, Coimbatore – 105.
4. Associate Professor, Department of Computer Science Engineering, Hindusthan Institute of Technology, Coimbatore – 32.

DOI: https://doi.org/10.63001/tbs.2024.v19.i02.S.I(1).pp590-596

KEYWORDS
Random Forest, Machine Learning, Accuracy

Received on:

20-07-2024

Accepted on:

14-12-2024

ABSTRACT

Diabetes is a significant worldwide health condition that affects millions of people regardless of demographic variations. Diabetes is characterized by increased blood glucose levels that are caused by inadequate insulin synthesis or activity. The purpose of this study is to investigate the use of machine learning and artificial intelligence approaches in the prediction of diabetes, specifically among pregnant women, who are a population that is at a higher risk. Through the use of sophisticated computational methods including Logistic Regression, Random Forest, and Naive Bayes models, the main objective of this work is to forecast the start of diabetes and reduce the difficulties that are connected with it. Out of all of them, the Random Forest method stands out because to its remarkable accuracy rate of 98% on the dataset, which demonstrates its potential for early identification. This research highlights the revolutionary role that machine learning plays in the delivery of prompt medical diagnosis, especially in underprivileged communities where delayed medical treatment often exacerbates health consequences. Through the simplification of the diagnosis procedure, these computational tools make it possible for medical professionals to perform diabetes management in a more effective and proactive manner.

INTRODUCTION

As a result of the fact that diabetes is a widespread worldwide health problem that impacts millions of people across all age groups, it has been given the nickname of the "silent pandemic" of our recent period. As a result of its extensive occurrence and the serious problems it causes, there is an immediate and pressing need for effective measures in the areas of prevention, early detection, and treatment. It is of the utmost importance to successfully address diabetes, as it continues to be one of the major causes of morbidity and death throughout the globe. A consistently increased blood glucose level is the defining characteristic of diabetes. This raised blood glucose level is caused by abnormalities in the insulin regulatory system. Insulin, a hormone that is generated by the pancreas, enhances glucose metabolism in the liver while also facilitating the absorption of glucose by muscle and fat cells. Insulin is an essential hormone at the same time. Together, these processes bring to a reduction in blood sugar levels and help to keep the metabolic system in balance. Individuals who have diabetes, on the other hand, have this route disrupted because they have insufficient insulin synthesis, poor secretion, or resistance to the effects of insulin. Type 1 and Type 2 diabetes are the two most frequent types of diabetes, and they are both caused by this disturbance. Diabetes type 1, which is an autoimmune condition, is responsible for around five percent of all occurrences of diabetes. The condition often presents itself at an early age, when the immune system erroneously targets the beta cells in the pancreas that are responsible for the synthesis of insulin. On the other hand, those who have type 2 diabetes account for about 95% of all instances and are intimately linked to lifestyle factors such as being overweight, having an unhealthy diet, and not being physically active. The development of this kind of diabetes occurs gradually as the cells of the body grow resistant to insulin. This resistance is often accompanied with inadequate insulin production to fulfill the needs of the metabolic process. Diabetes may result in a variety of serious problems if it is not well managed. These complications include neuropathy, cardiovascular disease, retinopathy, and renal failure. This highlights the crucial need for early diagnosis and appropriate intervention in the treatment of diabetes.

Artificial intelligence (AI) and machine learning (ML) are being used into diabetes research, which provides a revolutionary approach to the treatment of this ailment. The applications of these cutting-edge technologies include the analysis of complicated datasets, the forecasting of the course of diseases, and the optimization of treatment techniques. For the purpose of early diagnosis and prevention, machine learning models, in particular, are very useful tools because of their high level of effectiveness in recognizing patterns and correlations in medical data. Researchers hope that by using these skills, they will be able to improve the results for patients and reduce the risk of long-term consequences that are connected with diabetes. This project aims to evaluate the effectiveness of several supervised machine learning algorithms in predicting the early development of diabetes, especially among high-risk populations such as pregnant women. Specifically, the study will concentrate on predicting glucose levels in the blood. Not only does gestational diabetes have an effect on the health of the mother, but it also presents hazards to the development of the fetus, including macrosomia and newborn hypoglycemia. It is essential to detect this condition

at an early stage to prevent complications. Ultimately, the purpose of this study is to make use of machine learning models in order to lessen the likelihood of these dangers, enhance the results for health, and make progress in the general knowledge of diabetes prediction. There are several different categorization methods that have been investigated for their potential in the early diagnosis of diabetes-related conditions. J48 decision trees, Support Vector Machines (SVM), Naive Bayes (NB), and other models such as Decision Tables are examples of these types of models. Previous study (Dean L., 2004; Aishwarya, 2013; Kavakiotis, 2017) has shown that these approaches are very accurate when it comes to identifying a wide variety of illnesses. On the basis of this foundation, the current research intends to evaluate the predictive capacities of certain algorithms, such as Naive Bayes, Logistic Regression, and Random Forest, in order to establish the usefulness of these algorithms in identifying persons who are at risk for developing diabetes.

The study makes use of the Diabetes Risk Examination Dataset, which is obtained from the Kaggle Repository. This dataset includes a wide range of medical and demographic information. This dataset will serve as the foundation for conducting an analysis of the effectiveness of the machine learning models that were selected for the purpose of diabetes prediction. Attributes that are unique to each algorithm include:

In spite of the fact that it is based on the premise that features are independent of one another, the Naive Bayes algorithm is highly regarded for its ease of use and its speed, which makes it an efficient method for managing huge datasets.

Logistic regression is a common option for clinical predictions because it results in data that may be interpreted. This is accomplished by predicting the likelihood of a binary outcome.

The ensemble learning technique known as Random Forest is notable for its resilience and its capacity to deal with non-linear correlations that exist within the data. It constructs multiple decision trees during training and combines their outputs to deliver reliable and accurate predictions, reducing the risk of overfitting.

In this study, Random Forest demonstrates superior performance, achieving a high accuracy rate in predicting diabetes onset. This underscores its potential as a reliable tool for early diagnosis and its applicability in diverse clinical settings. The findings of this research emphasize the transformative role of computational methods in improving the management of diabetes. By enabling early and precise identification of high-risk individuals, these tools pave the way for timely interventions, including lifestyle modifications, dietary guidance, and personalized treatment plans. This is particularly critical in addressing gestational diabetes among pregnant women, where early management can significantly enhance maternal and neonatal outcomes. Moreover, the broader implications of this work are noteworthy. Machine learning models can empower healthcare systems by automating the diagnostic process, reducing dependency on extensive clinical expertise, and providing scalable solutions for resource-constrained settings. In underserved communities, where delayed diagnosis often exacerbates health outcomes, these technologies offer a practical and efficient means of delivering quality care. This research highlights the potential of AI and ML to revolutionize healthcare by bridging the gap between complex data and actionable insights. By leveraging these advanced computational tools, the study not only contributes to the understanding of diabetes prediction but also lays the groundwork for future innovations in disease prevention and management.

Methodology:

Literature Review:

2018 release of Deepti Sisodia In order to develop models that are capable of reliably predicting diabetes in patients, this research makes use of a number of different categorization approaches. These techniques include Decision Trees, Support Vector Machines (SVM), and Naive Bayes (NB). A number of evaluation criteria, including precision, accuracy, recall, and F-measure, were used, and Receiver Operating Characteristic (ROC) curves were utilized to illustrate the improved performance of the Naive Bayes algorithm. The accuracy of NB was much higher than that of the other models, coming in at 76.30%.

Bano Farhana, the year 2021: Multiple organs are profoundly impacted by diabetes mellitus, which is determined by a number of variables including genetics, family history, health, and environmental influence. The purpose of this research was to investigate the use of autonomous learning approaches for the prediction of diabetes. These techniques included Logistic Regression, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees, and the Farthest First algorithm. Early prediction is essential for both the well-being of individuals and the well-being of society. In comparison to these, the algorithm known as Farthest First displayed the best level of accuracy, highlighting the efficiency of this algorithm in further enhancing predicted results.

The year 2022: Khalid Mehboob The high levels of blood sugar that are linked with diabetes may result in serious problems such as blindness, renal failure, nerve damage, cardiovascular disease, and damage to the vascular system. This research sheds light on the significance of early prognosis in terms of efficient management of healthcare situations. The Random Forest, AdaBoost, and Bagging approaches were examined in this research when ensemble learning techniques were used. These techniques aggregate numerous models in order to enhance prediction accuracy. Among the ensemble models, the Random Forest Ensemble technique was found to have attained the best accuracy, which was 97%. This method outperformed the other ensemble models.

Archit Sharma, the year 2022: During the COVID-19 pandemic, diabetes mellitus appeared as a disorder that was associated with a high risk of death and a fatality rate that was considerable. As the country with the highest number of diabetes cases in the globe, India is known as the diabetes capital of the world. With a particular emphasis on logistic regression classifiers, this study highlights the importance of diabetes prediction and management performed in a timely manner. Using the PIMA Indian dataset, the research investigated 10 cognitive-learning approaches for categorization. The findings indicated that these methods had an accuracy of eighty percent in predicting diabetes, making them the most successful outcome.

Figure 1: Process Flow

It is possible to evaluate a variety of approaches by making use of the Healthcare Diabetes Dataset, which is obtained from the Kaggle Repository. There are 2,768 instances included in this dataset, and all of them include female patients. The dataset comprises thorough medical information. There are eight numerical characteristics that are included in it, with the result variable showing whether diabetes is present (labeled '1') or absent (labeled '0'). During the analysis, each and every characteristic is taken into consideration.

The features and their descriptions are provided below:

1. **Id**: A unique identifier for each record.
2. **Pregnancies**: The total number of pregnancies experienced.
3. **Glucose**: The concentration of plasma glucose measured over 2 hours during an oral glucose tolerance test.
4. **BloodPressure**: The diastolic blood pressure (measured in mm Hg).
5. **SkinThickness**: The thickness of the triceps skinfold (measured in mm).

6. **Insulin**: The serum insulin level after 2 hours (measured in μU/ml).
7. **BMI**: The body mass index, calculated as weight in kilograms divided by height in meters squared.
8. **DiabetesPedigreeFunction**: A genetic score representing the family history of diabetes.
9. **Age**: The age of the individual, measured in years.

Outcome: Binary classification indicating the presence (1) or absence (0) of diabetes

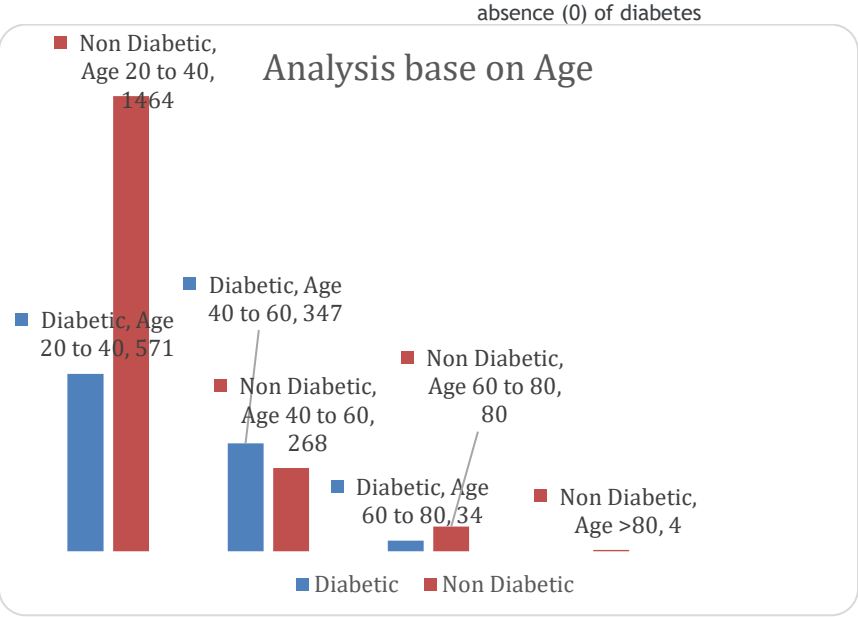

Figure 2: Analysis based on age

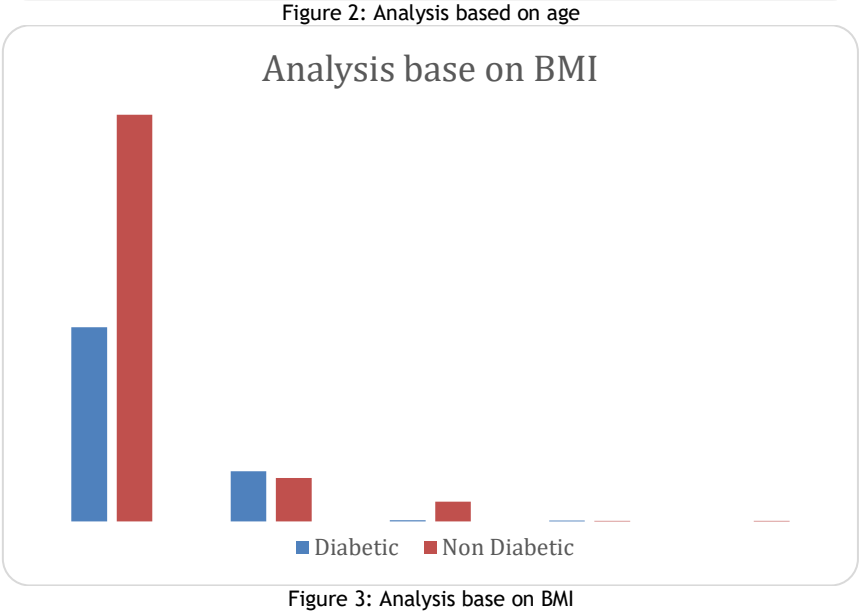

Figure 3: Analysis base on BMI

There is a considerable correlation between the number of pregnancies and the development of gestational diabetes mellitus (GDM), as shown by research conducted by the National Library of Medicine. The risk of getting gestational diabetes mellitus (GDM) is higher in women who have a body mass index (BMI) that is less than 24. A woman's risk of developing diabetes is increased if both of her parents have the illness, and if she has high blood pressure, the possibility of developing diabetes is also increased.

Along with the assessment metrics that are used to evaluate the performance of the model, this part offers an overview of the numerous categorization algorithms that are utilized in the machine learning technique.

Logistic Regression:

In the context of binary classification problems, logistic regression is a straightforward and easily interpretable approach that is often used. Additionally, it does an accurate calculation of the likelihood of a binary result.

For the purpose of diabetes prediction, logistic regression is an extremely helpful tool for constructing a model that establishes a connection between the chance of acquiring diabetes and independent variables such as glucose levels, body mass index (BMI), and age.

```
logi=LogisticRegression(random_state=1)
logi.fit(X_train,y_train)
y_pred_logi=logi.predict(X_test)
print("Logistic regression \n"+classification_report(y_test,y_pred_logi))
con_matrix=confusion_matrix(y_test,y_pred_logi)
con_matrix
accuracy=accuracy_score(y_test,y_pred_logi)
accuracy
```

```
Logistic regression
              precision    recall  f1-score   support

           0       0.73      0.83      0.77       367
           1       0.53      0.39      0.45       187

    accuracy                           0.68       554
   macro avg       0.63      0.61      0.61       554
weighted avg       0.66      0.68      0.66       554
```

0.6787003610108303

Logistic regression has limited ability to handle complex relationships or non-linear data, and its performance may suffer when the data is highly non-linear.

Random Forest:

The Random Forest method is a collaborative learning approach that employs multiple decision trees to make predictions or classifications. Known for its outstanding accuracy and robustness in managing complex datasets, it leverages the combined insights of many trees.

Random Forests are particularly effective in handling a large number of input variables and can manage non-linear relationships within the data. They are adept at capturing complex interactions between features like glucose levels, age, and family history, thereby improving the accuracy of diabetes prediction models.

```
rftree=RandomForestClassifier()
rftree=rftree.fit(X_train,y_train)
rypred=rftree.predict(X_test)
print("random forest\n "+classification_report(y_test,rypred))
accuracy = accuracy_score(y_test, rypred)
print("Accuracy:", accuracy)
```

```
    random forest
                  precision    recall  f1-score   support

               0       0.98      0.99      0.99       367
               1       0.98      0.96      0.97       187

        accuracy                           0.98       554
       macro avg       0.98      0.97      0.98       554
    weighted avg       0.98      0.98      0.98       554

    Accuracy: 0.98014440433213
```

While Random Forest is resistant to overfitting, it can be more complex and challenging to implement compared to the simpler logistic regression.

Naïve Bayes:

Naive Bayes is a probabilistic algorithm that excels in text classification tasks. Based on Bayes' theorem, it assumes conditional independence between features, which allows it to efficiently analyze and categorize text data, making it a useful tool for various classification problems.

In diabetes prediction, Naive Bayes can be effective if the assumption of feature independence is roughly accurate.

```
classifier = GaussianNB()
classifier.fit(X_train,y_train)
y_pred_NB=classifier.predict(X_test)
print("Naive Byes \n"+classification_report(y_test,y_pred_NB))
print("Accuracy is:",accuracy_score(classifier.predict(X_test),y_test))
```

```
Naive Byes
              precision    recall  f1-score   support

           0       0.81      0.86      0.83       367
           1       0.68      0.59      0.63       187

    accuracy                           0.77       554
   macro avg       0.74      0.72      0.73       554
weighted avg       0.76      0.77      0.76       554

Accuracy is: 0.7671480144404332
```

Although Naive Bayes is simple and computationally efficient, it assumes feature independence, which may not reflect real-world scenarios.

Results:

Several supervised machine learning algorithms were applied to predict diabetes using the available dataset, which includes the 'outcome' feature. Since the dataset contains both input and output variables, the models were trained using supervised learning techniques.

The performance of each algorithm was assessed using various evaluation metrics, as outlined below.

Precision:(true positives / (true positives + false positives)).

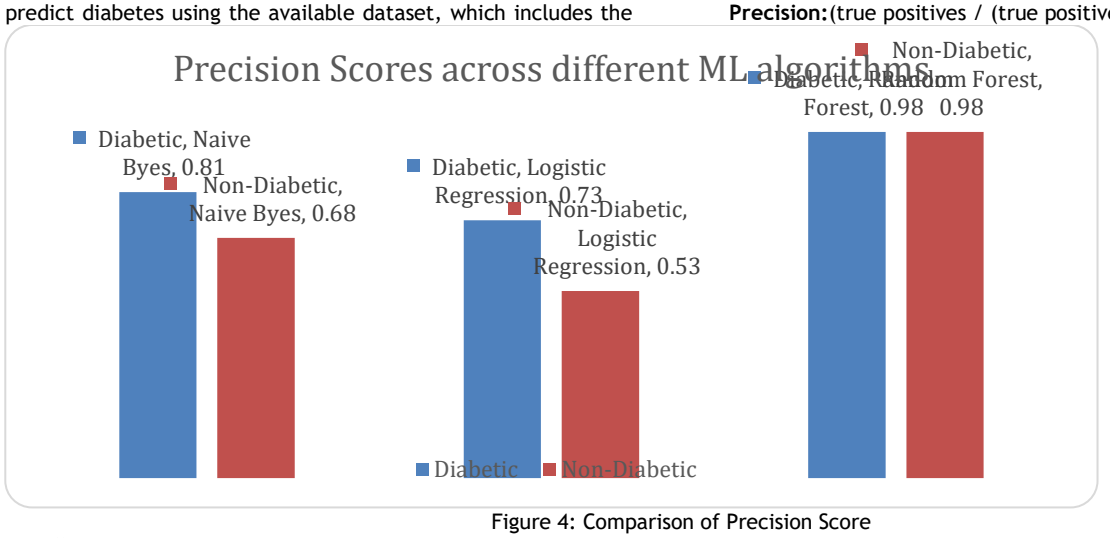

Figure 4: Comparison of Precision Score

Recall:(true positives / (true positives + false negatives))

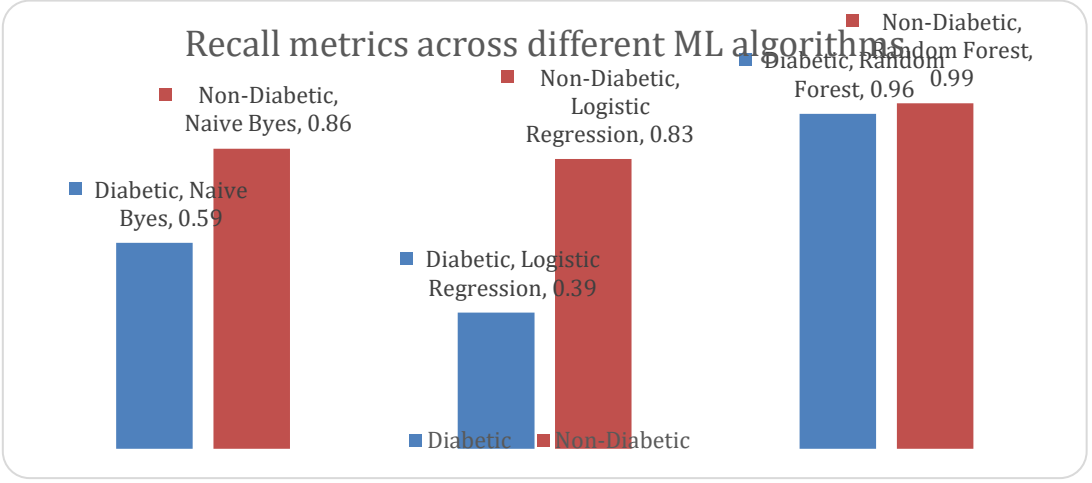

Figure 5: Comparison of metric Recall

F1Score:2 * ((precision * recall) / (precision + recall)).

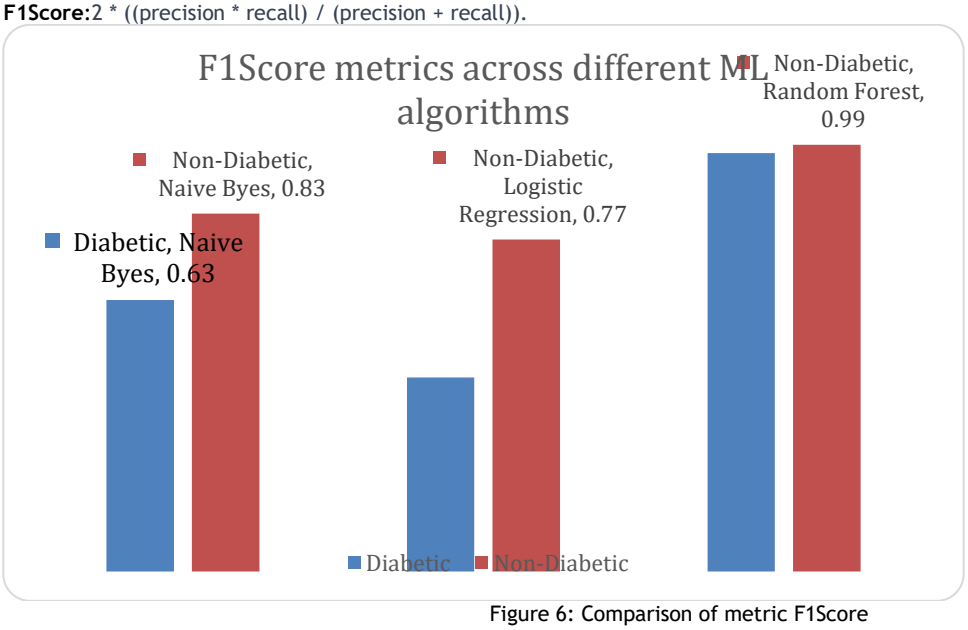

Figure 6: Comparison of metric F1Score

Algorithm	Accuracy
Naive Byes	0.77
Logistic Regression	0.68
Random Forest	0.98

Table 1: Algorithm and their Accuracy

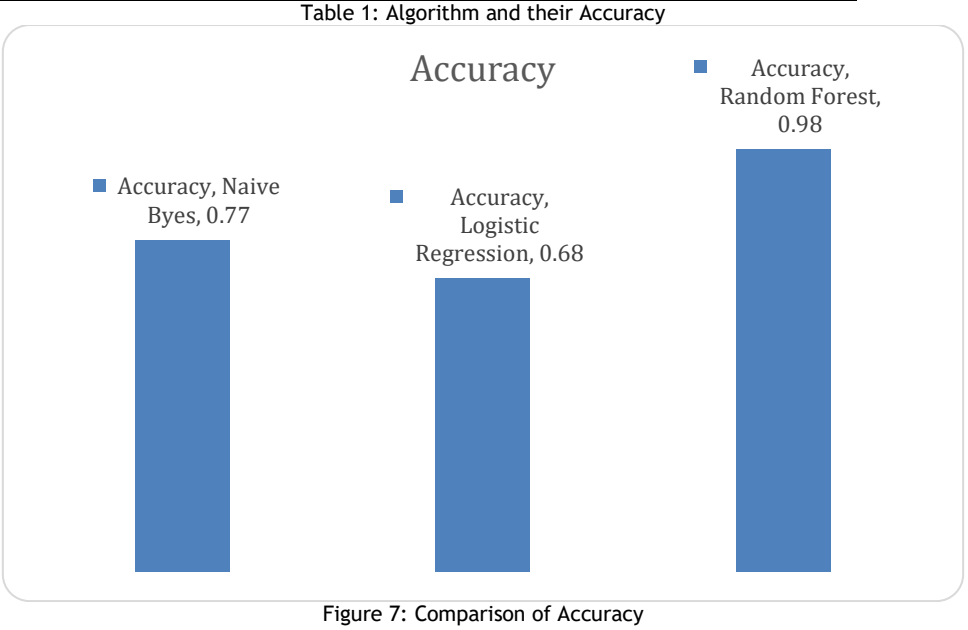

Figure 7: Comparison of Accuracy

Findings:
Using an 80-20 split for training and testing data from the dataset, the Random Forest Decision Tree (RFDT) method achieved an outstanding accuracy of 98%. It outperformed other classification algorithms, showcasing exceptional performance in predicting diabetes.
By using this approach, early diabetes detection becomes possible, allowing for proactive steps to reduce the complications linked to the condition. The application of advanced computational algorithms could further improve the accuracy of diabetes prediction.

CONCLUSION

In diabetes prediction, where the relationships between various health factors and the likelihood of diabetes are complex and non-linear, the Random Forest ensemble method proves to be highly effective. Its ability to handle intricate feature interactions, reduce overfitting, and deliver superior predictive accuracy makes it a preferred choice over simpler decision trees.
Furthermore, the focus on precision in the evaluation metrics highlights the strength of the Random Forest algorithm. In the realm of preclinical diabetes prediction, this machine-learning method shows considerable potential. Especially in rural areas,

where individuals may delay seeking medical help until symptoms worsen, this approach could be particularly valuable. By utilizing autonomous learning techniques, early diagnoses by healthcare providers become possible, potentially preventing the progression of health issues in their early stages.

REFERENCES

- **Comparing Machine Learning Algorithms for Diabetes Prediction** - This paper discusses the use of several algorithms, including Random Forest and Logistic Regression, to predict diabetes, highlighting their effectiveness in managing complex medical datasets. Available in the *Journal of Medical Systems* and other related healthcare research journals. Link: MDPI Article on Machine Learning in Healthcare MDPI
- **Syahri et al. (2024)** conducted a study comparing the performance of Logistic Regression, Random Forest, and AdaBoost for diabetes classification. Their findings suggest that the Random Forest algorithm outperforms the other two in terms of accuracy for diabetes prediction https://pdfs.semanticscholar.org/b857/e1fcdc3b896e5e59c5511e24ab531254c7c4.pdf
- **Machine Learning Approaches for Diabetes Prediction: A Comparative Study** - This study compares the performance of different machine learning algorithms, including Random Forest, Naive Bayes, and Logistic Regression, using real-world healthcare datasets. Available in: ResearchGate: Machine Learning for DiabetesBright Journal
- **Using Machine Learning for Diabetes Prediction** - This paper applies various algorithms such as Naive Bayes, Random Forest, and Logistic Regression to datasets related to diabetes prediction, with results showing varying performance based on the algorithm. Link: ResearchGate: Predictive Analysis Using MLMDPI

.