ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/04.2024-59689752/IJAEAST.2024.01.0005, January 2024, Volume-2, Issue-1, pp.43-48
International Journal of Advances in Engineering Architecture Science and Technology

Diabetes Prediction Using Random Forest

K. Divya Kalyani[1], P. Sirisha[2], A. Taniya[3*], N. Rahi[4], S. Venkat Sai[5]

[1,2,3,4,5]Dadi Institute of Engineering and Technology, Andhra Pradesh, India

divyakalyanik@diet.edu.in[1], sirishapentakota123@gmail.com[2], taniyaabbineni1428@gmail.com[3], rahinakka@gmail.com[4], sundrapusai16@gmail.com[5]

Received: 03-01-2024 **Accepted:** 29-01-2024 **Published:** 31-01-2024

Abstract

Background: Diabetes is a major health problem that affects many people around the world.
Objectives: To solve this problem, we use data from Kaggle, a popular data-sharing site.
Methods: We use computer intelligence called a random forest algorithm to test for diabetes based on age, diabetes level and other factors.
Statistical Analysis: Our random forest model has proven to be very effective at accounting for complex patterns in data, helping us accurately predict whether a person will develop diabetes.
Findings: Our method stands out because we focus on factual accuracy, which makes it different from other methods.
Applications and Improvements: This study, part of the growing use of computers in healthcare, shows that random forest tools may be a reliable and easy way to diagnose early diabetes.

Keywords: Machine learning, Diabetes, Disease Prediction, Health care, Random Forest.

1. Introduction

Diabetes is a chronic disease that can cause health problems worldwide. Based on the International Diabetes Federation, 38.2 crore people worldwide have diabetes. In 2035, this number will double to 59.2 crore. Diabetes is a disease caused by high glucose level. Insulin dependence and diabetes are the most common diseases, but other diseases occur during pregnancy, such as gestational diabetes and others. Machine learning is an emerging field in information science that studies how machines learn from experience. Disease management can be prevented more accurately in patients by using random forests.
Diabetes is a growing disease among people, even among young people. This is a chronic (long-term) health condition that affects how your body converts food into energy. Your body converts most of the food you eat into sugar (glucose) and releases it into your bloodstream. pancreas is the source for insulin. Insulin is like a key that allows sugar to enter the body cells for energy.

Types Of Diabetes

Insulin-dependent diabetes: Diabetes occurs when the body becomes weak and cells cannot produce enough insulin. No studies are proving what causes this type of diabetes, and there is currently no way to prevent it.

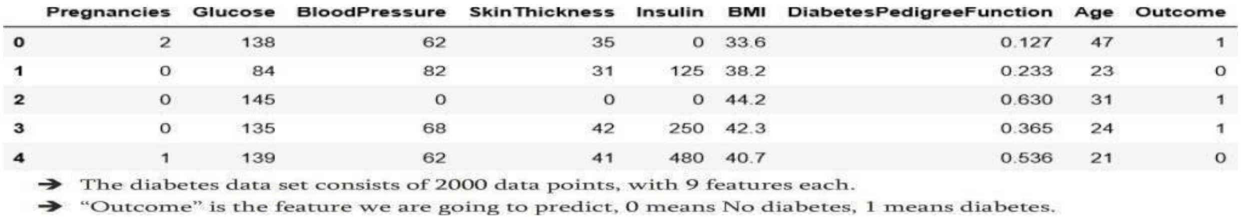
Diabetes: Diabetes is when cells produce less insulin or the body cannot use insulin as it should. This is the most common type of diabetes, affecting 90% of people with diabetes. It is due to genetics and lifestyle.
Gestational diabetes: Occurs when diabetes suddenly occurs in a pregnant woman. The incidence of diabetes is high after pregnancies affected by gestational diabetes.

2. Literature Survey

Diabetes Forecast Introduction: Diabetes is a global health problem affecting millions of people worldwide. Machine learning (ML) holds promise for early prediction of diabetes.
Zheng et al. (2018): “Deep learning methods to predict diabetes.”
Rajput et al. (2019): "Integrated Learning for Diabetes Prediction"
Smith et al. (2020); lifestyle. Advances such as normalization and imputation improve model performance.
Challenges and Future Recommendations: Limited datasets limit global modelling. wide Interpretation and interpretation of ML models in healthcare. The focus of the future is on the integration of real-time data and continuous monitoring.
Machine learning has great potential in predicting diabetes. Ongoing research aims to resolve issues and improve the robustness of the model. Collaboration between doctors and technologists is the key to success.

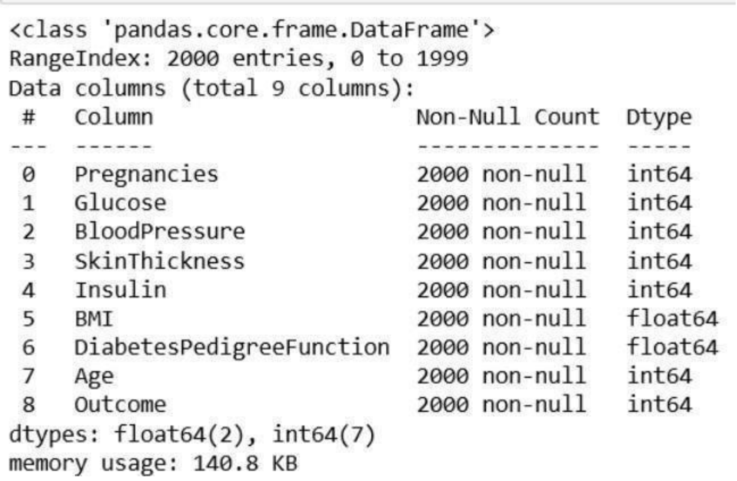
3. Methodology

In this project, we use the random forest algorithm to predict diabetes. Next, data collection is important and data with relevant characteristics and objective variables that indicate the presence of diabetes should be compiled. Data analysis and preprocessing are then important, including analysing the structure of the data set, resolving missing values, and outliers, and coding categorical variables accordingly.
The next step is specific selection, which often uses techniques such as factor analysis or correlation analysis to identify the most important features that make blood pressure estimates sweet. To evaluate the performance of the model, the data set is divided into the training set and the test set. The random forest algorithm was chosen as the prediction model because it is effective in cluster learning by combining multiple decision trees. The model is then trained on the training data and then evaluated using metrics such as the accuracy of the test data.
Cross-validation was performed to assess the model's ability to discriminate data to ensure robustness. Model interpretation is an important step in understanding the importance of different features in predicting diabetes. For those considering deployment, implementing the model in the application or system is an option, with complete documentation of the entire process. Along with continuity and improvement, ethical considerations, especially regarding health forecasts, should also be an important part of the project process.
Dataset Description: The diabetes data set was originated from https://www.kaggle.com/johndasilva/diabetes. Diabetes dataset containing 2000 cases.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

- ➔ The diabetes data set consists of 2000 data points, with 9 features each.
- ➔ “Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Figure 1. Diabetes Data Set

Bar Plot for Outcome Class

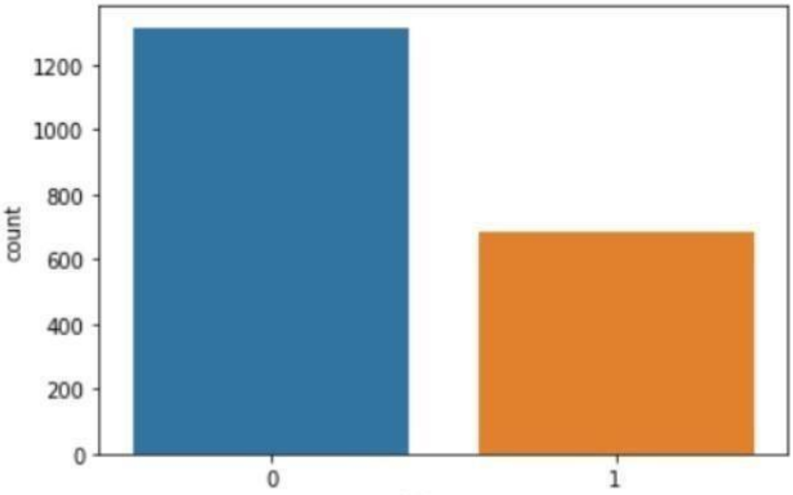


Figure 2. Outcome

Now let's take a look at these plans. It shows how each attribute and tag is broken down into different parts, further demonstrating the need for measurement. Then, wherever you see individual bars, that means each bar is a categorical variable. Before using machine learning, we need to deal with categorical variables. Our score has two categories; 0 means no infection, 1 means infection.

The above graph shows if data points are zero then it is non diabetic. The number of nondiabetics is almost twice the number of diabetic patients.

k-Nearest Neighbor: It can be said that the KNN algorithm is the simplest machine learning algorithm. Modelling only stores training information. To predict the new data, the algorithm looks for the closest data point in the data, its “nearest neighbor.”

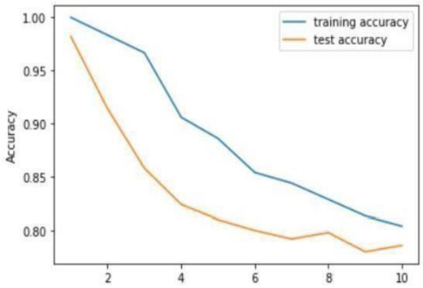


Figure 3. Graph Nearest Neighbor

The figure above shows the training process and testing accuracy on the y- axis compared to n neighbors set on the x-axis.

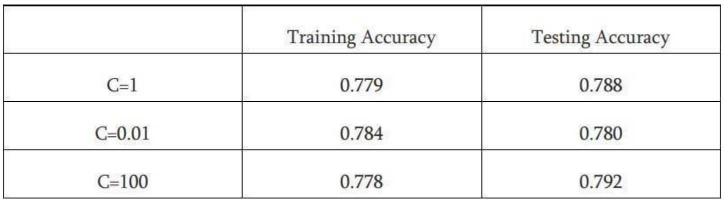
Table 1. Accuracy in k-Nearest Neighbor

Training Accuracy	0.81
Testing Accuracy	0.78

Imagine that the prediction for training will be perfect if we choose the nearest neighbor. However, training decreases as more neighbors are considered, indicating that using a single nearest neighbor makes the model too complex. Best performance is around 9 neighbors.

Logistic regression: Logistic Regression is one of the most common classification algorithms

Table 2. Accuracy in Logistic Regression



	Training Accuracy	Testing Accuracy
C=1	0.779	0.788
C=0.01	0.784	0.780
C=100	0.778	0.792

In the first column, the preset value C = 1 gives 77% accuracy of the training set and 78% accuracy of the test. - Using C = 0.01 in the second-row results in an accuracy of 78% in both training and testing. - Using C = 100 leads to slightly more accuracy of the training process and slightly more accuracy of the testing process; It must be acknowledged that regular and more samples will not improve the value preset. Therefore, we must choose the default value C = 1.

Decision Tree: This classifier creates a decision tree and assigns ranking values to all data points along the tree. Here we can change the maximum number that should be considered when creating the model.

Table 3. Accuracy in k-Nearest Neighbor

Training Accuracy	1
Testing Accuracy	0.99

Feature Importance in Decision Trees: Feature importance measures the importance of each feature for the decision tree. Each attribute is a number between 0 and 1; 0 means "never used" and 1 means estimated.

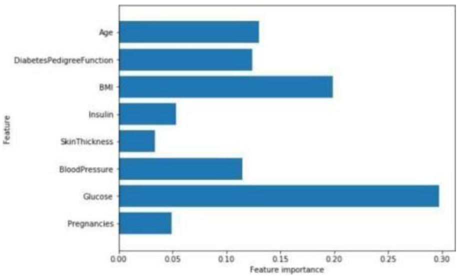


Figure 4. Decision Tree Features

Proposed System Architecture:

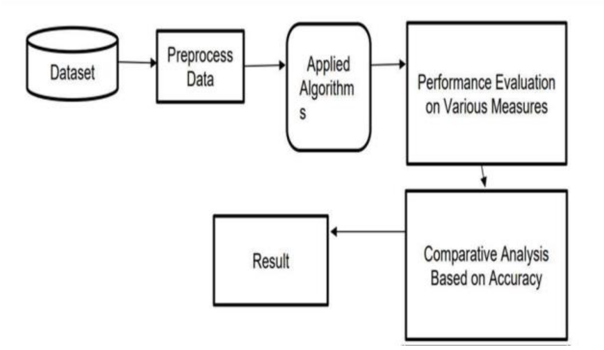

Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

Figure 5. Architecture

Random Forest: This classification takes the concept of decision trees to the next level. It creates a forest of trees where each tree consists of randomly selected features from all the features.

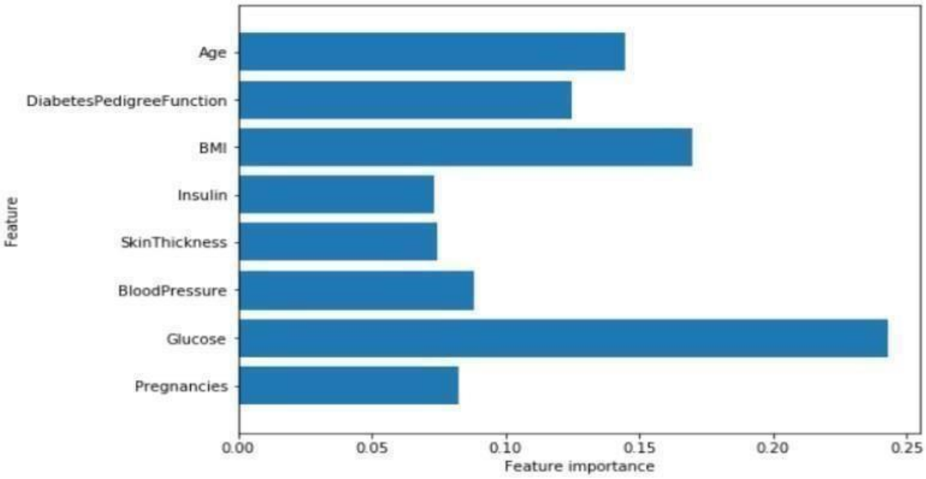


Figure 6. Feature Importance of Random Forest

Random Forest is a powerful machine learning algorithm known for its efficiency and robustness in classification and propagation. Each tree was trained on a subset of the dataset using randomly selected features. In the prediction phase, the algorithm aggregates the predictions of each tree through voting (for classification) or averaging (for regression) to make a final decision. This combination helps reduce the workload and increase the capacity of the model. Random forests are particularly good for processing large datasets that contain long and noisy data. Its ability to handle missing values and maintain accuracy even across unequal data sets makes it a popular choice in many fields such as finance, healthcare, and business.

One of the advantages of random forests is their ability to provide valuable information. By assessing how much each factor contributes to each tree's prediction accuracy, analysts can understand underlying patterns in the data. Additionally, compared to other complex algorithms, random forest requires little hyperparameter tuning, making them easier to use and deploy. However, its interpretation may be limited compared to simple models such as decision trees. But random forests still remain the first choice of many data scientists due to their robustness, scalability, and ability to provide good predictions across different layers in nature.

4. Conclusion

In conclusion, the random forest algorithm in machine learning for diabetes prediction provides a powerful and scalable method. It provides a reliable model for identifying potential diabetes through a combination of learning and critical analysis. The underlying robustness of this algorithm reduces the risk of overfitting and ensures fit with new data. Its scalability makes it efficient to process larger data sets important for medical applications. Additionally, the minimal hyperparameter tuning required by random forests makes it a practical choice. Overall, it is becoming an important tool for early detection and effective management of diabetes as well as improving public health outcome.

References

1. Liu, J., Zhang, S., & Sun, C. (2018). A Novel Random Forest Based Approach for Diabetes Prediction. (pp. 3833-3837). IEEE.
2. Al-Fuqaha, A., Javed, H., Guizani, M., & Rayes, A. (2018). Machine Learning for IoT Big Data and Streaming Analytics: A Survey. & Tutorials, 20(4), 2923-2960.
3. Tave, G. M., & Zhang, S. (2019). A predictive analysis of the USA diabetes population using random forest. Informatics in Medicine Unlocked, 16, 100203.
4. Patel, D. A., Patel, A. A., & Pandya, S. N. (2016). Prediction of Diabetes using Random Forest Algorithm. International Journal of Computer Applications, 135(10), 23-26.
5. Kumar, P. B., & Srinivas, K. (2018). Diabetes Prediction using Random Forest Classification Algorithm. Procedia Computer Science, 133, 157- 162.