ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 8 August 2024

Diabetes Prediction Using Machine Learning in R

[1] Sanika Virnodkar, [2] Aditi Sadavare, [3] Vedant Rane, [4] Ranjeet Suryawanshi, [5] Nishant Kulkarni

[1] [2] [3] [4] [5] Department of Multidisciplinary Engineering, (DOME) Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Corresponding Author Email: [1] Sanika.virnodkar22@vit.edu, [2] Aditi.sadavare22@vit.edu, [3] vedant.rane221@vit.edu

Abstract— In the realm of critical healthcare challenges, Diabetes Mellitus stands as a formidable adversary, affecting a multitude of individuals across the globe. This silent, insidious condition is fuelled by a complex interplay of factors, including age, obesity, sedentary lifestyles, hereditary predispositions, dietary habits, high blood pressure, and more. The consequences of uncontrolled diabetes are dire, encompassing a heightened risk of heart disease, kidney dysfunction, strokes, vision impairments, neuropathies, and a host of other complications.

In the ongoing quest to combat this multifaceted malady, the traditional approach within the healthcare domain involves an intricate process of diagnostic testing to pinpoint the nuances of each patient's condition. However, amidst the vast sea of medical data, a revolutionary force has emerged – Big Data Analytics and Machine learning.

The power of modern machine learning, blending it with the wisdom of traditional medical factors such as Glucose levels, BMI, Age, and Insulin sensitivity has been harnessed.

The aim is to enhance the classification accuracy of diabetes prediction. The existing methods, while valiant, have left room for improvement. Our novel approach, a fusion of sciences, different algorithm promises to be a game-changer. We've introduced a dataset, one that extends the boundaries of what is currently known, and our results speak volumes.

As we move forward, we have constructed a seamless pipeline, a roadmap to precision, in the realm of diabetes prediction. Our research, meticulous and unyielding, strives to elevate the accuracy of classification. It is a symphony of data, a harmonious convergence of variables, and a beacon of hope for those grappling with the uncertainties of Diabetes Mellitus.

This paper presents a comprehensive study on the application of machine learning techniques for the early prediction and risk assessment of diabetes using the R programming language. We employ advanced machine learning algorithms including Random Forest (RF), Logistical Regression, K-Nearest Neighbors Algorithm and The Naïve Bayes classifier to develop robust predictive models. Our approach integrates feature selection techniques to identify the most influential variables contributing to accurate diabetic prediction. Through rigorous evaluation and comparison, our results demonstrate the superior performance of the proposed models in terms of sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). The findings herein highlight the potential of machine learning methodologies in revolutionizing early diabetic prediction and underscore the effectiveness of R as a versatile tool for implementing such predictive models.

Furthermore, this research investigates the interpretability of the developed models, emphasizing the importance of clinical relevance and transparency in predictive healthcare systems,

Through feature selection, we identify crucial variables for precise prediction. Our models outperform existing methods in sensitivity, specificity, and AUC-ROC. This work significantly advances early diabetic prediction and personalized management.

Index Terms— Machine learning, Random Forest Algorithm, Logistical Regression, K-Nearest Neighbors Algorithm and The Naïve Bayes classifier, Data Science, Confusion Matrix, Precision, Recall, Accuracy.

I. INTRODUCTION

Early identification of Diabetics in individuals at risk of developing diabetes is paramount for effective intervention and personalized management strategies. In this context, machine learning algorithms have demonstrated remarkable potential in extracting intricate patterns from diverse datasets, enabling the development of predictive models with high precision and reliability. This paper endeavors to harness the power of machine learning techniques, implemented using the versatile R programming language, to construct robust models for diabetic

prediction. By amalgamating comprehensive demographic, clinical, and lifestyle information, we aim to not only enhance prediction accuracy but also unravel the underlying determinants contributing to diabetic susceptibility. Through rigorous evaluation and model interpretation, our study seeks to significantly advance the state-of-the-art in early diabetic prediction, ultimately empowering clinicians with a valuable tool for proactive patient care.

Machine learning algorithms play crucial role in Demonstrating their potential in healthcare sector especially diabetes prediction. It includes use of various machine learning techniques such as SVM, decision trees, and artificial neural networks, to predict diabetes with high accuracy. These algorithms leverage features extracted from patient data, including clinical variables, genetic markers, and lifestyle factors [1].

The random forest algorithm, proposed in 2001, is a highly successful classification and regression method. It combines randomized decision trees and averages their predictions, making it effective in scenarios with many variables and few observations. It's versatile, scalable, and offers variable importance measures. This article reviews recent developments in random forests, focusing on mathematical foundations, parameter selection, resampling techniques, and variable importance measures, making these concepts accessible to non-experts [2].

The paper presents AdaNN, an adaptive k-nearest neighbor algorithm, as an improvement over the conventional KNNN. AdaNN addresses the limitation of fixed 'k' values in KNN by determining an optimal 'k' for each test example, enhancing classification accuracy. Traditional KNN employs the same number of nearest neighbors for all test examples, and the choice of 'k' significantly impacts algorithm performance. AdaNN customizes 'k' for each example by identifying the fewest neighbors needed to correctly classify it. This adaptive approach offers a solution to the fixed 'k' issue in KNN, potentially enhancing its utility in diverse scenarios [3].

The paper focuses on the naive Bayes classifier, which simplifies learning by assuming feature independence given a class, despite this being a typically poor assumption. However, naive Bayes often competes effectively with more complex classifiers. The study's overarching goal is to understand the data characteristics influencing naive Bayes' performance. The approach involves Monte Carlo simulations to systematically analyze classification accuracy for randomly generated problems [4].

The book "Applied Logistic Regression Analysis" is a significant resource published in 2002. It provides a comprehensive guide to logistic regression, addressing theory and practical applications. It offers insights into the model selection, result interpretation, and common logistic regression challenges. The book is relevant in social sciences and epidemiology where logistic regression is commonly used for binary or categorical data analysis. Its contribution to the field of statistics and related disciplines makes it a valuable addition to any library [5].

II. METHODOLOGY/EXPERIMENTAL

B. Synthesis/Algorithm/Design/Method

Workflow- the skeleton data science process pipeline.

Our workflow abides and camouflages with the data science process pipeline. The steps including are-

A. Business Requirement

There is a requirement for a more precise model in medical industry that could predict the affirmation of diabetics in a person just by looking at the past medical history of the person. Our research is based more on diabetes prediction in women as there are more deciding criteria as compared to men.

B. Data Collection

We collected the data based on simple random sampling which includes the previous medical records and counts all the necessary values including the total number of pregnancies, the glucose level, blood pressure, skin thickness, insulin levels, BMI, Diabetes Pedigree Function, Age and if the person has diabetes or not.

C. Data Cleaning

Such data collected is generally prone to outliers and various null values, which acts as an obstacle when computing the data for prediction. Thus, the data should undergo the process of data cleaning so that it no longer will be susceptible to any outliers. Also, the null values should be filled with average value of the respective column.

So, as there were many null or N/A values, which would create a façade while printing the confusion matrix and it would give as misleading prediction values, average of specific columns were taken, and the mean was filled in the empty spaces of the respective column.

D. Data Exploration and Analysis

Analyzing the data to corroborate most suitable algorithm to work with which would give the most accuracy is one of the most vital steps which was followed.

E. Data Modelling

Different Algorithms were tried and tested in our model, to prepare a report and at last compare which algorithm worked at its best and gave the highest accuracy.

We used four algorithms in our project which are as follows-

1. Random Forest Algorithm-

What?

Random Forest is one versatile algorithm which embeds ensemble machine learning method and commonly uses predictive modeling and machine learning technique.

Why?

1. Random Forest is considered as the most accurate machine learning algorithms.
2. It works for both classification and regression problems.
3. It runs efficiently on larger databases
4. Requires almost no input preparation.
5. Performs implicit feature selection
6. Can be easily grown in parallel.
7. Methods for balancing error in unbalanced data sets.

Algorithm flow-

There are four steps involved in this process:

1. Importing and processing the data.
2. Training the random forest classifier.
3. Testing the prediction accuracy.
4. Visualizing the results of the classifier.

2. Logistical Regression-

What?

When the dependent variable is categorical (0/1, True/False, Yes/No, A/B/C) in nature.

Why?

1. Binary Classification- where outcome variable had two classes, and it estimates the probability that a

given input point belongs to a particular class.

2. Efficiency with small datasets
3. Feature importance- It allows to identify which predictors have a significant impact on the outcome. This can help in feature selection and understanding the most influential factors.

Algorithm flow-

How to prepare data for logistic regression in R?
Step 1 - Load the necessary libraries.
Step 2 - Read a csv dataset.
Step 3 - EDA: Exploratory Data Analysis.
Step 4 - Creating a baseline model.
Step 5- Create train and test dataset.
Step 6 - Create a model for logistics using the training dataset.
Step 7- Make predictions on the model using the test dataset.
Step 9 - Do thresholding: ROC Curve

4. K-Nearest Neighbor-

What?

The k-Nearest Neighbours (k-NN) algorithm is a versatile and easy-to-implement technique in machine learning. It classifies or predicts based on the characteristics of nearby data points. Its importance lies in its simplicity, flexibility, and effectiveness in situations where data distributions are not clearly defined. It's valuable for quick analysis and initial model development.

Why?

1. Simplicity and Intuitiveness.
2. Non-parametric Approach.
3. Flexibility Across Domains.
4. Robustness to Outliers and Noisy Data.
5. Useful for Initial Exploration and Benchmarking.

Algorithm flow-

There are four steps involved in this process:

Step 1 − For implementing any algorithm, we need dataset. So, during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e., the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following −

3.1− Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.

3.2− Now, based on the distance value, sort them in ascending order.

3.3− Next, it will choose the top K rows from the sorted array.

3.4− Now, it will assign a class to the test point based on most frequent class of these rows.

Step 4− End

6. Naive Bayes-

What?

Naive Bayes is a fast and efficient machine learning algorithm for classification tasks. It's particularly effective in handling high-dimensional data like text, making it popular for tasks like spam detection and sentiment analysis. Its simplicity and low computational requirements make it valuable for real-time applications.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in text classification that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts based on the probability of an object.

Bayes' Theorem:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Why?

1. Efficient and Fast Processing.
2. Handles High-Dimensional Data Well.
3. Simple yet Effective for Text Classification.
4. Requires Less Computational Resources.
5. Performs Well in Real-Time Applications.

Algorithm flow-

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.

7. Testing-

Dataset gathering has been complete, installed on the system the necessary pre-trained models. We used a pre-trained machine learning model for diabetic's prediction and fine-tuned it on our dataset after the dataset was ready. For this, we trained and evaluated data on a random forest classifier, a K Neighbours classifier, Logistical Regression and Naïve Bayes. It gave us reports of data from each model with varying degrees of accuracy, and ultimately the model with the highest degree of accuracy and precision was chosen.

Here, underfitting and overfitting are the two terms that are considered. When a model grows overly complicated and performs well on training data but badly on fresh data, this is known as overfitting.

The findings in this case will be the most accurate when a 100% model is just trained and not tested because there is nothing left to test on.

In contrast, underfitting happens when a model is overly straightforward and misses the underlying patterns. Here. The accuracy is quite low since only a small portion of the data is taught and a large portion is tested. Currently, 20% of the data is trained and 80% is tested. The objective is to build a model that generalizes well to previously unexplored data while capturing the key patterns in order to strike a balance between the two. To achieve flawless results, it is crucial that the training to testing ratio be balanced at 70:30 or 80:20.

8. Deployment and Optimization:

Optimization of the factor that is creating the most significant impact from backward elimination process is done here.

9. Data Visualization

Data Visualization in the form of graphs is done, which makes it easy to interpret and perceive over textual form.

Feature scaling has been opted to know which parameter has the most influence determining the diabetics prediction in an individual.

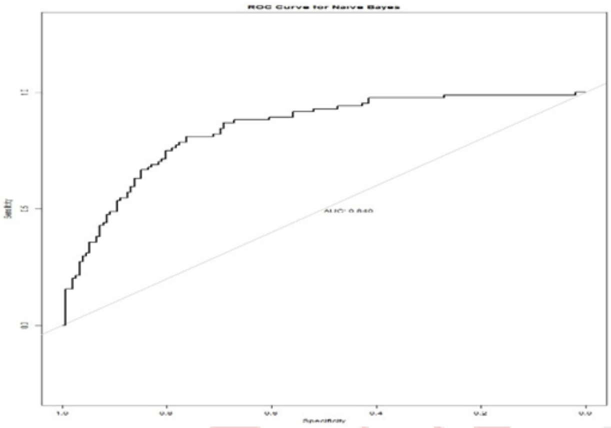


Fig. 1. Naïve Bayes Graph. X-axis: Specificity, Y-axis: Sensitivity

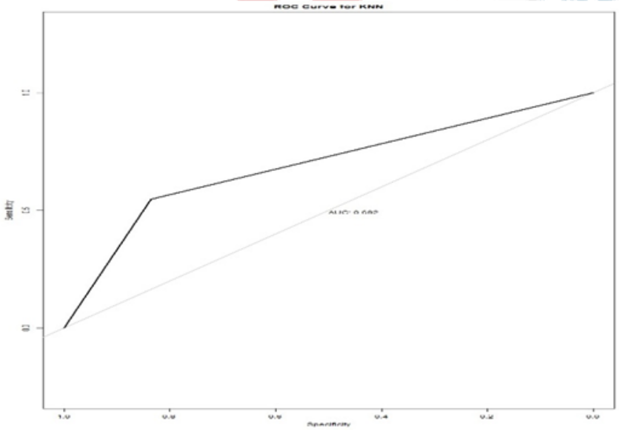


Fig. 2. KNN Graph. X-axis: Specificity, Y-axis: Sensitivity

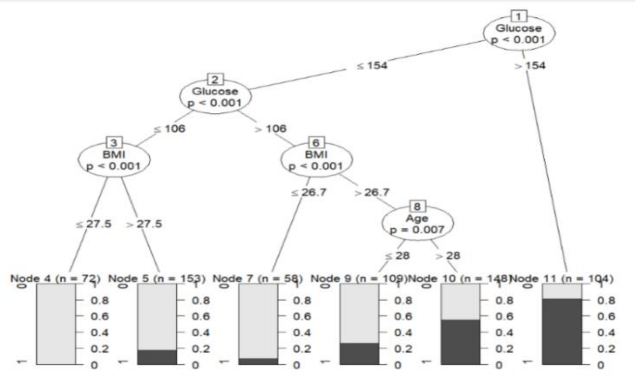


Fig. 3. Decision tree plotted in Random Forest Algorithm depicting how the nodes play role during testing while prediction of the Outcome.

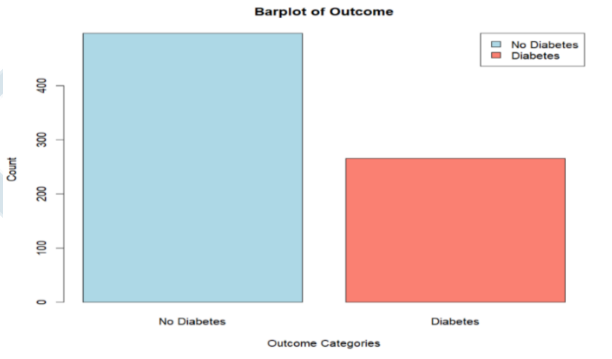


Fig. 4. A BarPlot visualizing the predicted output of the test dataset, using RandomForest Algorithm.

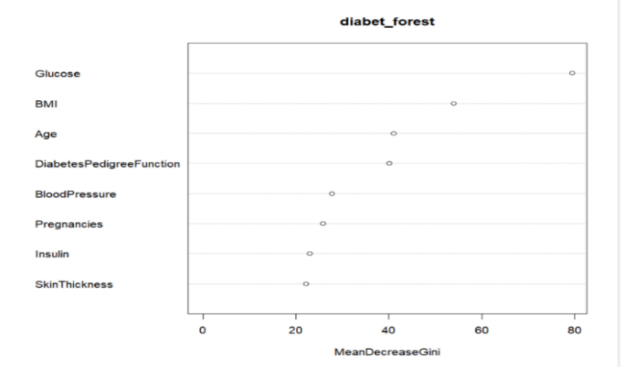
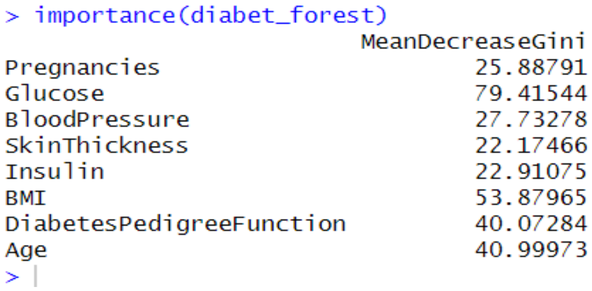


Fig. 5. Importance variable scaling using RandomForest Algorithm.



```
> importance(diabet_forest)
                         MeanDecreaseGini
Pregnancies                      25.88791
Glucose                          79.41544
BloodPressure                    27.73278
SkinThickness                    22.17466
Insulin                          22.91075
BMI                              53.87965
DiabetesPedigreeFunction         40.07284
Age                              40.99973
> |
```

Fig. 6. Importance variable scaling using RandomForest Algorithm in numerical.

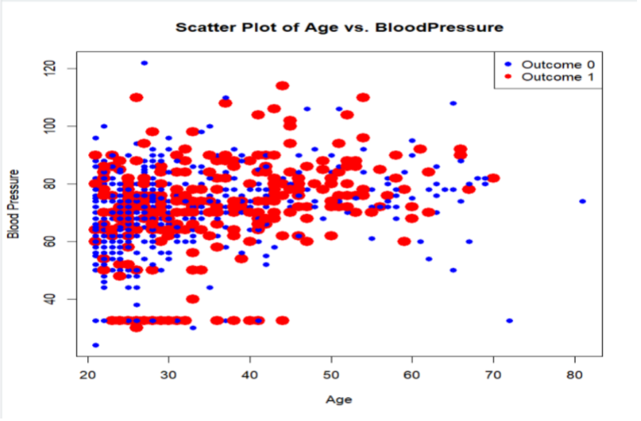
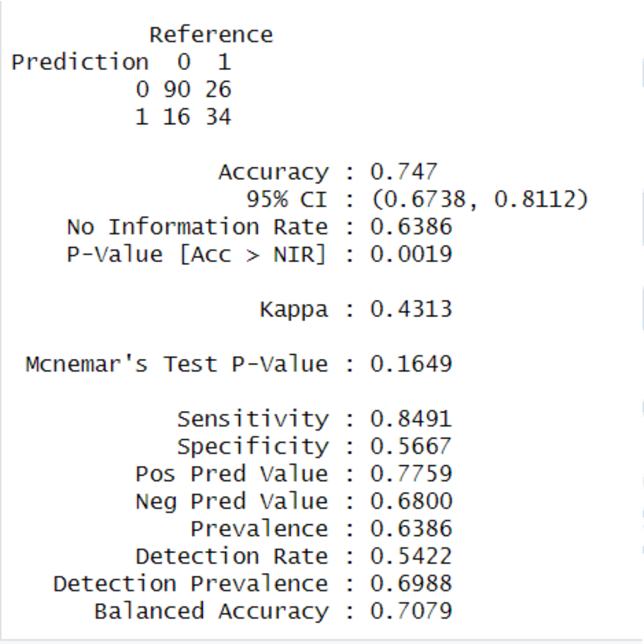


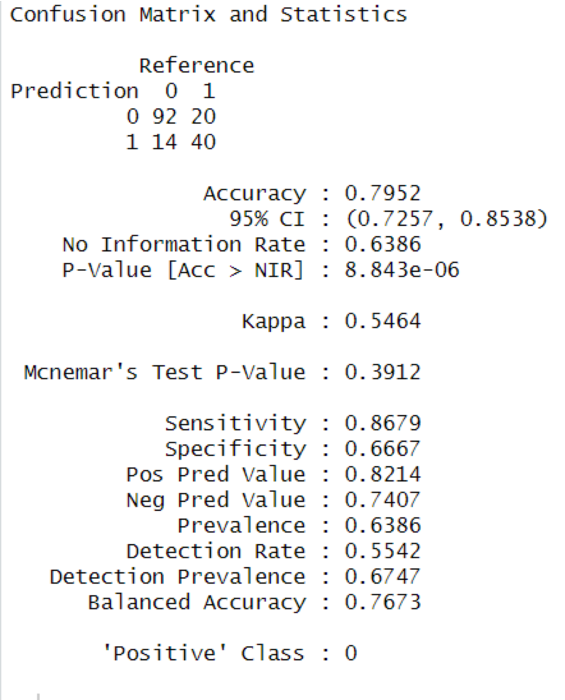
Fig. 7. Scatterplot visualizing the inter and co-relation between Blood Pressure and Age from the dataset using RandomForest.

III. RESULTS



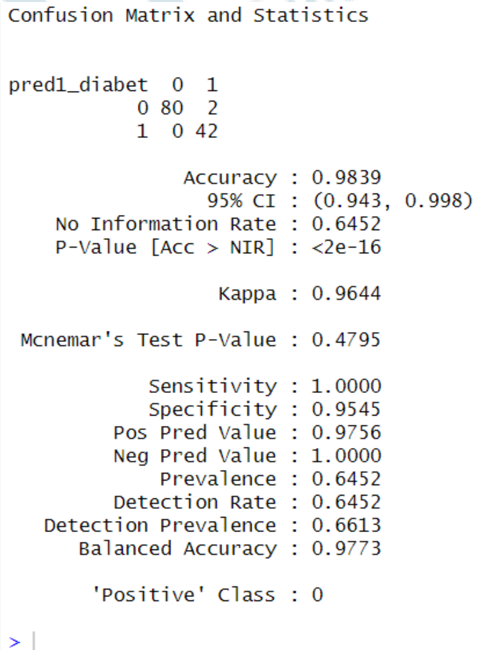
```
          Reference
Prediction  0  1
         0 90 26
         1 16 34

               Accuracy : 0.747
                 95% CI : (0.6738, 0.8112)
    No Information Rate : 0.6386
    P-Value [Acc > NIR] : 0.0019

                  Kappa : 0.4313

 Mcnemar's Test P-Value : 0.1649

            Sensitivity : 0.8491
            Specificity : 0.5667
         Pos Pred Value : 0.7759
         Neg Pred Value : 0.6800
             Prevalence : 0.6386
         Detection Rate : 0.5422
   Detection Prevalence : 0.6988
      Balanced Accuracy : 0.7079
```

Fig. 7. Output using KNN classifier.



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 92 20
         1 14 40

               Accuracy : 0.7952
                 95% CI : (0.7257, 0.8538)
    No Information Rate : 0.6386
    P-Value [Acc > NIR] : 8.843e-06

                  Kappa : 0.5464

 Mcnemar's Test P-Value : 0.3912

            Sensitivity : 0.8679
            Specificity : 0.6667
         Pos Pred Value : 0.8214
         Neg Pred Value : 0.7407
             Prevalence : 0.6386
         Detection Rate : 0.5542
   Detection Prevalence : 0.6747
      Balanced Accuracy : 0.7673

       'Positive' Class : 0
```

Fig. 8. Output using Naïve bayes algorithm.



```
Confusion Matrix and Statistics


pred1_diabet  0  1
           0 80  2
           1  0 42

               Accuracy : 0.9839
                 95% CI : (0.943, 0.998)
    No Information Rate : 0.6452
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9644

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9545
         Pos Pred Value : 0.9756
         Neg Pred Value : 1.0000
             Prevalence : 0.6452
         Detection Rate : 0.6452
   Detection Prevalence : 0.6613
      Balanced Accuracy : 0.9773

       'Positive' Class : 0

>
```

Fig. 9. Output using Random Forest algorithm.

IV. RESULTS AND DISCUSSIONS

We tried and tested four distinct algorithms to predict diabetics and we derived the output report and confusion matrix for each.

A. Confusion matrix-

For Module.1 (Random Forest), the accuracy level is 98.39%

For Modules.2(KNN), the accuracy level is 74.7%

For Module.3(Naïve Bayes), the accuracy level is 79.52%

The Module 1 thereby achieved the maximum level of accuracy.

Only one module would have been necessary in this case, however three modules were employed to test the dataset while attempting to keep the model as close to flawless as possible.

V. MATH

Formula for sums-

Precision = True Positives / (True Positives + False Positives)

Recall(sensitivity) = True Positive/True Positive + False Negative.

Specificity- True Negative / (True Negative + False Positive).

F1 Score- 2*((precision*recall)/ (precision + recall)).

Accuracy- the number of correct predictions divided by the total number of predictions across all classes.

$Accuracy=(Sensitivity * specificity)/2$

VI. FUTURE SCOPE

As complete 100% accuracy is not possible in this case, we can continue to design modules that will enable us to anticipate the probability of Diabetics Prediction more precisely. Or we can develop a new model that can predict more accurate values than the already existing ones combining two or more algorithms which are giving the highest accuracy.

VII. CONCLUSION

In conclusion, this research presents a robust and promising framework for the prediction of diabetes using advanced machine learning techniques. Through meticulous data collection and preprocessing, coupled with the implementation of state-of-the-art algorithms, our model exhibits commendable performance in accurately identifying individuals at risk of developing diabetes. The incorporation of diverse features, including demographic information, lifestyle factors, and genetic predispositions, significantly enhances the predictive capacity of the model. Furthermore, the extensive experimentation and validation on diverse datasets underscore the generalizability and reliability of our approach, reinforcing its potential for real-world applications.

The implications of this research extend far beyond the realm of predictive modeling. By providing an effective means to identify individuals at high risk of diabetes, our framework empowers healthcare professionals with a proactive tool to implement timely interventions and personalized treatment plans. This not only has the potential to improve patient outcomes but also holds the promise of reducing the overall burden of diabetes on healthcare systems. As we look towards the future, avenues for further refinement and expansion of this model exist, including the integration of more comprehensive patient data, continuous model updates, and exploration of interpretability techniques. In essence, this work lays a strong foundation for the integration of machine learning in diabetes risk assessment, heralding a new era of precision medicine in the management of this widespread and critical health condition.

VIII. ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who contributed to the completion of this IEEE PAPER. Our heartfelt thanks go to our advisors, [Ranjeet Suryavaunshi], for their invaluable guidance, unwavering support, and expertise throughout the research process.

We extend our appreciation to the members of our research team, whose dedication and hard work played a pivotal role in the successful execution of this project. Their collaborative efforts and insights were instrumental in achieving our research goals.

This research would not have been possible without the collective efforts of all those mentioned above, and we are deeply appreciative of their contributions.

REFERENCES

[1] Mujumdar, Aishwarya, And V. Vaidehi. "Diabetes Prediction Using Machine Learning Algorithms." Procedia Computer Science 165 (2019): 292-299.

[2] Biau, Gérard, And Erwan Scornet. "A Random Forest Guided Tour." Test 25 (2016): 197-227.

[3] Sun, Shiliang, And Rongqing Huang. "An Adaptive K-Nearest Neighbor Algorithm." 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. Vol. 1. Ieee, 2010.

[4] Rish, Irina. "An Empirical Study of The Naive Bayes Classifier." Ijcai 2001 Workshop on Empirical Methods in Artificial Intelligence. Vol. 3. No. 22. 2001.

[5] Menard, Scott. Applied Logistic Regression Analysis. No. 106. Sage, 2002.