Diabetic Prediction using Random Forest Classifier

R.M. Sanduni Kanishka Chamodani
University of Dalarna
Borlänge, Sweden
e-mail: v23skrmu@du.se

Abstract— **The objective of this study is to use a random forest classifier to predict diabetes. The goal is to create a model that can accurately recognize people who are at risk for diabetes based on a variety of demographic and medical characteristics. Because it can capture complicated interactions between predictors and produce reliable predictions, the random forest method is used. Age, gender, body mass index, hypertension, heart disease, smoking habit, HbA1c level and blood glucose level are the factors in the dataset. The random forest model has been tested and shown to have great accuracy in predicting diabetes. The feature importances draw attention to the important predictors, such as HbA1c level, blood glucose level, age, BMI and smoking history. This study advances the use of machine learning to detect diabetes and offers insights for preventive healthcare interventions and individualized management approaches.**

Keywords-Random Forest; diabetes; analysis; data visualization;

I. INTRODUCTION

A significant section of the world's population suffers from diabetes, a chronic metabolic illness that places a heavy load on healthcare systems everywhere. The efficient management of the disease, the avoidance of complications, and the enhancement of patient outcomes all depend on the early detection and accurate prediction of diabetes. With the capacity to examine complex patterns and connections within big datasets, machine learning approaches have demonstrated considerable promise in this area.

The main objective of this research is to use the random forest algorithm to create a diabetes predicting model. The task of identifying people at risk for diabetes based on multiple medical and demographic factors is what the initiative attempts to address. Healthcare providers can proactively interfere, carry out targeted interventions, and customize treatment plans for specific patients by correctly anticipating diabetes.

The possibility for better healthcare outcomes and a decrease in the financial burden of diabetes make this issue crucial. It is possible to more effectively deploy healthcare funds and put preventative measures in place to slow the spread of the disease by identifying those who are at high risk. A prompt diagnosis can also encourage people to change their lifestyles and seek the right medical attention, which improves health management.

The following research questions will be addressed in this project:

1. Based on a collection of demographic and medical characteristics, can a random forest classifier accurately predict the chance of diabetes?
2. Which factors have the biggest impact on diabetes prediction?
3. How does the performance of the random forest model?

The objective of this research is to create a reliable and accurate diabetes predictive model using the random forest algorithm and to evaluate its performance. Additionally, the study aims to pinpoint the major risk factors for diabetes and reveal the risk variables that contribute to the disease.

II. LITERATURE REVIEW

An active topic of research in recent years has been the prediction of diabetes using machine learning algorithms. With an emphasis on papers released in the recent few years, this literature review seeks to give a brief summary of the state of the art in diabetes prediction using machine learning algorithms.

The significance of feature selection in diabetes prediction has been emphasized in a number of recent research. For instance, a study by [1] employed feature selection methods including recursive feature elimination (RFE).

Several machine learning methods have been used in the recent years to predict diabetes. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two deep learning techniques, have drawn attention for their capacity to automatically identify complex patterns in data. For instance, the study conducted by [2], which used an RNN-based model, successfully predicted diabetes with high accuracy.

The interpretability of machine learning models has received more attention in recent years, particularly in the healthcare industry. Methods to explain model predictions in diabetes prediction tasks have been investigated by researchers. Decision trees and rule-based models are two examples of explainable machine learning techniques that have been used to improve model interpretability.

The evaluated literature provides evidence of the state-of-the-art in machine learning methods for diabetes prediction. It emphasizes the value of feature selection, the effectiveness of various methods, and the emphasis on model interpretability. These results are in line with the project's research aims, which include creating a diabetes prediction model utilizing the high-performance Random Forest algorithm and identifying the crucial factors affecting the prediction.

III. METHOD DESCRIPTION

A. The Dataset

Dataset for diabetes prediction was taken from [3]. The dataset consists of 100000 rows and 9 columns. Columns are as follows:

1. gender: the individual's biological sex, which is divided into the categories of male, female, and other.
2. age
3. hypertension: the blood pressure which has values a 0 or 1 where 0 indicates that they don't have hypertension and for 1 it means they have hypertension.
4. heart_disease: heart disease is a medical condition which has values 0 and 1 where 1 indicates they have heart disease and otherwise not.
5. smoking_history: History of smoking which consists 5 categories such as 'not current', 'former', 'no info', 'current', 'never' and 'ever'.
6. bmi: measure of body fat based on weight and height.
7. HbA1c_level: average blood sugar level during the previous two to three months of the person.
8. blood_glucose_level: amount of glucose in the bloodstream at a given time.
9. diabetes: target variable which has values of 1 and 0 where 1 indicates diabetes positive and otherwise not

For data preprocessing, first null values were checked and there were no null values. Then duplicate rows were dropped. Then label encoding was done for 'gender' to convert it to numerical value. Then data type of 'age' column is converted into int. Since the 'smoking history' consists of six labels, then that variable is converted into numerical format by converting 'no info' into -1, 'never' and 'not current' into 0 and 'current', 'former' and 'ever' into 1.

Then data analysis was done using visualization. Visualizations were done for categorical variables and continuous variables separately. For categorical variables, gender, hyper_tension, heart_disease, smoking_history is visualized using countplot. It is shown in figure 1.

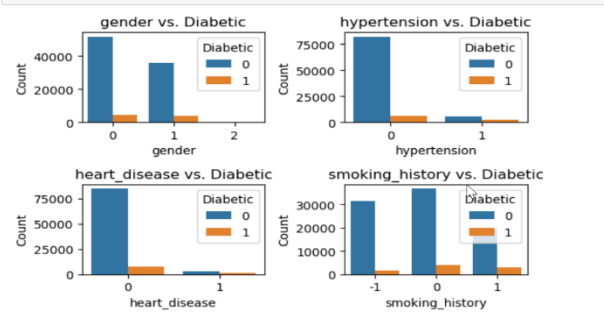

Figure 1: Categorical data analysis

As continuous variables, 'age', 'bmi', 'HbA1c_level' and 'blood_glucose_level' were visualized using histograms and lineplots and it is shown in figure2.

According to the figure 3, it is obvious that blood glucose level abd HbAIc level have direct association with diabetes.
To find out the correlation of variables correlation map was created.

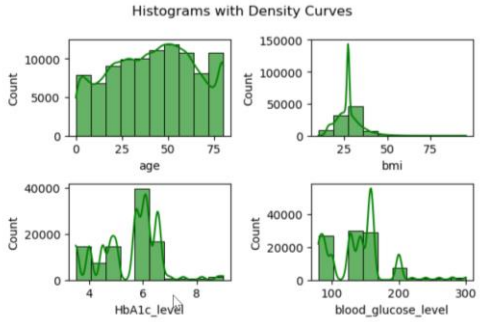

Figure 2: Continuous variables analysis

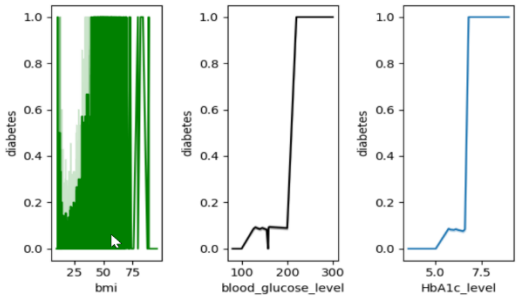

Figure 3: Behaviour of BMI, HbA1c_level and blood glucose level with diabetes

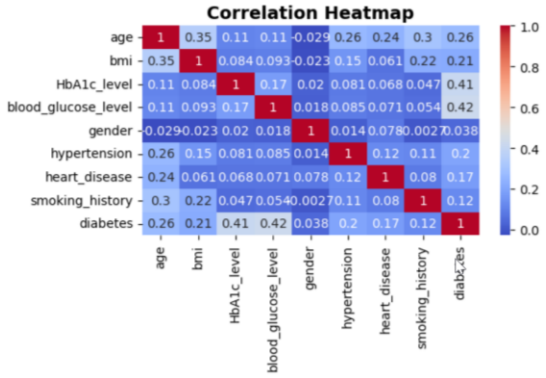

Figure 4: Correlation Heatmap

According to the correlation map of figure 4, 'gender' has the minimum importance towards 'diabetes'. Therefore, that column was removed from the dataset.

B. Data Mining Method

RandomForestClassifier is used as the method to create the model. Before creating the model, values of target variable ('diabetes') are counted and it shows that there are RandomForestClassifier is used as the method to create the model. Before creating the model, values of target variable ('diabetes') are counted and it shows that there are 87664 of non-diabetic records and 8482 of diabetic records. So, it is obvious that dataset is not balanced. Therefore, SMOTE is used to balance the data. First, data is split into testing and training and then SMOTE is used to balance them. Then the model is created using this balanced data.

Then to measure the performance, confusion matrix is created. Also, accuracy_score, mean_squared_error, f1_score and classification_report are calculated. Finally, the factors that have biggest impact on diabetes were selected using the model.

IV. Results and Analysis

The performance of this model is as follows. Accuracy score of this model is 95.42% and the mean squared error is 0.04. The F1 score of this model is 0.74. Those results are shown in figure 5.

```
Accuracy Score:  95.42381695267811

Mean Squared Error:  0.045761830473218926

F1 Score:  0.7443346891342244

Classification Report
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     17509
           1       0.74      0.74      0.74      1721

    accuracy                           0.95     19230
   macro avg       0.86      0.86      0.86     19230
weighted avg       0.95      0.95      0.95     19230
```

Figure 5: Performance of the model

```
Confusion Matrix
 [[17069   440]
 [  440  1281]]
```

Figure 6: Confusion Matrix of the model

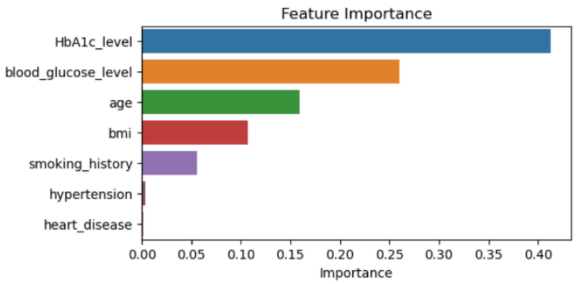

Figure 7: Important features of the model

Since the accuracy is high and MSE is low, it implies that this model is an accurate model. Also, F1 score implies that this model has achieved a relatively good level of performance in terms of both precision and recall. With a score of 0.74, the model has demonstrated good precision and recall performance. It implies that a suitable trade-off between limiting false positives (precision) and minimizing false negatives (recall) has been accomplished by the model.

According to the confusion matrix in figure 6, TN and TP are high than FP and FN. That means number of instances that are correctly predicted as diabetic positive and the number of instances that are correctly predicted as diabetic negative are higher than false predictions.

Then the results that taken from important features from the model are shown in figure 7. According to the figure 7, it implies that, HbA1c level, blood glucose level, age, bmi and smoking history have high impact on diabetes.

V. Conclusion

In this study, we were able to create a Random Forest classifier that accurately predicted diabetes. The model performed successfully, obtaining high levels of accuracy and F1 score. Therefore, it is obvious that a random forest classifier can accurately predict the chance of diabetes based on a collection of demographic and medical characteristics. Also, those scores implies that this model performed well.

We identified important variables, including HbA1c level, blood glucose level, age, bmi and smoking history, that have a substantial impact on the prediction of diabetes by feature importance analysis.

The findings of this study have important implications for the healthcare industry. Accurate diabetes prediction can result in early detection, enabling prompt interventions and individualized treatment protocols. Healthcare resources can

be used more effectively, improving patient outcomes and lowering costs, by identifying high-risk individuals.

More improvements can be discovered, such as adding new features, collecting larger and more varied datasets, and utilizing cutting-edge approaches like hyperparameter tuning or ensemble methods to increase the model's performance.

Overall, this study contributes to the growing body of knowledge with regard to the use of machine learning in the healthcare industry. The discoveries show promise for improved diabetes prediction and eventually patient care.

VI. References

[1] X. Li and J. Zhang, "Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm," *SpringerLink,* vol. 1, no. 1, pp. 153-169, 2021.

[2] B. Ljubic and A. A. Hai, "Predicting complications of diabetes mellitus using advanced machine learning algorithms," *JAMIA,* vol. 27, no. 9, p. 1343–1351, 2020.

[3] M. MUSTAFA, "Diabetes prediction dataset," [Online]. Available: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-datase.