Diabetes Prediction using Machine Learning Techniques Final Report

Team Members:

- Name: **ABDULLAH EID**, ID: 1821221254
- Name: **Muhammed Yahya Avar**, ID: 1921221004
- Name: **Taha Demir**, ID: 1921221354

Introduction

Diabetes is a chronic disease that affects millions of individuals worldwide. Early detection and proper management of diabetes can significantly reduce the risk of mortality and associated complications. In this project, our main objective is to develop a machine learning model application that accurately predicts the likelihood of a patient developing diabetes based on their medical history and demographic information.

To accomplish this, we utilized a dataset comprising health-related information from 100,000 individuals. Using their data as input, we employed three machine learning algorithms: Random Forest Classification, Gaussian Naive Bayes, and Decision Tree. In the data preprocessing phase, we performed data cleaning, feature selection, and normalization to ensure the quality and suitability of the data for our models.

Related Work

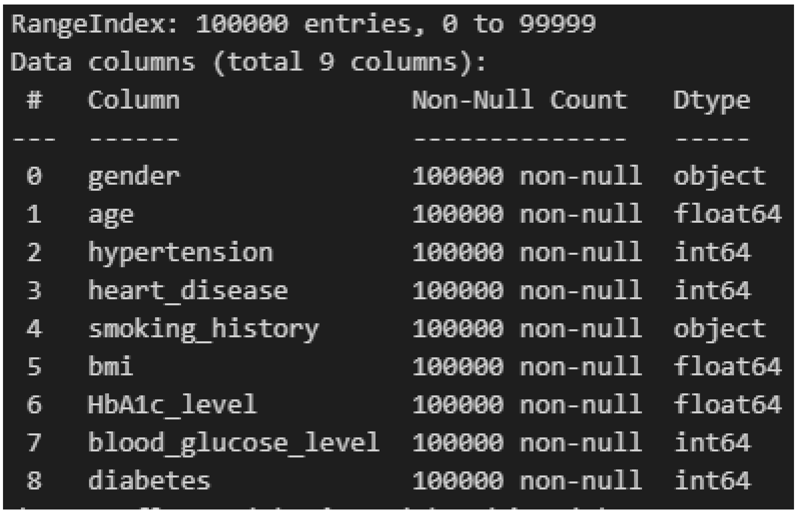
In the field of diabetes prediction, several related works have been explored. Priyadarshinee, S., Panda, M. (2022) conducted a comparative study of machine learning algorithms, highlighting their strengths and weaknesses.N. Nnamoko, A. Hussain and D. England(2018) focused on ensemble learning techniques. S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika(2021) proposed a feature selection framework. Azad, C., Bhushan, B., Sharma (2021) introduced an improved decision tree algorithm for diabetes classification. These approaches demonstrate the use of various algorithms and techniques for diabetes prediction tasks.

Approaches in diabetes prediction also involve the application of deep learning techniques. However, traditional machine learning algorithms such as random forest, support vector machines, and decision trees are still widely used. Machine learning algorithms have proven effective in capturing patterns and relationships within the data.

In our project, we contribute by employing multiple machine learning algorithms for diabetes prediction. By leveraging the neural network's ability to learn complex patterns and nonlinear relationships, we aim to improve prediction accuracy. Additionally, we incorporate data preprocessing techniques, including data cleaning and feature engineering to ensure high-quality and suitable input for our models.

Dataset and Features

Dataset Link: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset



```
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
```

Figure 1- Dataset Features

gender

Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. There are three categories in it male ,female and other.

age

Age is an important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0-80 in our dataset.

hypertension

Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It has values a 0 or 1 where 0 indicates they don’t have hypertension and for 1 it means they have hypertension.

heart_disease

Heart disease is another medical condition that is associated with an increased risk of developing diabetes. It has values a 0 or 1 where 0 indicates they don’t have heart disease and for 1 it means they have heart disease.

smoking_history

Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes. In our dataset we have 5 categories i.e not current, former, No Info, current, never and ever.

bmi

BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese.

HbA1c_level

HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes.

blood_glucose_level

Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes.

diabetes

Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes.

We found out that our data is imbalanced. Afterwards we performed preprocessing in our dataset including handling missing values. We filled the null values with the mode of the respective column. Initially, we also attempted to remove outliers from the data. However, we discovered that removing outliers resulted in the loss of all patients diagnosed with diabetes. So, we decided to retain the outliers in our dataset.

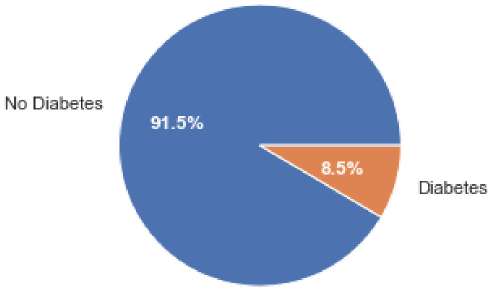


Figure 2- Original Dataset

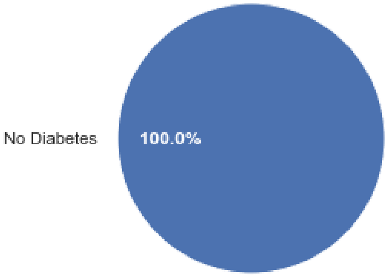


Figure 3- Dataset with outliers removed.

Next, we split the dataset into training and testing sets to evaluate the performance of our models. To standardize the data and bring all features to a similar scale, we employed feature scaling using the StandardScaler.

Methods

We trained our machine learning models of Naive Bayes, Decision Tree, and RandomForest. We also implemented Naive Bayers model from scratch.

- **Naive Bayes Classifier:**

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with an assumption of independence between features. It calculates the probability of a particular class given the input features and assigns the class label with the highest probability. Naive Bayes assumes that all features are conditionally independent of each other, which simplifies the calculation. Despite this simplifying assumption, Naive Bayes can still achieve good performance, especially in text classification and spam filtering tasks.

- **Decision Tree Classifier:**

Decision Tree is a hierarchical tree-based algorithm that uses a flowchart-like structure to make decisions. It starts with a root node representing the entire dataset and recursively splits the data based on feature values, creating branches and nodes. The splits are determined by selecting the feature that provides the most information gain or Gini impurity reduction. Each leaf node represents a class label or a decision. Decision Trees are interpretable, as they can be visualized, and can handle both categorical and numerical features. However, they are prone to overfitting if not properly pruned.

- **Random Forest Classifier:**

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It works by creating a set of decision trees and aggregating their predictions through voting or averaging. Each tree is trained on a random subset of the training data (bootstrapping) and a random subset of features (feature bagging). This randomness helps reduce overfitting and improves generalization. Random Forests are robust, handle high-dimensional data well, and provide feature importance measures. They are widely used in various domains, including healthcare and finance, due to their accuracy and versatility.

Results

We compared our methods’ performance by using F1-score metric. The higher F1-score point to better results.

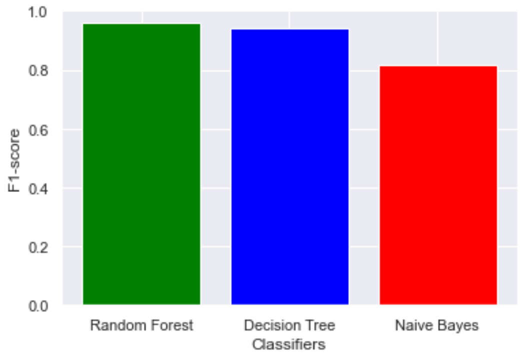


Figure 4- F1 Score comparison

Utilizing the default parameters of the RandomForest classifier yielded promising outcomes. By implementing optimizations and ensuring we steer clear of overfitting, we can attain exceptional results with this model.

Naive Bayes assumes that all features are independent, but it is highly unlikely to find a set of features that are completely independent. This limitation of Naive Bayes can result in suboptimal performance. On the other hand, Random Forest tends to provide better results because it employs decision trees in its algorithm. Decision trees consider the relationships between features and the target variable when constructing the model, which can lead to more accurate predictions.

We also used RandomUnderSampler, Near Miss Under Sampling and Over Sampling by SMOTE to mitigate the impact of class imbalance, allowing our model to better learn from the minority class and improve its performance.

Conclusion

In conclusion, our report focused on the development of a machine learning model for diabetes prediction. We explored the performance of three algorithms: Naive Bayes, Decision Tree, and RandomForest. Among these, the RandomForest algorithm emerged as the highest-performing algorithm. This can be attributed to its approach where it combines multiple decision trees to make predictions through voting or averaging. The aggregation of predictions from individual trees helps improve the overall accuracy and robustness of the model.

In future work, we could investigate the performance of other classification algorithms such as Support Vector Machines, Gradient Boosting, or Deep Learning models. Comparing their performance against the RandomForest algorithm could provide insights into the suitability of different algorithms for diabetes prediction. Additionally, expanding the dataset to include more samples and a wider range of features could further enhance the model's efficiency and generalizability. Moreover, conducting feature engineering and incorporating domain-specific knowledge may also contribute to improving the predictive power of the model.

Contributions:

- Abdullah EID: Model Definition and Training
- Muhammed Yahya AVAR: Preprocessing and Feature Engineering
- Taha DEMIR: Dataset Description and Visualization

References:

1. https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb
2. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6
3. https://towardsdatascience.com/outlier-detection-methods-in-machine-learning-1c8b7cca6cb8
4. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
5. Priyadarshinee, S., Panda, M. (2022). Machine Learning Algorithms for Diabetes Prediction. In: , et al. Innovations in Intelligent Computing and Communication. ICIICC 2022. Communications in Computer and Information Science, vol 1737. Springer, Cham. https://doi.org/10.1007/978-3-031-23233-6_22
6. N. Nnamoko, A. Hussain and D. England, "Predicting Diabetes Onset: An Ensemble Supervised Learning Approach," 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil, 2018, pp. 1-7, doi: 10.1109/CEC.2018.8477663.
7. S. Sivaranjani, S. Ananya, J. Aravinth and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 141-146, doi: 10.1109/ICACCS51430.2021.9441935.
8. Azad, C., Bhushan, B., Sharma, R. et al. Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. Multimedia Systems 28, 1289–1307 (2022). https://doi.org/10.1007/s00530-021-00817-2