Diabetes detection

Sufyan

June 2023

1 Introduction

Diabetes is a chronic condition that effects the way the body processes food into energy. The condition can have life long complications and must be treated. Type two diabetes can appear later in life, thus it is important to be able to predict if someone has diabetes.

Using the dataset found at https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset, In this write-up and accompanying notebook, I look at and compare a few different models that try to predict if someone has diabetes based on other health indicators.

2 Detailos about the dataset

The dataset comprises nine variables, including eight health-related features and one binary classification feature indicating the presence or absence of diabetes. The subsequent sections of this report will delve into these features in greater detail, discussing their relevance and influence on our predictive models. Note that the detailed description of these features is available on the original dataset page.

- **Gender**: Gender represents the differences in biological sex, and can impact the susceptibility of someone to develop diabetes.

- **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 0 or 1.
- **Heart Disease**: Heart disease is another medical condition that is associated with an increased risk of developing diabetes. It has values 0 or 1.
- **Smoking˙history**: Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes.
- **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.
- **HBA1c level**: HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months.
- **Blood glucose level** Blood glucose level refers to the amount of glucose in the bloodstream at a given time.

There are a total of $100,000$ entries

3 Exploratory analysis and data prep

In this section, I will embark on an in-depth exploration of the most influential features of the dataset, elucidating potential correlations and their implications for diabetes prediction. Furthermore, I will outline the data preparation processes implemented prior to model training, including data cleaning and pre-processing.

While our goal is to assess patterns and relationships within our data, it's essential to understand that correlation does not always imply causation. Therefore, any associations I detect should be considered in terms of their potential predictive value for our models, rather than any direct biological cause-effect relationship.

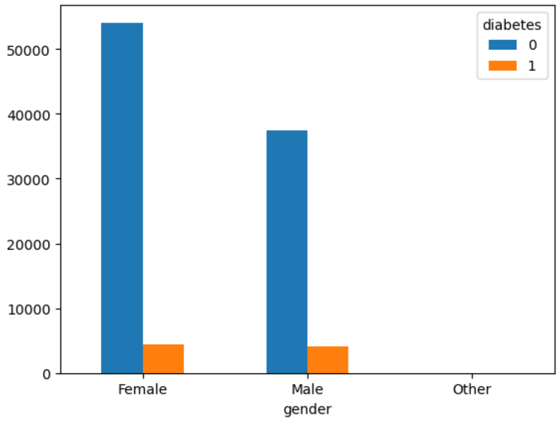
In the data preparation phase, I aim to ensure our data is devoid of any inaccuracies, inconsistencies or missing values that could potentially compromise our model's performance. This step is crucial for ensuring the robustness and reliability of our subsequent predictive analysis. Detailed explanations of these steps will be provided as I navigate through this phase.

3.1 Gender

A cursory review of the dataset, as visualized in Figure 1, indicates a higher prevalence of diabetes among males compared to females. However, this observation is based on the assumption of identically and independently distributed (i.i.d) data, and the absence of any significant sampling bias.

Sampling bias can significantly skew our analysis, as certain subsets of the population might be underrepresented or overrepresented. In this case, it's worth considering potential bias arising from health-seeking behaviors. For instance, if men who are asymptomatic are less likely to get tested for diabetes, the data may show an inflated incidence of diabetes in males. This could affect our model's predictions and interpretations, leading us to incorrect conclusions.

Figure 1: Diabetes against Gender

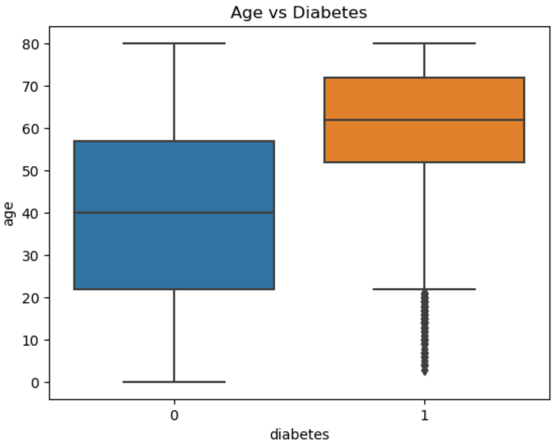


3.2 Age

Age is widely acknowledged as a significant risk factor in the onset of Type II diabetes. In our dataset, as illustrated in Figure 2, this correlation becomes evident: the age distribution of individuals with diabetes is skewed towards older age groups when compared to the non-diabetic population.

This suggests a higher prevalence of diabetes among the elderly, aligning with the commonly accepted understanding of age-related risk in Type II diabetes. It's important to note that while age is a crucial component in our analysis, the nature of this disease is multifactorial, with various other factors potentially contributing to its onset.

Figure 2: Diabetes against Age

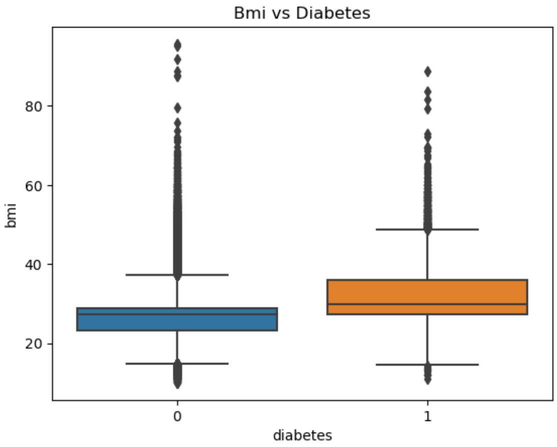


3.3 BMI

Body Mass Index (BMI) is another important factor to consider in the context of diabetes. As we might anticipate, our dataset reveals a higher average BMI amongst individuals diagnosed with diabetes. As seen in figure 3.

This observation aligns with established medical research indicating a strong association between elevated BMI and increased diabetes risk. Overweight or obese individuals tend to have a higher probability of developing Type II diabetes, making BMI a crucial feature in our prediction model.

Figure 3: BMI against Age

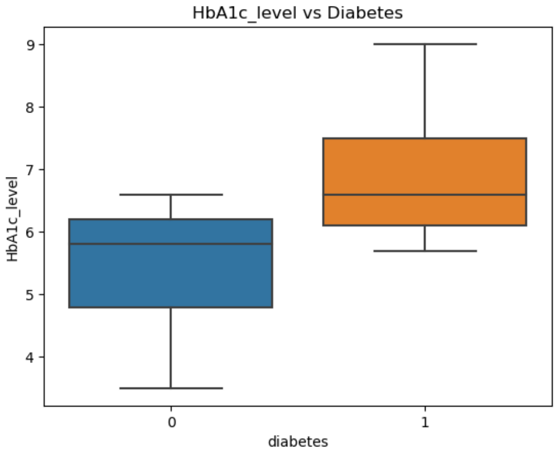


3.4 HbA1c Levels

HbA1c, or glycated hemoglobin, is a form of hemoglobin that is chemically linked with glucose. It's a significant marker for assessing long-term blood glucose control, providing an average glucose level over a span of about three months.

As diabetes is characterized by elevated and often erratic blood glucose levels, we would anticipate a strong correlation between HbA1c levels and diabetes, as confirmed by our dataset. Therefore, HbA1c is an extremely relevant feature for our predictive model. As the data shows those with diabetes have higher HbA1c and therefore higher average blood suger over a three-month period.

Figure 4: HbA1c Levels against Age



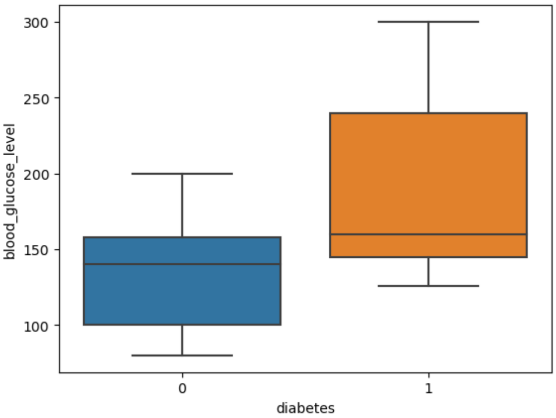
3.5 Blood Glucose levels

Blood glucose levels serve as a crucial parameter in diabetes detection and management. As our dataset confirms, individuals diagnosed with diabetes generally exhibit higher blood glucose levels than those without the condition.

Blood glucose provides a snapshot of an individual's current blood sugar level. If consistently elevated over time, these levels will also result in an increase in HbA1c, indicating poor long-term glucose control. Thus, while blood glucose and HbA1c measurements are distinct, they may exhibit a degree of correlation due to their common link to blood sugar levels.

Our analysis reveals a weak yet statistically significant correlation between these two variables. This suggests that while blood glucose and HbA1c levels do relate to one another, other factors also play a role in determining HbA1c levels.

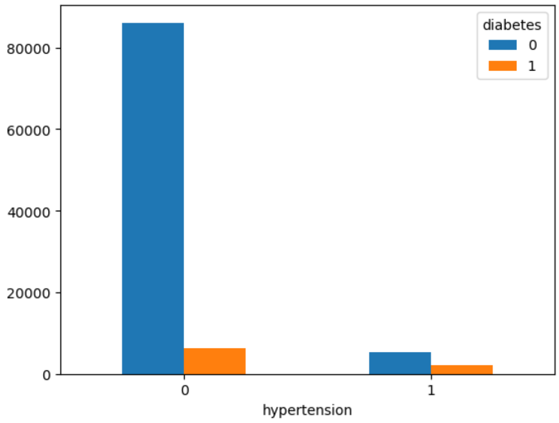
Figure 5: Blood Glucose Levels



3.6 Hypertension

Hypertension, commonly known as high blood pressure, has been recognized as a potential risk factor for the development of diabetes. In accordance with this understanding, our dataset shows that individuals with hypertension demonstrate a higher prevalence of diabetes compared to their normotensive counterparts.

Figure 6: Hypertension and Diabetes



3.7 Smoking

The relationship between smoking and diabetes is complex and warrants careful examination. From our initial dataset review, we observe some complications with the 'Smoking' category. Firstly, a considerable portion of the data is missing or not recorded, which could potentially skew our analysis and predictions.

Secondly, the category seems to contain repetitive or unclear classifications, creating some ambiguity. For instance, the category 'ever' might be a typographical error, likely intended to be 'never'. Such inconsistencies need to be addressed during the data cleaning process to ensure accuracy and clarity.

Figure 7 provides a visual representation of the current status of the 'Smoking' data. Our subsequent steps will include addressing these issues, standardizing the data, and assessing the possible impact of smoking on diabetes risk. Given the established detrimental health impacts of smoking, its inclusion in our model could enhance the comprehensiveness of our predictions. However, data quality and integrity must be ensured first to avoid introducing error or bias.

To improve the 'Smoking' category's data quality, one of the preliminary steps we took was to consolidate synonymous categories. This data clean-up process helped enhance clarity and avoid redundancies, making the dataset more efficient for our analysis. The outcomes of these adjustments are depicted in Figure 8.

Figure 7: Smoking and Diabetes

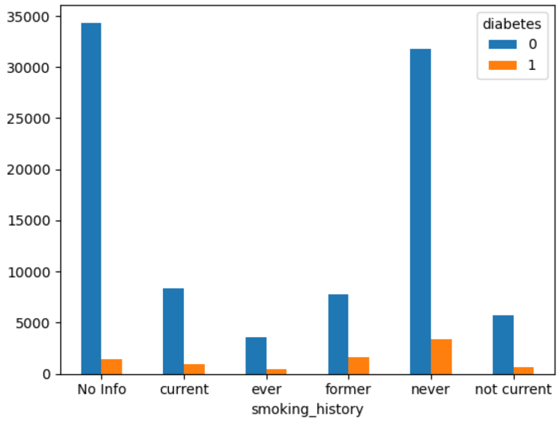
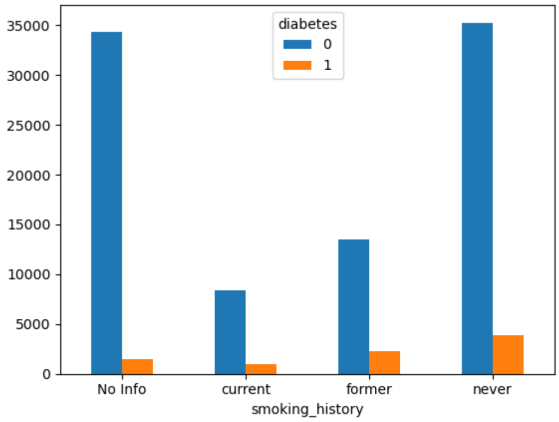


Figure 8: Smoking and Diabetes



Given the categorical nature of the 'Smoking' data, it was essential to transform this information into numerical representations to facilitate our predictive modelling. One possible method involved assigning values of 0.5 to 'previous smokers' and 'no info' categories, 1 to 'current smokers', and 0 to 'never smoked'.

The underlying assumption here is that being a former smoker is somewhat "in between" being a current smoker and having never smoked, and might therefore influence the chances of having diabetes accordingly. Despite this assumption not being directly confirmed by the calculated conditional probabilities, I decided to retain it, acknowledging that our data might not be independently sampled.

An alternative approach would have been to maintain 'smoked in the past' and 'smoked currently' as separate categories. However, this could introduce some collinearity, potentially compromising the accuracy of our regression models.

It's important to note that assigning a value of 0.5 to 'previous smokers' is a non-linear transformation. This is based on the premise that the effect of past smoking is half of the current smoking effect - an assumption that may not hold true, thus potentially impacting the performance of linear models.

Nevertheless, upon running all the models with and without the 'Smoking' category, I found the inclusion or exclusion of this feature to have negligible impact on the model results. This might indicate that while smoking is an important lifestyle factor, its role as an isolated predictor of diabetes in our dataset might be relatively minor compared to other factors. Further analysis and modeling might help validate this observation.

4 The models

Having performed our exploratory data analysis and data preparation, I now proceed to the model implementation phase. The primary aim is to evaluate different predictive models and identify the one that performs optimally in our context.

In terms of model evaluation metrics, I will primarily focus on 'recall'. Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives (in this case, people with diabetes) that are correctly identified. The rationale for this choice is rooted in the nature of diabetes as a medical condition. A false positive (a person incorrectly identified as having diabetes) is less harmful as they can undergo further tests and health monitoring. However, a false negative (a person incorrectly identified as not having diabetes) can be extremely risky, given the potential for untreated diabetes to lead to severe health complications.

To ensure the reliability and consistency of our analysis, we've divided our data into training and test sets. This split remains consistent across all model evaluations in this section, providing a fair basis for model comparison and selection. In the following subsections, I will delve into the specifics of the different models I will evaluate.

4.1 Logistic Regression

The first model I implemented was a Logistic Regression, often used as a baseline in binary classification tasks. This model, which uses a logistic function to model a binary dependent variable, is particularly useful in medical diagnostics, including diabetes prediction, due to its ability to provide probabilities and classify new samples using continuous and categorical input data.

The Logistic Regression model I trained yielded a precision of 0.92 and a recall of 0.81. Precision represents the proportion of true positives out of the predicted positives, implying that our model has a high accuracy in predicting actual diabetes cases. Recall, our primary metric of interest, at 0.81 indicates that the model correctly identified 81% of the actual diabetes cases.

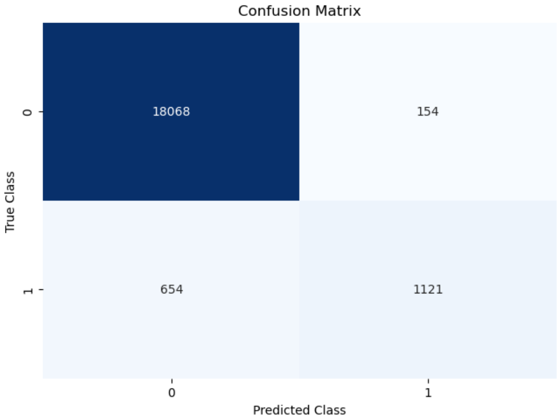
For our analysis, I chose to focus on the macro average instead of the weighted average. This decision is grounded in the fact that the group with diabetes is smaller and thus more crucial to correctly predict due to the risk associated with under-diagnosis. Using a macro average allows us to treat each class equally, providing a better indication of how our model performs on the minority class, in this case, the group with diabetes.

In the following sections, I will evaluate the performance of other models and compare their results against this Logistic Regression baseline.

4.2 Decision Tree

The Decision Tree is another model I employed for our analysis. Decision trees are flexible, powerful, and interpretable models that can handle both classification and regression tasks. They work by creating a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Figure 9: Logistic Regression



The trained Decision Tree model yielded a recall of 0.86 and a precision of 0.85. As our primary focus is on recall, the Decision Tree model outperforms the Logistic Regression model in capturing more true positive cases.

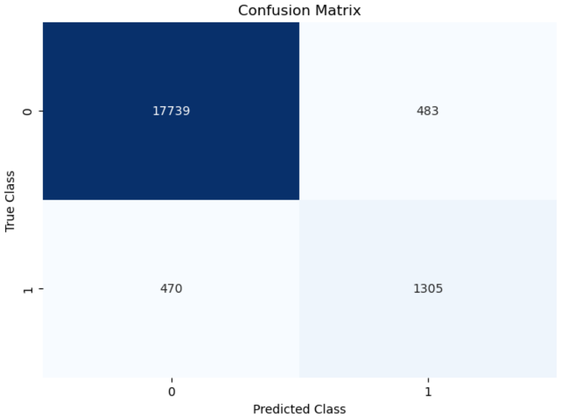
While the precision is slightly lower than that of the Logistic Regression model, meaning it identified more false positives, we accept this trade-off. Given the context of our study, we are more concerned about missing true positive cases (false negatives) due to the serious health risks associated with untreated diabetes. Therefore, from the perspective of recall, the Decision Tree model is superior to the Logistic Regression model.

The Decision Tree's superior performance suggests that the relationship between the variables might be non-linear, and the Decision Tree is better able to capture this complexity.

4.3 Deep Learning

Ialso explored a Deep Learning model for our prediction task. Deep Learning models are highly flexible and capable of learning complex patterns, making them a potentially powerful tool for our analysis. However, these models require careful tuning of numerous parameters and can be computationally expensive, as reflected in the longer training time compared to the previously discussed models.

Figure 10: Decision Tree



After adjusting various parameters, such as adding additional layers and increasing the training time, our Deep Learning model achieved a recall of 0.87 and a precision of 0.92. Even though the tuning process brings with it the risk of overfitting to the test set, the comparative performance to the other models suggests that overfitting is unlikely in this case.

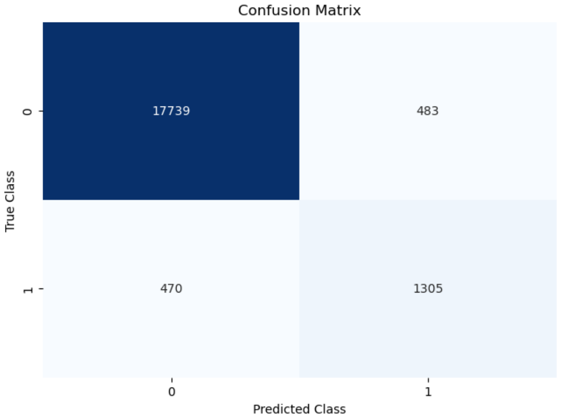
In terms of recall, which is our primary metric of interest, the Deep Learning model outperformed both the Logistic Regression and Decision Tree models. Its precision also matched the highest precision we've achieved so far with the Logistic Regression model.

Although the superior performance of the Deep Learning model is paired with a longer training time, this trade-off can be justified if the focus is on achieving the highest possible recall. The marginally better results could have significant implications in a real-world medical context where early and accurate detection of diabetes is crucial. Nevertheless, the interpretability and computational efficiency of the model are factors that need to be considered alongside performance.

4.4 Random Forests

Finally, I implemented a Random Forest model. Random Forest is a versatile machine learning method capable of performing both regression and classification

Figure 11: Deep learning



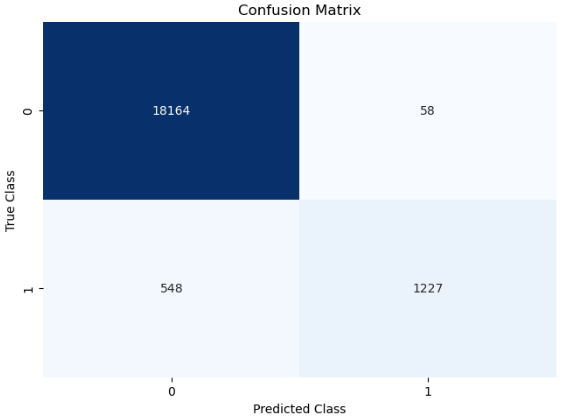
tasks. It operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes or mean prediction of the individual trees.

The Random Forest model achieved a recall of 0.87 and a precision of 0.92. Interestingly, these performance metrics matched those of the Deep Learning model, suggesting comparable prediction capabilities.

However, the Random Forest model holds a significant advantage over the Deep Learning model in terms of training speed. The Random Forest was able to train and provide predictions much faster, offering a substantial efficiency gain. This benefit makes the Random Forest model an attractive choice when considering computational resources and time constraints.

Although the Random Forest model is not as interpretable as a single Decision Tree, it usually provides a better generalization performance, making it a powerful tool for such prediction tasks. Given its performance and efficiency, the Random Forest model proves to be a strong contender as the optimal predictive model for our dataset.

Figure 12: Random Forest



5 Conclusion

In my quest to develop a model that can accurately predict diabetes, I evaluated several models: Logistic Regression, Decision Tree, Deep Learning, and Random Forest. Of these, the Random Forest model emerged as the most efficient and effective, balancing excellent performance with relatively low training time. The highest recall I achieved across all models was 87%, which implies that 13% of diabetes cases were still missed. While this performance is robust, there remains room for improvement.

In terms of data quality, one area that stood out was the representation of smoking habits. Although I made assumptions to transform this data into a numerical format for our analysis, having more detailed and accurate information about smoking habits could potentially improve the models' performance.

However, when I reevaluated the models excluding the smoking column, the difference was marginal. Less than 30 out of 100,000 individuals were classified differently due to the absence of smoking data. This finding suggests that, in our specific context, the smoking data had a minimal impact on overall model performance.

In conclusion, this study demonstrates the potential of machine learning models, particularly Random Forest, in predicting diabetes based on health indicators. Further improvements could be made by enriching the dataset with additional relevant features or fine-tuning the models with advanced techniques. Such efforts could help enhance early diagnosis and treatment of diabetes, ultimately contributing to improved patient outcomes.