Name: Shinza Jabeen

Student Number: 5012207852

Supervisor name: Ceni Babaoglu

Project Title: Diabetes Prediction Using Machine Learning

Date: 27 November,2023

Table of Contents

Abstract:

Diabetes prediction using Machine learning:

Based on the medical data of patients, I would like to predict the likelihood of diabetes in patients. This is a classification problem. I am using the Diabetes prediction dataset that is publicly available on Kaggle. It can be downloaded. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. The dataset consists of: Age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, blood glucose level, and diabetes. It has 100,000 rows and 9 columns. I will be primarily working with python as my programming language. My objective is to predict the likelihood of diabetes in patients, so I'll utilize machine learning methods focused on classification. The libraries I plan to use include matplotlib for data visualization, scikit-learn for constructing machine learning algorithms, and pandas for data ingestion and manipulation. The models I will be using are Logistic Regression, Random Forest, and Support Vector Machine. Using random forest, I can get the importance of each feature and identify the most important factors in determining the likelihood of diabetes. I am using classification metrics like Accuracy, Precision, Recall, F1 score, to measure how good the model is doing. After applying all the measures, I came up with the conclusion that all three of the algorithms have performed well in the terms of the accuracy, but the Random Forest has the higher accuracy as compared to logistics regression and the SVM classifier.

Introduction:

Diabetes is a long-term health issue in Canada, affecting people of all ages and causing severe health problems like heart disease, vision loss, kidney failure, nerve damage, and amputation.

Diabetes has three types: type 1, 2, and gestational diabetes. In 2013-2014, 3 million Canadians had diabetes, with older males having higher rates. Managing diabetes involves maintaining a healthy lifestyle, taking medication if needed, and monitoring blood sugar levels to prevent complications. Type 1 diabetes, a childhood condition requiring insulin, is prevalent in adults, while type 2 is more prevalent in adults and often linked to obesity or inactivity. Gestational diabetes increases the risk of type 2 diabetes, with 3 million Canadians diagnosed in 2013-2014, higher among older males. Maintaining a healthy lifestyle (eating healthy, exercising, and average weight) is crucial for reducing risk. According to the study, females aged 10 to 54 are diagnosed with diabetes 120 days before or 180 days after conception. The Canadian Public Health Agency works with provinces to track occurrences and establish policy. (Pelletier et al., 2012)

The following are my research questions for this project: the first research question is what are the medical and demographic factors that influence the likelihood of developing diabetes? The second research question is Based on patients' history, identify if they might be at a risk of developing diabetes?

Literature Review:

Previous Research:

During my research on diabetes prediction using machine learning (ML) and deep learning (DL) techniques, I went over a few research papers. The first article by Soni & Varma (2020) demonstrated that the Random Forest algorithm yielded the highest accuracy on the Pima Indian Diabetes Dataset. It Emphasizes the use of machine learning algorithms, particularly Random Forest, showcasing the significance of different features in predicting diabetes. It provides insights into machine learning that can aid in early diagnosis, leading to improved patient

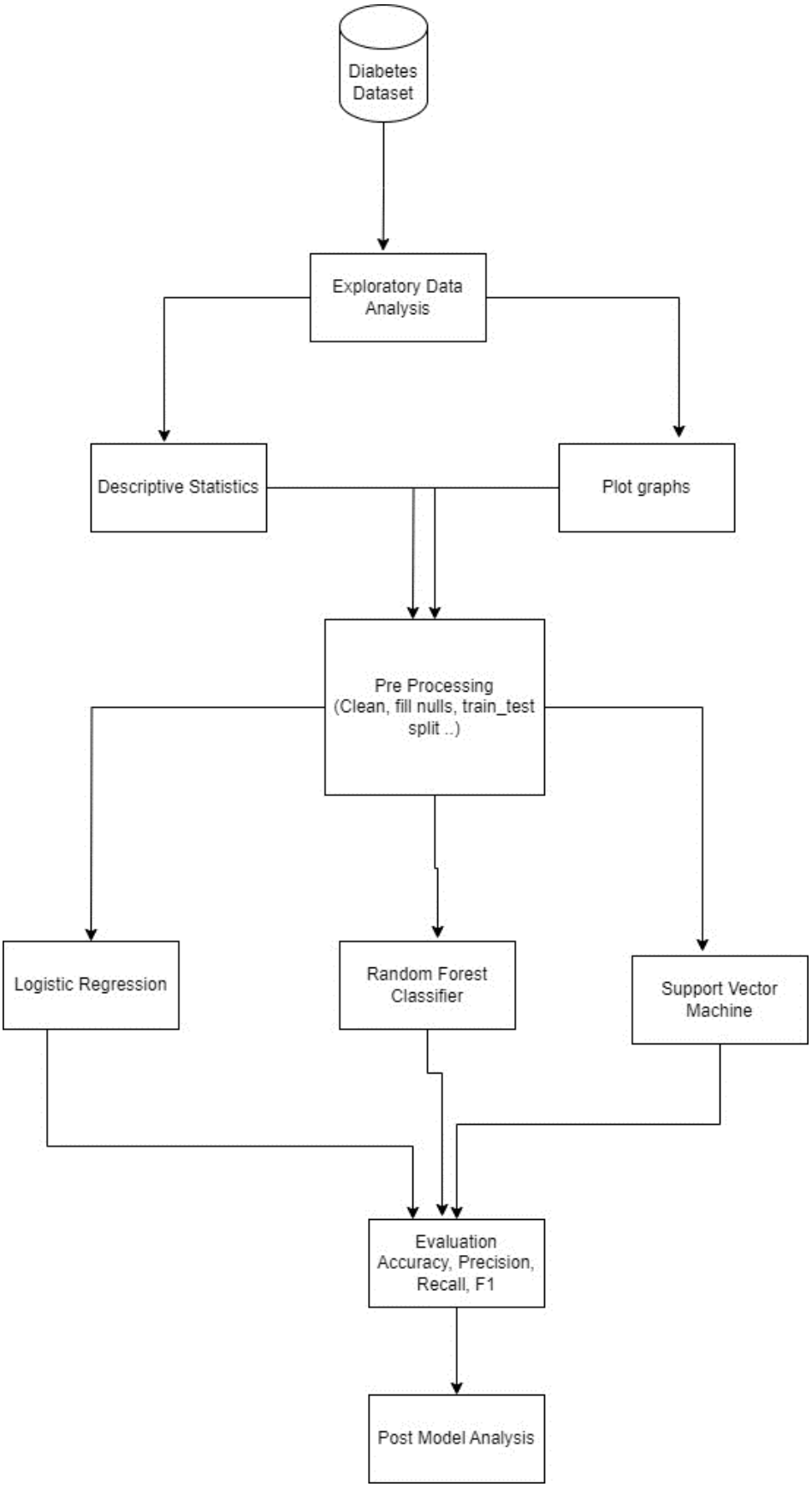
outcomes. Mujumdar & Vaidehi (2019) improved diabetes prediction accuracy to 97.2% by combining traditional features with external factors and employing Logistic Regression. It addresses the need for minimizing false diagnoses and automating the diagnosis procedure using machine learning, highlighting its potential impact on healthcare. Marie-Sainte et al. (2019) reviewed over 40 studies and suggested using lesser-known classifiers like REPTree for diabetes prediction. It delves into the advantages and disadvantages of various strategies, suggesting the fusion of infrequently used classifiers with other methods to enhance prediction accuracy. Hasan et al. (2020) developed a superior ensemble model, utilizing AUC as a weighting method. It uses the complex ensemble model, emphasizing preprocessing and feature selection methods to improve prediction quality. It highlights the use of AUC as a weighting method and demonstrates the efficacy of ensemble models in predicting diabetes. Sisodia & Sisodia (2018) found Naive Bayes to be the most effective classifier on the Pima Indians Diabetes Database. It underscores the potential for automating disease analysis and expanding the use of machine learning in healthcare for early disease identification. Azrar et al. (2018) emphasized the effectiveness of data mining in early diabetes detection, with Decision Tree being the most accurate. It emphasizes the value of data mining in healthcare for pattern identification and early diagnosis. Lastly, Ayon & Islam (2019) demonstrated the potential of deep learning, particularly a deep neural network, in achieving high prediction accuracy, showcasing its viability for early diabetes detection, and contributing to global healthcare advancements.

All the articles are presenting novel methodologies, algorithm comparisons, and insights into improving diabetes prediction, thereby contributing to the advancement of healthcare practices aimed at early disease diagnosis and patient care.

How my work builds on it:

To build on my work I have used the model's random forest, logistic regression, and Support Vector machine in my project, with oversampling using SMOTE and evaluated the performance by accuracy, precision, recall and F1- Score metrics.

Methodology:


Diabetes Dataset
Exploratory Data Analysis
Descriptive Statistics
Plot graphs
Pre Processing (Clean, fill nulls, train_test split ..)
Logistic Regression
Random Forest Classifier
Support Vector Machine
Evaluation Accuracy, Precision, Recall, F1
Post Model Analysis

Exploratory analysis:

In the exploratory data analysis, I will do the 2 parts. One is doing the visualization through the graphs like bar plots and box plots. In the descriptive statistics I will find the mean, median, mode, value count, standard deviation, and quartiles to analyze the data and to know the relation of between the variables.

Preprocessing:

In the preprocessing, I will clean the data and check for the null values and train the dataset for further machine learning algorithms.

Algorithms:

I will use the algorithms logistic regressions, Random Forest Support Vector to predict the diabetes based on features. In Ensemble I will combine the predictions and find the mode of it by combining them.

Evaluation:

After performing the machine learning algorithm, I will evaluate which algorithm performed better by Accuracy, Precision, Recall and F1-Score, through post model analysis, I will analyze the feature selection and feature importance.

Implementation:

- I have run the info on the diabetes dataset. I have got the information about the number of columns, label of columns, data types and the range index and get count of non-null values.

```
<class 'pandas.core.frame.DataFrame'>
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
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

- I have run the describe on the diabetes dataset and I got the summary statistics. .in this dataset the mean age is approximately 41.89 years, with a standard deviation of about 22.52 years. The minimum age recorded is 8 months, and the maximum is 80 years. Quartiles indicate that 25% of the data falls below 24 years, 50% below 43 years, and 75% below 60 years. The hypertension feature is binary, indicating whether an individual has hypertension (1) or not (0). On average, around 7.49% of the dataset has hypertension. Similar to hypertension, heart disease is also a binary feature denoting the presence (1) or absence (0) of heart disease. Approximately 3.94% of the dataset has heart disease on average. The bmi (Body Mass Index): mean BMI is approximately 27.32, with a standard deviation of about 6.64. The BMI ranges from a minimum of 10.01 to a maximum of 95.69. The HbA1c represents the HbA1c level, an indicator of average blood sugar levels over time. The mean HbA1c_level is around 5.53% with the standard deviation of about 1.07. The range spans from a minimum of 3.5% to a

maximum of 9.0%. blood_glucose_level Indicates the blood glucose level. The mean glucose level is about 138.06 with the standard deviation of 40.71, ranging from minimum 80 to maximum 300. Similar to hypertension and heart disease, diabetes feature is binary, indicating the presence (1) or absence (0) of diabetes. On average, around 8.5% of the dataset has diabetes. And diabetes is our target variable. These statistics offer a summary of the dataset, providing insights into the distributions, central tendencies, and variability of various attributes associated with diabetes prediction within the 100,000 records.

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

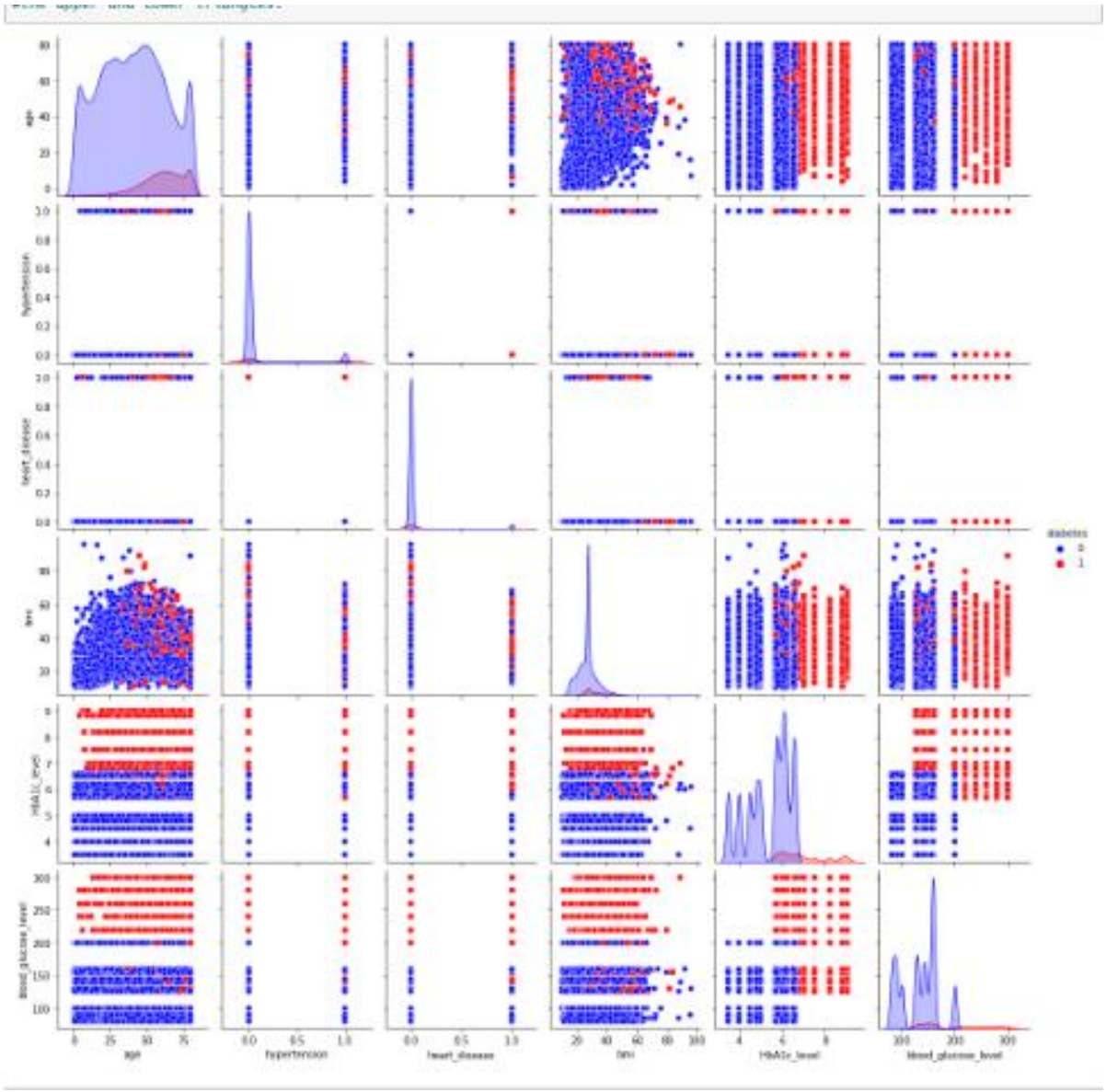
- I checked for the null values for the dataset, and it shows there is no null values in the dataset.

```
gender               0
age                  0
hypertension         0
heart_disease        0
smoking_history      0
bmi                  0
HbA1c_level          0
blood_glucose_level  0
diabetes             0
dtype: int64
```

- I have used the run the value count on the dataset, which indicates an imbalance in the dataset regarding the target variable "diabetes," where the 91500 instances labeled as "0" (indicating no diabetes) are significantly more prevalent than 8500 instances labeled as "1" (indicating diabetes). In cases of imbalanced datasets, predictive models may exhibit biases toward the majority class, potentially affecting their ability to accurately predict the minority class. Handling such class imbalances is essential for building robust and accurate predictive models. so, when I do the modelling, I will use the smote technique to balance the dataset.

PairPlot:

- I have checked the relationship between the numerical variables by using the Pairplot graph. Pairplots help to identify patterns or relationships between different variables and understand the distribution of individual variables. Following is the summary of a few of the variables.
- As this graph shows, some of the variables do not have any relationship. For example, hypertension has no relationship with age, heart disease, BMI, blood glucose level, HbA1c level. Similarly, heart disease does not have a relationship with age, heart disease, BMI, blood glucose level, HbA1c level.
- In this graph it shows the people who have HbA1c level less than 5 do not have diabetes and people who have HbA1c-level 7 or higher than this has diabetes.
- Some variables have strong relationships, for example people with less than 200 levels and less than the age of 20 or 25 does not have diabetes. And people who have a blood glucose level more than 200 and are 25 or older have more chances of getting diabetes. Below is the screen shot of the pairplot.

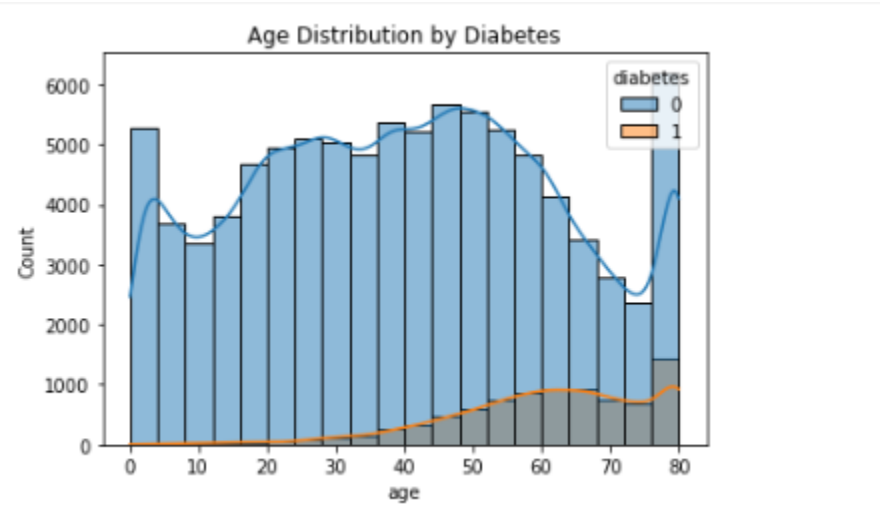


Histogram:

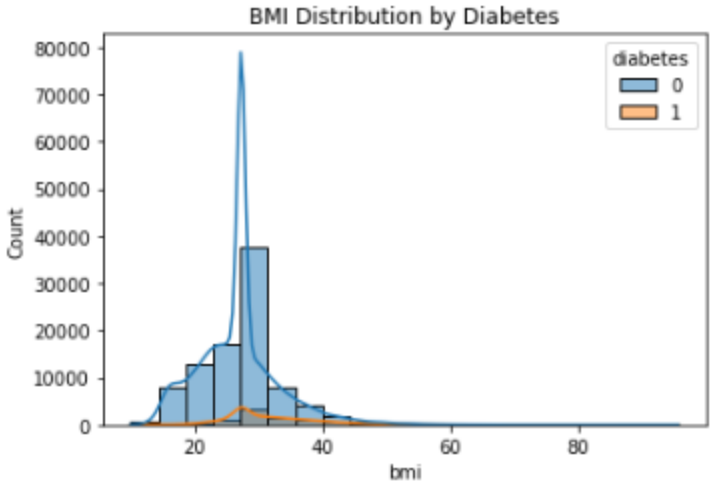
I have plotted the histogram of 3 variables, age BMI and blood glucose level with target variables.

The histogram depicts the age distribution among individuals, segmented by their diabetic and non-diabetic status. A notable trend observed is that the likelihood of diabetes tends to increase with age. Among individuals aged 0 to 30, there is a lower prevalence of diabetes, while as the

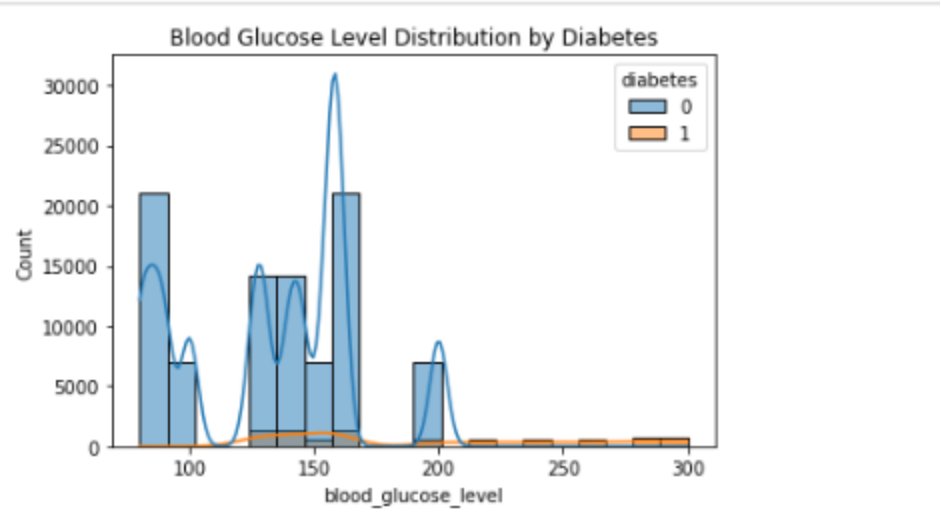
age progresses beyond this range, there is a noticeable rise in the occurrence of diabetes. This trend suggests an age-related correlation with the likelihood of developing diabetes, indicating a potential association between age and the onset of the condition."



BMI, a numerical measure representing body fat relative to weight and height, is a continuous variable. This visualization provides insights into the distribution of BMI among individuals, categorized by their diabetic and non-diabetic status. Notably, individuals within the BMI range of 0-20 demonstrate lower occurrences of diabetes. Conversely, as the BMI surpasses 20, there is an observable increase in the likelihood of having diabetes. This pattern suggests a potential correlation between higher BMI values and an elevated probability of diabetes within the dataset. It implies that higher BMI values might be associated with an increased propensity for diabetes onset.



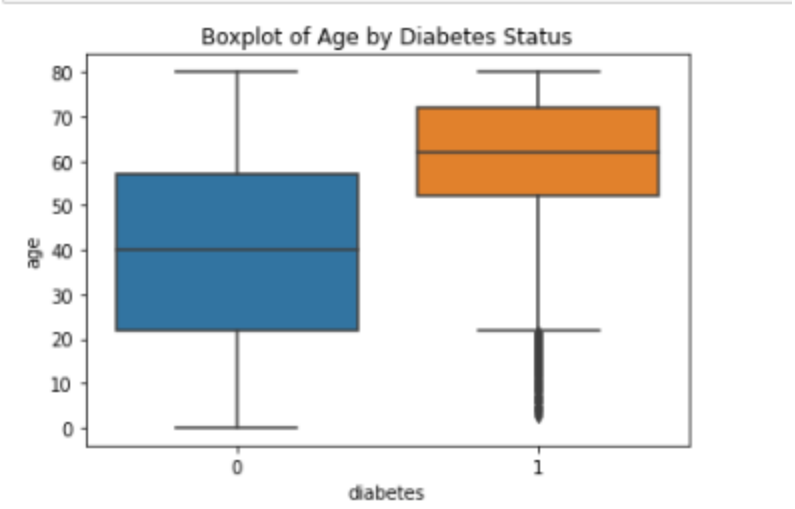
This histogram is segmented into 20 bins, each representing distinct ranges of blood glucose levels. The visualization offers a comparative view of blood glucose level distributions across individuals categorized by their diabetic and non-diabetic status. Notably, a clear trend emerges, indicating that as blood glucose levels increase, there is a concurrent rise in the likelihood of being diagnosed with diabetes. Moreover, elevated blood glucose levels beyond 100 suggest a higher probability of prediabetes or diabetes onset. This graph effectively illustrates how blood glucose levels vary among individuals with and without diabetes, highlighting the potential association between higher blood glucose levels and the presence of diabetes.



Box Plot:

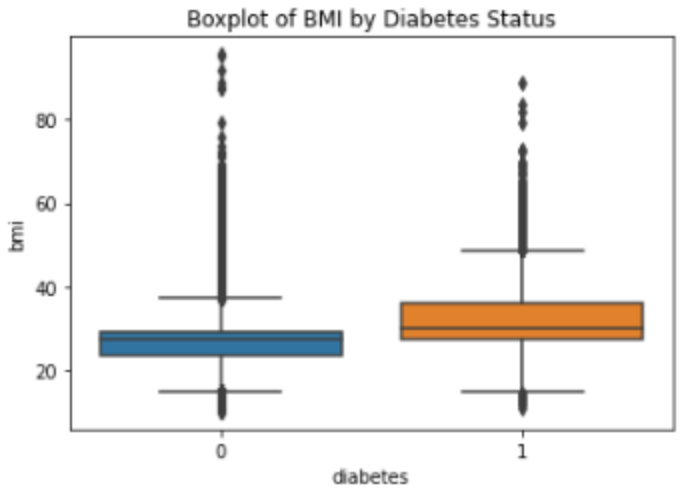
In this box plot of age and diabetes the 0 indicates non-diabetic individuals, and 1 represents diabetic individuals. A clear distinction is evident in the median ages between the two groups. The median age for individuals diagnosed with diabetes is 60, significantly higher compared to the median age of 40 observed among non-diabetic individuals.

The dataset reveals a notable concentration of individuals with diabetes in older age groups. This observation suggests a trend where individuals diagnosed with diabetes tend to be older on average, as depicted by the higher median age compared to the non-diabetic group.



The box plot of BMI and diabetes exhibits that dataset has numerous outliers within the BMI variable. Among non-diabetic individuals, the median BMI stands at 30, whereas among diabetic patients, the median surpasses 30, indicating a slightly higher median BMI for those diagnosed with diabetes.

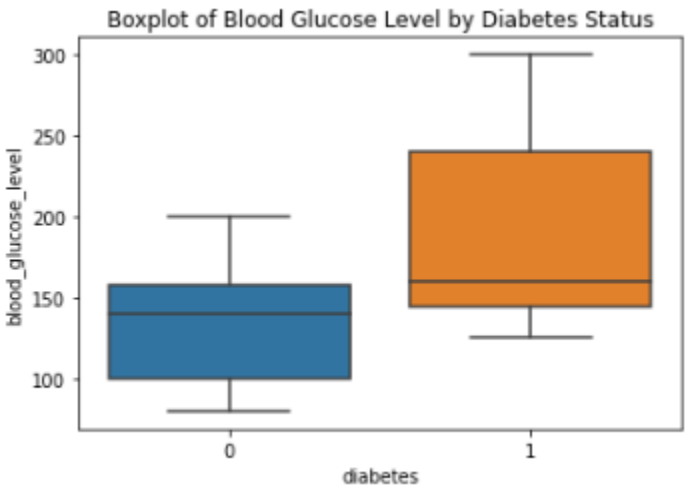
The BMI range for non-diabetic individuals spans between 25 to 35, while for diabetic patients, it ranges slightly higher, typically between 35 to 40. This distinction showcases a tendency for diabetic individuals to have a somewhat elevated BMI compared to non-diabetic individuals, as indicated by the wider range of BMI values and the higher median BMI among the diabetic group.



The box plot of blood glucose levels in relation to diabetes status reveals notable observations. The absence of outliers in blood glucose levels among diabetic and non-diabetic groups is noted. The median blood glucose level for non-diabetic individuals tends to be under 150, whereas for diabetic individuals, it tends to be higher, surpassing 150.

The boxplot representing diabetic patients exhibits a wider distribution, indicating higher variability and higher blood glucose levels among individuals with diabetes. Specifically, the median blood glucose level for non-diabetic individuals' hovers between 100 to 150, whereas for those diagnosed with diabetes, the median blood glucose level extends to around 150 to 250.

This comparison underscores a distinct pattern wherein diabetic individuals generally showcase elevated blood glucose levels, with the boxplot's width indicating a broader range of values. The median values further reinforce this trend, illustrating a noticeable difference in blood glucose levels between diabetic and non-diabetic populations.



GitHub link: **https://github.com/ShinzaJabeen/Diabetes_prediction**

Results:

- I have run the code for selecting key variables crucial for predictive modeling. These selected features encompass various aspects: "gender" denotes the individuals' gender, "age" signifies their age, while "hypertension" indicates the presence or absence of hypertension, distinguished by binary values (0 for absent, 1 for present). Similarly, "heart disease" is a binary variable denoting the presence of heart disease. Additionally, the "Smoking history" feature captures categorical data related to smoking habits. The remaining features encompass critical health indicators such as "bmi" (Body Mass Index), "HbA1c_level" representing blood sugar control levels, and "blood_glucose_level," which signifies the actual blood glucose level. These features collectively form the basis for predicting or modeling the target variable, "diabetes," which itself is a binary variable distinguishing between non-diabetic (coded as 0) and diabetic individuals (coded as 1).

- In the diabetes data set the "gender" and "smoking history" features contain categorical information. To enable their utilization in a machine learning model, I'm transforming these categorical variables into numerical labels by using the LabelEncoder. This conversion is necessary as machine learning models typically require numerical inputs rather than categorical data for processing and analysis. This will allow machine learning models to perform effectively.
- Then I have run the train_test_split function to divide the dataset into training and testing subsets. The "features" encompass the independent variables utilized for making predictions, while the "target" refers to Diabetes(X), the dependent variable (Y) I seek to predict and analyze. The dataset division involves allocating 80% of the data for training the machine learning model and reserving 20% for testing its performance. This approach aims to effectively analyze the model's predictive capabilities.
- I have used the Standard Scaler function to standardize specific columns within the training (X_train) and testing (X_test) datasets. The StandardScaler() from is used to scale numerical features ("age," "bmi," "HbA1c_level," and "blood_glucose_level") to have a mean of 0 and a standard deviation of 1. This ensures that these features are on a similar scale, preventing any particular feature from dominating the model training due to larger values.
- I am using the 3 machine learning classifiers: Random Forest, Logistic Regression, and Support Vector Machine. Each classifier is paired with a distinctive name or label. The purpose is to systematically implement each classifier on the data to evaluate their individual performances. By iterating through this list, I'll be able to compare their

effectiveness and determine which classifier is most suitable for the machine learning task at hand."

- I have used the code of Stratified K_Fold in cross-validation, which is beneficial, especially with imbalanced datasets. It’s configured to generate five splits or folds while maintaining the proportion of the target variable's classes in each fold. By preserving the class distribution within each fold, this strategy ensures a fair representation of all classes. This approach contributes to more dependable and impartial model evaluation.
- For the comparative analysis of the model, I go through each classifier, which measures the Accuracy precision and recall of the model and all three of them are important measure for the model performance.
- Precision, denoted by TP/(TP+FP), measures the likelihood that the model's prediction of a diabetes outcome truly corresponds to a diabetic condition. In this specific model, precision is calculated at 0.97.
- Recall, represented by TP/(TP+FN), gauges the model's ability to accurately predict actual diabetes outcomes. For this particular model, the recall value is 1.00, indicating a high rate of correctly identifying true diabetic cases.
- Accuracy, a measure of correctly classified records, is calculated as (TP+TN)/(TP+FP+TN+FN). In this model, the accuracy stands at 0.97, showcasing the model's proficiency in accurately classifying both positive and negative instances.
- These classifiers evaluate multiple classifiers through cross-validation on the training data and then apply them to the test data to assess their performance. The Random Forest classifier achieves an accuracy of 97% with precise identification of non-diabetic cases (97%) and reasonably high precision for diabetic cases (95%). For diabetic cases, it

exhibited a precision of 95%, implying that among all the cases it predicted as diabetic, 95% were indeed diabetic. This signifies that the Random Forest model performed well in correctly identifying both non-diabetic and diabetic cases. The precision values showcase the accuracy of the classifier's positive predictions for both classes, emphasizing its ability to minimize false positive predictions for non-diabetic cases and diabetic cases, leading to an overall high accuracy rate. However, the recall for diabetic cases at 69% suggests that the Random Forest model correctly identified only 69% of all actual diabetic cases present in the dataset. In other words, 31% of the actual diabetic cases were misclassified as non-diabetic or missed by the model. The F1-score is a combined measure of precision and recall, offering a balanced assessment of the model's performance. For non-diabetic cases, the F1-score of 0.98 indicates a strong balance between precision and recall, reflecting high accuracy and reliability in identifying non-diabetic instances. On the other hand, the F1-score of 0.80 for diabetic cases highlights a slightly lower balance between precision and recall, implying that while the model performs well, there might be a compromise between correctly identifying diabetic cases and avoiding false positives. In essence, the Random Forest model excels in accurately predicting non-diabetic cases but shows a comparatively lesser balance in correctly identifying all diabetic cases while avoiding misclassification.

- The Logistic Regression model achieved an overall accuracy of 96%, showing consistent precision of 96% for non-diabetic cases, indicating that 96% of the predicted non-diabetic instances were actually non-diabetic. However, its precision for diabetic cases is slightly lower at 86%, indicating that among the instances classified as diabetic, 86% were actually diabetic.

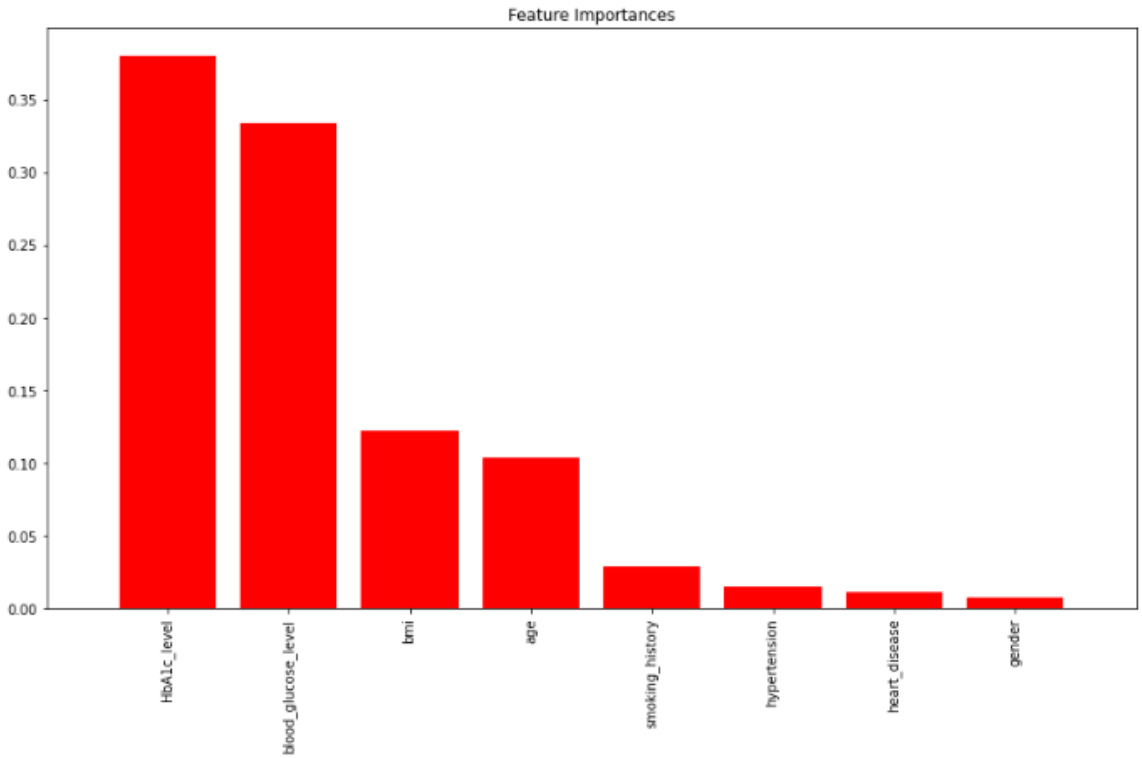
- Regarding recall, it's high for non-diabetic cases at 99%, meaning the model correctly identified 99% of the actual non-diabetic cases. But for diabetic cases, the recall drops to 61%, indicating that the model correctly identified only 61% of the actual diabetic cases present in the dataset. The F1-scores, which consider both precision and recall, are 0.98 for non-diabetic cases and 0.72 for diabetic cases. A high F1-score for non-diabetic cases suggests a balanced performance between precision and recall in identifying non-diabetic instances. However, the lower F1-score for diabetic cases suggests a relatively lower balance between precision and recall, implying a trade-off between correctly identifying diabetic cases and avoiding false positives in the predictions.
- The Support Vector Machine (SVM) classifier exhibits a mean cross-validation accuracy of 96%, signifying that it accurately predicts outcomes for around 96% of the total dataset. In terms of precision, both non-diabetic and diabetic cases demonstrate fairly high precision rates: 96% for non-diabetic cases and 92% for diabetic cases. This indicates that the SVM model correctly identifies around 96% and 92% of non-diabetic and diabetic cases, respectively, among the instances it classifies. The recall for non-diabetic cases is at the highest level, standing at 100%, indicating that the model identifies all actual non-diabetic instances. However, the recall for diabetic cases is notably lower at 57%, which implies that the SVM model misses around 43% of the actual diabetic cases present in the dataset. Considering the F1-scores, the SVM classifier has a high score of 0.98 for non-diabetic cases, suggesting a good balance between precision and recall for non-diabetic predictions. However, for diabetic cases, the F1-score drops to 0.71, indicating a trade-off between correctly identifying diabetic cases and avoiding false positives, which might result in missing some true diabetic cases.

- Overall, all three classifiers exhibit robust accuracy, with Random Forest achieving the highest, followed closely by Logistic Regression and SVM. However, their performance varies in terms of precision, recall, and F1-score, where each classifier shows specific strengths and weaknesses in predicting diabetic and non-diabetic cases.

```
Classifier: Random Forest
Cross-Validation Accuracy: 0.97 (± 0.00)
Test Accuracy: 0.97
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98     18292
           1       0.95      0.69      0.80      1708

    accuracy                           0.97     20000
   macro avg       0.96      0.84      0.89     20000
weighted avg       0.97      0.97      0.97     20000

----------------------------------------
Classifier: Logistic Regression
Cross-Validation Accuracy: 0.96 (± 0.00)
Test Accuracy: 0.96
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98     18292
           1       0.86      0.61      0.72      1708

    accuracy                           0.96     20000
   macro avg       0.91      0.80      0.85     20000
weighted avg       0.96      0.96      0.96     20000

----------------------------------------
Classifier: Support Vector Machine
Cross-Validation Accuracy: 0.96 (± 0.00)
Test Accuracy: 0.96
Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98     18292
           1       0.92      0.57      0.71      1708

    accuracy                           0.96     20000
   macro avg       0.94      0.78      0.84     20000
weighted avg       0.96      0.96      0.96     20000

----------------------------------------
```

- In a Random Forest classifier, the feature_importances_ attribute indicates the relative importance of each feature in predicting the target variable, in this case, diabetes. When examining the feature importance graph, higher values suggest a more influential role in

the model's predictions. The prominence of HbA1c_level and blood_glucose_level as the most important features signifies their significant impact on predicting diabetes. These variables, being measures of blood sugar control and glucose levels, are crucial indicators often associated with diabetes. Their higher importance implies they carry substantial information or patterns that strongly correlate with diabetic outcomes. Essentially, the Random Forest model relies more heavily on these features when making predictions, suggesting their strong predictive power in identifying diabetic cases within the dataset. Below is the screenshot of it.



Discussion:

- I have tried the model's Logistic regression, Support vector machine and the which has been used for the previous research articles. However, I have employed the SMOTE

(Synthetic Minority Oversampling Technique) approach that has not been used by these previous research articles. In this technique variables X_train_resampled and y_train_resampled encompass the resampled training data, ensuring a balanced distribution concerning the target variable. Through the SMOTE technique, synthetic samples are generated to augment the minority class, aiming to maintain similarity with existing examples while introducing some variability to prevent model overfitting.

- The utilization of SMOTE in this context aims to ameliorate the performance of machine learning models when faced with imbalanced datasets. By mitigating the bias toward the majority class, this technique fosters a more equitable representation of classes, thereby enhancing the model's ability to generalize and make accurate predictions across both minority and majority classes.
- The Random Forest classifier exhibits a mean cross-validation accuracy of 98%, displaying remarkable consistency across folds. It demonstrates precision rates of 98% for classifying non-diabetic cases and 76% for diabetic cases, showcasing its ability to predict each class. Notably, its recall stands at 98% for non-diabetic cases and 75% for diabetic cases, affirming its proficiency in correctly identifying both classes. The F1-scores, amalgamating precision and recall, achieve values of 0.98 for non-diabetic cases and 0.75 for diabetic cases, the former being the best possible score.
- Comparatively, Logistic Regression and Support Vector Machine (SVM) classifiers both attain a mean cross-validation accuracy of 89%. They display precision rates of 99% for non-diabetic cases and 42% for diabetic cases, indicating a bias towards effectively predicting the non-diabetic class. While their recall rates for non-diabetic cases stand at 89%, they achieve relatively higher recall rates of 88% for diabetic cases. However, the

F1-scores for Logistic Regression and SVM are 0.93 for non-diabetic cases and 0.57 for diabetic cases, denoting a lower efficacy in correctly identifying diabetic cases compared to Random Forest.

- In summary, these classifiers exhibit differing performance levels. Random Forest stands out with the highest accuracy and F1-scores for both classes, showcasing its overall effectiveness. On the contrary, Logistic Regression and SVM display lower precision and F1-scores for the minority class (diabetic cases), indicating a comparatively reduced ability to accurately identify instances of diabetes.
- After applying the smote techniques, the accuracy of the Logistic Regression has dropped slightly.

Shortcomings:

- The diabetes data set was highly imbalanced as mentioned earlier, it contains the 91500 observations that are of non- diabetic people and 8500 of diabetic's patients. Data imbalance refers to a scenario in a dataset where the distribution of classes is significantly skewed. One class (the majority class) might have a much larger number of instances compared to another class or classes (the minority class). This imbalance can lead to biased learning patterns in machine learning models, where the model tends to favor the majority class due to its abundance, leading to poorer predictive performance for the minority class. the other short coming was there was 100k observations in our dataset, the higher volume might prompt the use of more complex algorithms.

Ethical Considerations:

- The patient data must be kept private and confidential. Patient data must be anonymized to prevent identification. Protecting sensitive health information is paramount to comply with regulations.
- Algorithms must be developed to minimize biases that could disproportionately impact certain demographic groups. Biases could lead to unequal access to healthcare resources or inaccurate predictions for specific populations.
- The information should be gathered with Proper consent and protocols must be followed, ensuring that individuals are informed about how their data will be used and have given explicit consent for its utilization in research or analysis.
- There should be Beneficence and Non-maleficence by using data to improve healthcare outcomes should be the primary aim. Ensuring that the analysis or use of data doesn't cause harm is essential.
- There should be transparency regarding how the data is collected, used, and shared. Additionally, accountability measures should be in place to address any misuse or breaches.
- The data security measures are crucial to safeguard data against unauthorized access, breaches, or cyber threats.
- The Findings from the dataset analysis should be reported accurately and responsibly, avoiding sensationalism or misrepresentation of results that could lead to unnecessary panic or stigma.

Conclusion and Future Work:

In conclusion, I have the diabetes dataset from the Kaggle dataset. On this dataset I have done the exploratory data analysis and I have done the modeling by preprocessing the dataset and

evaluations. From that I got the results that Random Forest give the best Accuracy. However, after applying the smote technique to balance the imbalance dataset the Accuracy of the Random Forest model has dropped slightly. On the other hand, in terms of consistency both models' evaluations show high accuracy in predicting non-diabetic cases (class 0) but with slightly different recall scores for diabetic cases (class 1). The second set exhibits a better ability to identify actual diabetic cases, reflected in its higher recall and improved F1-score for class 1 compared to the first set. Overall, the Random Forest classifier seems adept at identifying non-diabetic cases but may vary in its ability to accurately predict diabetic cases, with the second set showing a slight improvement in this aspect.

Future research could entail conducting more comprehensive investigations into the parameters involved. With access to larger datasets, the utilization of neural networks for more intricate modeling becomes feasible. Additionally, employing ensemble techniques could further enhance the depth and accuracy of the analysis.

References:

Ayon, S., & Islam, M. (2019). Diabetes Prediction: A Deep Learning Approach. *I.J. Information Engineering and Electronic Business, 2*. Retrieved from https://j.mecs-press.net/ijieeb/ijieeb-v11-n2/IJIEEB-V11-N2-3.pdf

Azrar, A., Awais, M., Ali, Y., & Zaheer, K. (2018). Data Mining Models Comparison for Diabetes. *International Journal of Advanced Computer Science and Applications, 9*. Retrieved 2023, from https://pdfs.semanticscholar.org/3be5/c172b0c99e83cfd7a6e0b01ad89ed7363cfd.pdf

Hasan, K., Alam, A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access , Vol.8*. Retrieved 2023, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9076634

Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019, 10 29). Current Techniques for Diabetes Prediction. *Applied Sciences*. Retrieved 2023, from https://www.mdpi.com/2076-3417/9/21/4604/pdf?version=1572920776

Soni, M., & Varma, S (2020, 09). Diabetes Prediction using Machine Learning. *International Journal of Engineering Research & Technology, Vol. 9*(Issue 09). Retrieved 2023, from https://d1wqtxts1xzle7.cloudfront.net/64739619/diabetes_prediction_using_machine_learning_techniques_IJERTV9IS090496-libre.pdf?1603351804=&response-content-disposition=inline%3B+filename%3DIJERT_Diabetes_Prediction_using_Machine.pdExpires=1697291953&Sig

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science, Volume 165*. Retrieved 2023, from https://www.sciencedirect.com/science/article/pii/S1877050920300557?ref=pdf_download&fr=RR-2&rr=8169344f7ff1a253

Pelletier, C., Dai, S., Roberts, K. C., Bienek, A., Pelletier, L., & Onysko, J. (2012). Report Summary - Diabetes in Canada: facts and figures from a public health perspective. Retrieved 2023, fr

om https://www.researchgate.net/publication/234085745_Report_Summary_-_Diabetes_in_Canada_facts_and_figures_from_a_public_health_perspective

Sisodia, D., & Sisodia, D. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*. Retrieved 2023, from https://www.sciencedirect.com/science/article/pii/S1877050918308548?ref=pdf_download&fr=RR-2&rr=8169ba691849a232