DIABETES PREDICTION USING MACHINE LEARNING

CIND 820: Capstone Project

Prepared by: Muhammad Maqsud Hussain
m5hussain@torontomu.ca
Student #500978290

Toronto Metropolitan University

Supervisor: Dr. Ceni Babaoglu

Date of submission: 27 November 2023

Abstract

Diabetes mellitus refers to a group of diseases that affect how the body uses blood glucose. Glucose is an important source of energy for the cells that make up the muscles and tissues. It's also the brain's main source of fuel. Diabetes can lead to excess sugar in the blood. Too much sugar in the blood can lead to serious health problems, such as heart disease, vision loss, and kidney diseases etc. As per the International Diabetes Federation, 537 million adults (20-79 years) are living with diabetes and there are more than 5.7 million Canadians living with diabetes. Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. In this project, the aim is to develop a system to predict diabetes early for a patient using different machine learning techniques. The dataset is called "Diabetes prediction dataset" and it has 100,000 rows that consists of medical and demographic data collected from different patients along with their diabetes status (positive or negative). Besides, early detection of diabetes, analysis is performed to find some of the related research questions such as *Which factors are highly correlated with positive diagnosis of diabetes? How smoking history is related to the probability of becoming a diabetic patient in future? Is blood sugar level of an obese/overweight person always higher than a person with normal BMI? and How accurately can diabetes be predicted, and which model works best?* Five Machine Learning techniques will be adopted for this dataset to predict diabetes which are: *logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier and support vector model classifier.* Python with its different software libraries such as pandas, scikit-learn were used predominantly for performing the analysis. The analysis of results shows that Random Forest (RF) classifier model performs best and provides highest accuracy in predicting diabetes.

Table of Contents

Introduction

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose. Hyperglycaemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels (WHO, 2023). In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. According to Diabetes Canada, 30% of Canadians live with diabetes or prediabetes.10% of Canadians live with diagnosed diabetes which raises to 15% if we consider undiagnosed type-2 diabetes. People with diabetes are over three times more likely to be hospitalized with cardiovascular disease, 12 times more likely to be hospitalized with end-stage renal disease, and almost 20 times more likely to be hospitalized for a non-traumatic lower limb amputation compared to the general population (Diabetes in Canada n.d).

These numbers and figures are alarming which heightens the importance of early detection of diabetes which can save millions of lives all over the world. Researchers around the world has been using Machine Learning (ML) techniques to predict diabetes based on different attributes of a person such as weight, height, age blood glucose, BMI etc. Machine learning is a discipline of artificial intelligence (AI) that provides machines with the ability to automatically learn from data and past experiences while identifying patterns to make predictions with minimal human intervention. In this project, efforts have been made explore different machine learning techniques to predict diabetes from the diabetes dataset that consists of 100,000 instances.

The data set was retrieved from:

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

The raw data and processes for this study can be accessed from:

https://github.com/HussainM19/CIND-820-Predicting-Diabetes-using-MAchine-Learning

Literature Review

Several scholars used the machine learning (ML) method to predict diabetes using different datasets over the time. Due to the increasing trend of diabetes patients all over the world, researchers and data scientists from different part of the world have taken different approach to find a method for early prediction of this disease.

Soni & Varma, 2020 used six machine learning techniques in their project which are Which are K-Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB) and Random Forest (RF). They used a dataset with eight attributes and 768 instances. The dataset is called Pima Indian Diabetes (PID) and it contains attributes about female only. After applying the machine learning techniques, they measured the accuracy of the results of each method and Random Forest classifier achieved highest accuracy (77%). Also, importance of each feature has been measured and Glucose is found to the most important among the eight features considered in their project.

However, the paper titled *"A comparison of machine learning algorithms for diabetes prediction."* (Khanam & Foo, 2021) took a more detailed approach with the same dataset. They additionally used Neural Network (NN) methods in their research besides seven ML techniques to predict diabetes. They adopted feature reduction method and finally kept five input features (Glucose, BMI, Insulin, Pregnancy, and Age) from the PIMA dataset. All of their ML models provided an accuracy greater than 70%. However LR and SVM provided highest accuracy for both train/test split and K-fold cross-validation method. They used NN models with 1, 2, 3 hidden layers varying the epochs 200, 400, 800. Hidden layer 2 with 400 epochs provided highest (88.6%) accuracy. Finally, they recommended the Neural Network with two hidden layers as it has an accuracy rate of approximately 86% for all varying epochs (200, 400, 800).

Similar approach was taken with more detailed approach such as measuring feature importance for each of the adopted machine learning methods but using a different dataset of 2000 instances (Rani, 2020). In this paper, five different ML techniques were used, and an accuracy comparison table was presented. In this case, Decision tree showed highest accuracy for both the training (98%) and testing datasets (99%). Also, similar to the results found by Soni & Varma, 2020, Glucose has been identified as the most important feature for this dataset too.

Zou et al., 2018 adopted similar methods (decision tree, random forest and neural network) to predict diabetes mellitus. Their dataset originated from physical examination of patients in a hospital in a in Luzhou, China. It contains 14 attributes. They used five-fold cross validation to examine the models. To neutralize the impact of data imbalance, the researchers randomly extracted 5 times data and adopted the final result as the average of the five experiments. They also used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. Finally, their results showed that the prediction with random forest achieved highest accuracy (0.8084) when all the attributes were used.

Project performed by Tigga & Garg, 2020 was focused on predicting Diabetes type-2 in people. They have applied six machine learning methods and their analysis resulted Random Forest (RF) as the most accurate method after comparison. The dataset originated from data collected through online and offline questionnaires consisting of 18 questions relevant to diabetes. Also, after observing significance of all the variables, ‘Age’, ‘Family diabetes’, ‘Physically active’, ‘Regular Medicine’ and ‘Pdiabetes’ or gestation diabetes has highest significance among all variables and have greater impact on predicting diabetes than the rest.

Two main algorithms were applied on a dataset that has 800 instances with 10 attributes instances by Mujumdar & V, 2020 in their project. Algorithm 1 is diabetes prediction using various machine learning algorithms and Algorithms 2 is diabetes prediction using pipeline. For algorithm 1, the results show that logistic regression has the highest accuracy of 96%. However, they also compared the result of their diabetes dataset with the PIMA Diabetes dataset (which was the dataset used in the project by Soni & Varma, 2020 and Khanam & Foo, 2021). The comparison shows that their models give significantly better accuracy for the diabetes dataset than the PIMA dataset. For algorithm 2 (Pipelining), they also got highest accuracy for logistic regression (97.2%).

The project titled” Machine learning in precision diabetes care and cardiovascular risk prediction” by Oikonomou & Khera, 2023 is one of the latest related scholarly works. Their research is different from others as it is focused on how to use AI and ML to find new ways to optimize the management of diabetes and care and predicting cardiovascular risk to reduce related to diabetes. The paper is a comprehensive review of the various ML methods in the context of ML in the diagnosis, prognostication, phenotyping, and treatment of diabetes and its cardiovascular complications. Furthermore, the authors tried to identify key issues on equity and bias mitigation in healthcare and ways in which the regulatory framework can ensure the efficacy and safety of ML and AI in transforming cardiovascular care and outcomes in diabetes (Oikonomou & Khera, 2023).

Data Description

For the project, the Diabetes prediction dataset is chosen which is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. There are 100,000 instances/rows in this dataset. The dataset can be retrieved from the following link:

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

The dataset has total eight features or independent variables and one target feature/dependent variable.

The type of data features is described in Table-1.

Feature Name	**Type**
Gender	Categorical
Age	Numerical
Hypertension	Numerical
Heart disease	Numerical
Smoking History	Categorical
BMI (Body Mass Index)	Numerical
HbA1c (Hemoglobin A1c)	Numerical
Blood glucose	Numerical
Diabetes (Target Variable)	Numerical

Table-1: Feature type (Categorical or Numerical)

However, Brief description of each feature is depicted below:

1. **Gender**:

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

2. **Age:**

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

3. **Hypertension:**

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

4. **Heart disease:**

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

5. **Smoking History:**

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

6. **BMI (Body Mass Index):**

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.

7. HbA1c (Hemoglobin A1c):

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.

8. Blood glucose:

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.

9. Diabetes (Target Variable):

Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of it.

Exploratory Data Analysis

The explanatory data analysis section is divided into three parts. First, initial analysis is performed through finding some descriptive statistics of the features of the dataset. Then, Univariate analysis is performed for each of the independent variables and finally, bivariate analysis was performed in pair on some of the important variables.

(i) Initial Analysis:

Missing Values:

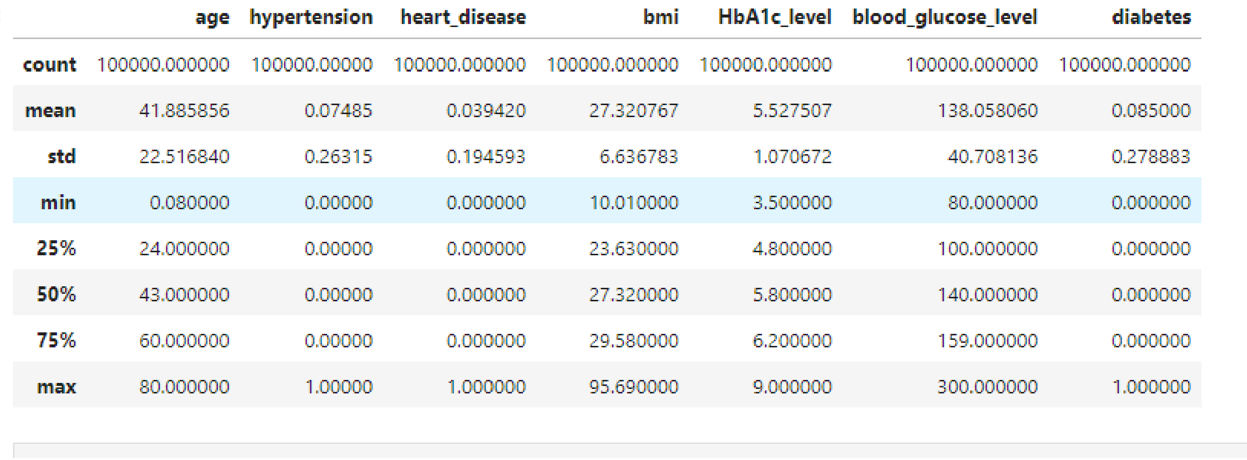
It has been found that there are no missing values for any of the variables. This indicates that there is not any chance of biases related to missing values in the analysis.

Duplicate Records:

There are 3854 records among the 100,000 rows have been found as duplicate. As these records are anonymous, duplicate records are acceptable which means there are people who may have same gender, age, BMI, blood glucose etc. among the 100,000 data records collected. Therefore, duplicate records are not removed from the dataset.

Statistical Measures:

Different statistical measures such as mean, mode, standard deviation of the numerical variables are summarized in Table-2.



	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

Table-2: Statistical measures for numerical features.

(ii) Univariate Analysis:

1) Gender:

There are 58,552 instances of female and 41,430 instances of male in the gender column. There are only 18 instances of “other” gender which is very low. Therefore, records with “other” are removed from the dataset.

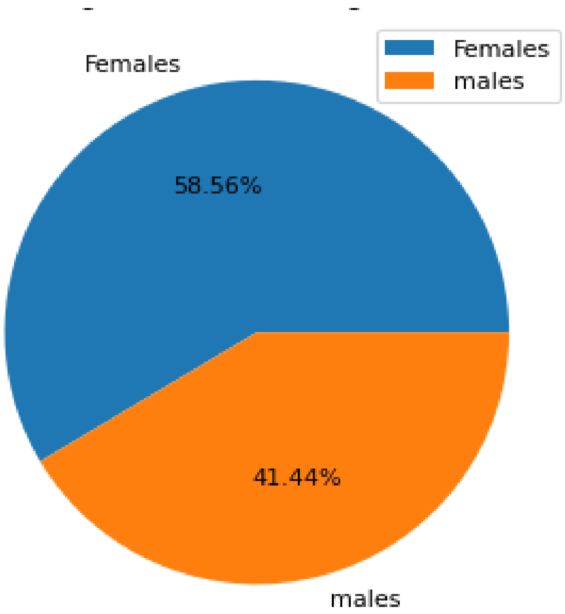


Fig-1: Ratio of male and female in the dataset.

2) Age:

The analysis of the age variable indicates that the dataset consists of records of newborn baby to a person aged 80 years. The mean age is 41.88 years for this dataset. A bar chart showing distribution of age is shown in Fig.3 at the following page.

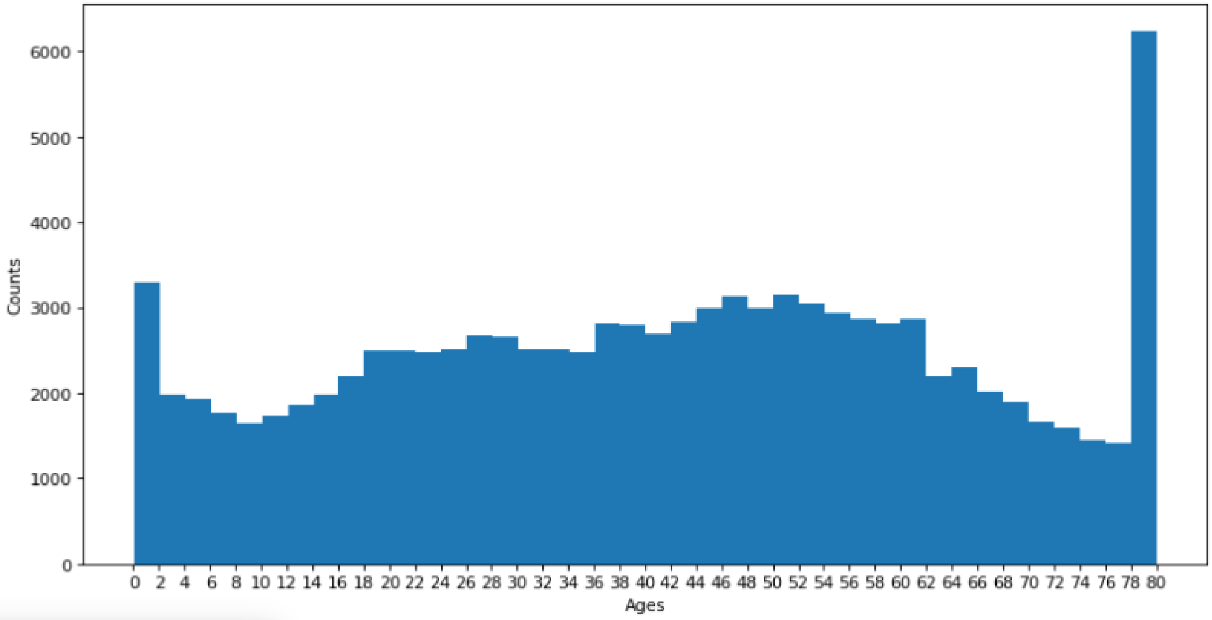


Fig.2: Distribution of "age" in the dataset

3) Hypertension:

There are several research that indicate that hypertension or high blood pressure may worsen complications in diabetes patients. In the dataset of this project, the percentage of patients with hypertension is 7.49% in comparison with 92.51% with no diagnosed hypertension.

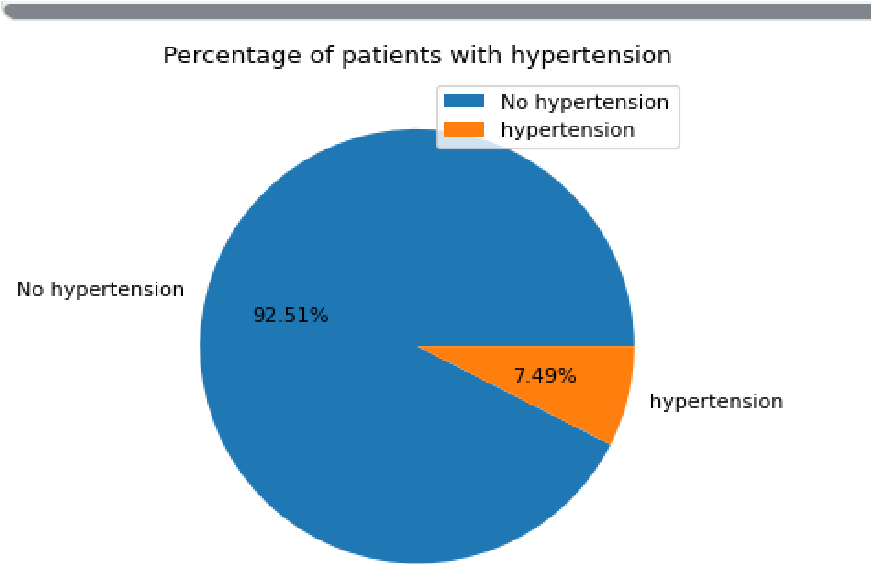


Fig.3: Ratio of persons with "Hypertension" in the dataset

4) Heart Disease:

High blood glucose from diabetes can damage a person's blood vessels and the nerves that control heart and blood vessels. This damage may gradually lead to heart disease (*Diabetes, heart disease, & stroke – NIDDK*, n.d). People with diabetes tend to develop heart disease at a younger age than people without diabetes. In the dataset, number of persons with heart disease is 3942.

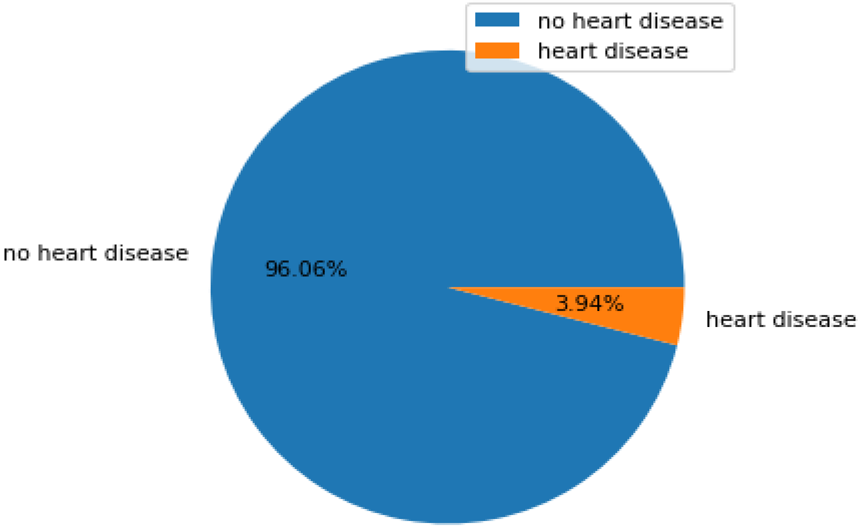
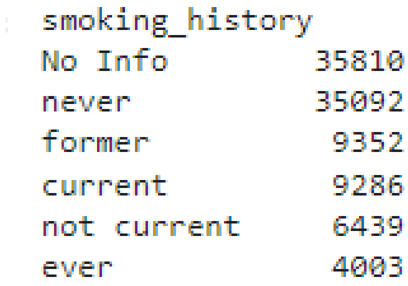


Fig.4: Ratio of persons with heart disease in the dataset

5) Smoking History:



```
smoking_history
No Info        35810
never          35092
former          9352
current         9286
not current     6439
ever            4003
```

Table-3: Count of Smoking history column in the dataset

35810 counts or 35.81% of instances in the smoking history feature has *No info* about smoking which is high and may impact the result significantly. Therefore, this column is removed from the dataset before performing further analysis.

6) Body Mas Index (BMI):

Both the box plot and violin plot of BMI indicate that there are outliers in the in this feature.

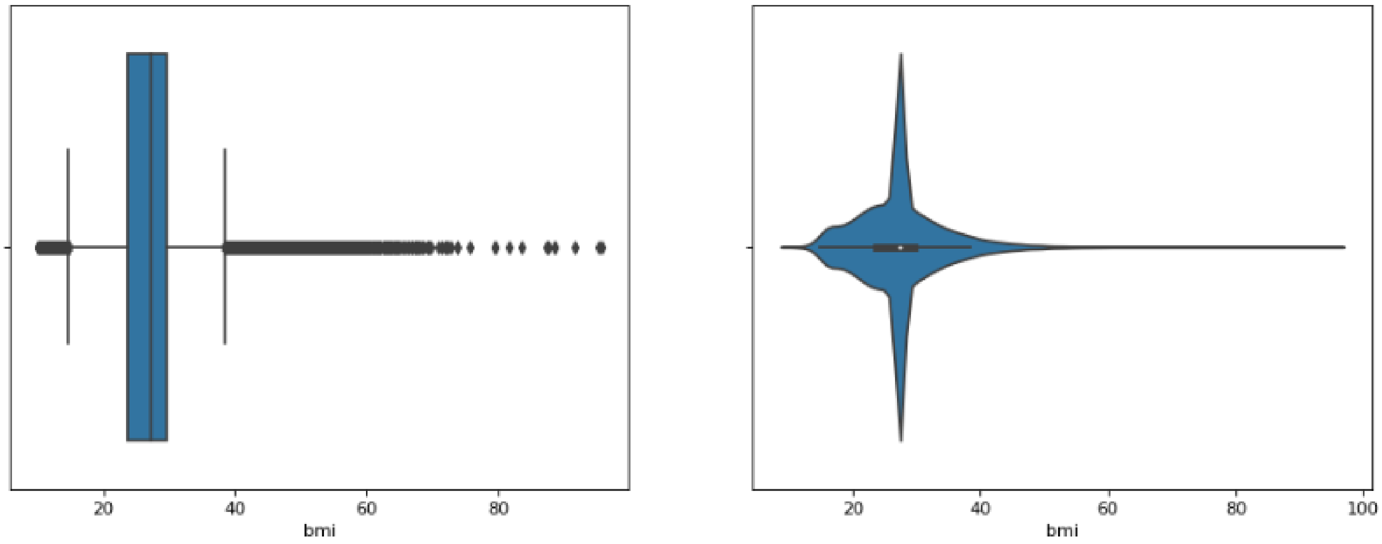


Fig.5: Boxplot and Violin plot for BMI showing outliers.

Therefore, removing outliers is an essential step to ensure better modelling and outcome. Removing those outliers will help in getting better results.

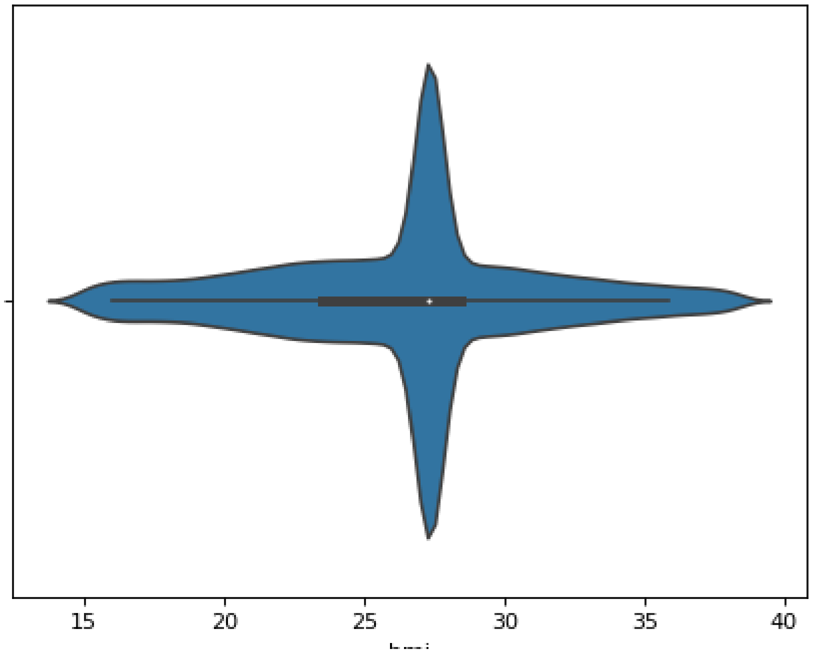


Fig.6: Violin plot for BMI after removing outliers.

7) Hba1c level:

Hba1c level for the patients in the dataset ranges from 3.5% to 9%. Any patient having 5.*7% or below can be classified as non-diabetic* while *5.7%-6.4% is prediabetic* and *6.5% or above is a diabetic patient.*

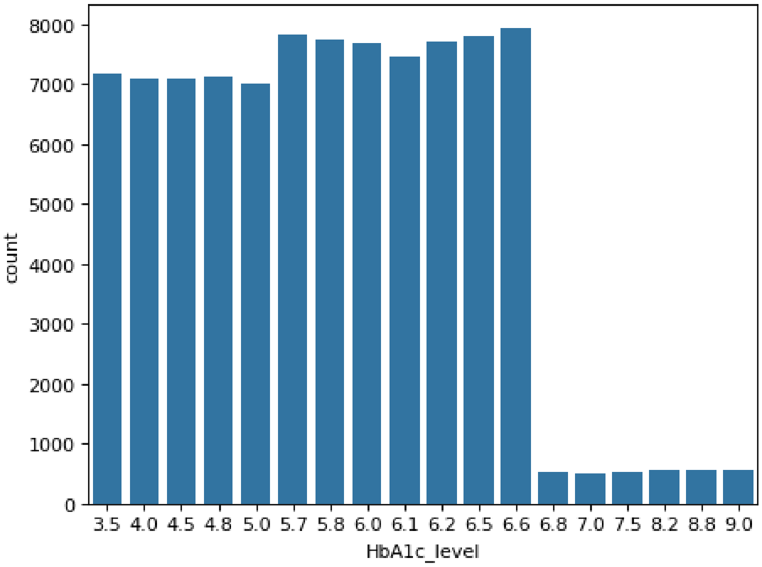


Fig.7: Distribution of HbA1c level in the dataset.

8) Blood glucose level:

Blood glucose level column ranges from 80 mg/dl to 300 mg/dl for the patients of this dataset. A fasting blood sugar level (measured in the morning before eating or drinking) of *99 mg/dL or lower is normal*, *100 to 125 mg/dL indicates having prediabetes, and 126 mg/dL or higher indicates that the person is a diabetic.*

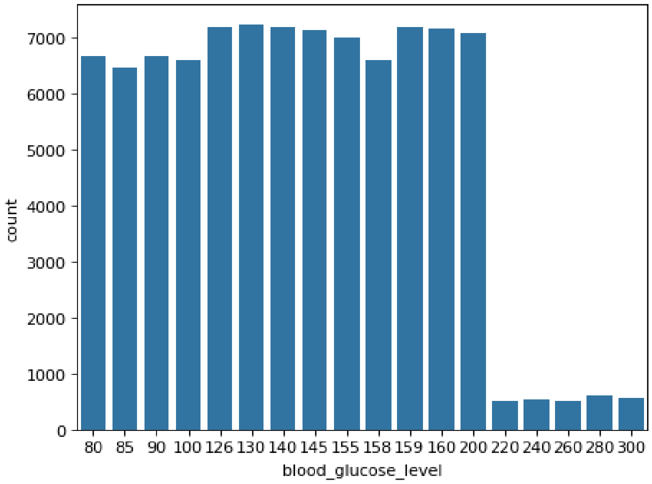


Fig.8: Distribution blood glucose level of persons in the dataset.

9) Diabetes (Target Feature):

Diabetes is the dependent variable or the target feature in this dataset. The ultimate goal of this project is to develop a methodology that will determine whether a person has diabetes or not based on the dependent variables. T

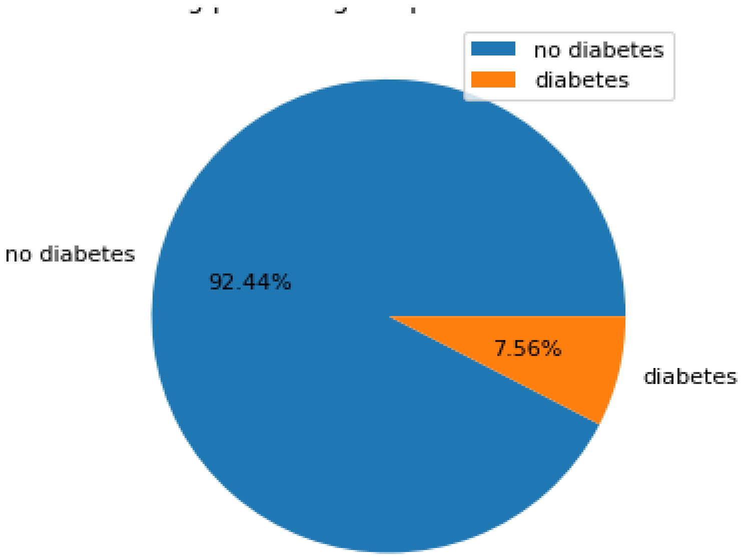


Fig.9: Distribution blood glucose level of persons in the dataset.

The percentage of patients having diabetes in this dataset is 7.56% which may seem to be an imbalance. However, the global diabetes prevalence is estimated to be 9.30% (Saeedi et al., 2019). As there is no geographic location of the source of the dataset, considering the global percentage of diabetic patients, it may be accepted.

(iii) Bivariate Analysis:

Bivariate analyses have been performed between the target variable and three other important features.

- **HbA1c_level vs. diabetes:**

The analysis of initial and actual diagnosis of diabetes based on HbA1c level shows that the value of HbA1c_level affects highly on the prediction of diabetes. If the value of HbA1c lies below the normal value, no person has an actual diagnosis of diabetes. However, As HbA1c value increases, the percentage of actual diagnosis of diabetes increases and for HbA1c level 6.5 or greater, it is 23.64%.

HbA1c level	**Initial diagnosis**	**Actual diagnosis**
< 5.7	Normal	100% no diabetes
5.7 – 6.4	Prediabetes	7.47% have diabetes
>= 6.5	Diabetes	23.64% have diabetes

Table-4: HbA1c level vs initial vs actual diagnosis.

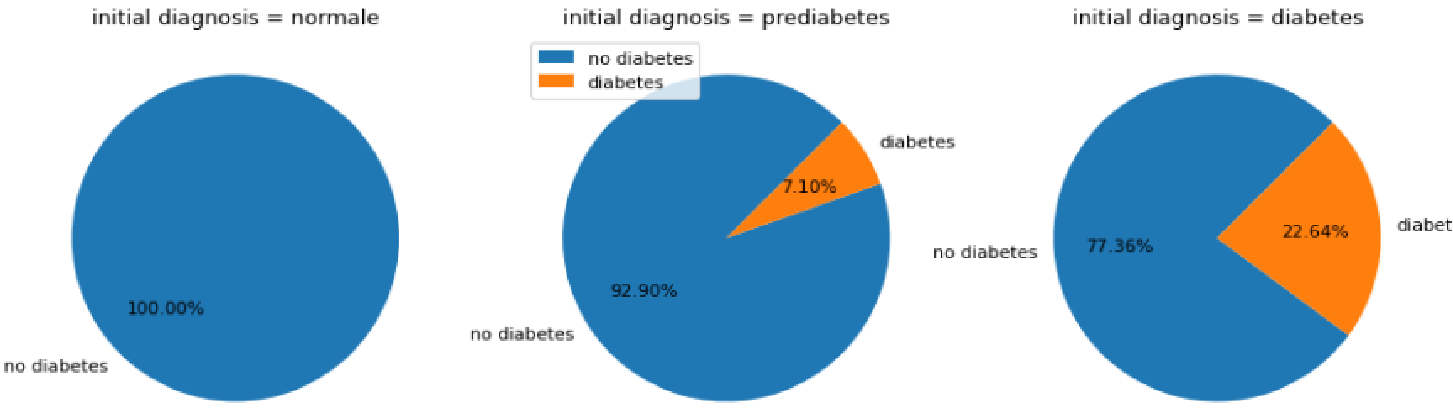


Fig.10: Pie Chart for HbA1C vs Diabetes

- **BMI vs diabetes:**

Body mass index is a measure of body fat based on height and weight that applies to adult men and women. It is calculated as weight in kilograms divided by height in meters squared. According to ordinal category [underweight, normal, overweight, obesity], as weight category increases, percentage of patients with diabetes increases.

BMI	Category	Prediction
=< 18.5	Underweight	0.7% have diabetes
18.5 – 24.9	Normal	3.85% have diabetes
25 – 29.9	Overweight	7.88% have diabetes
>= 30	Obesity	15.94% have diabetes

Table-5: BMI vs Prediction.

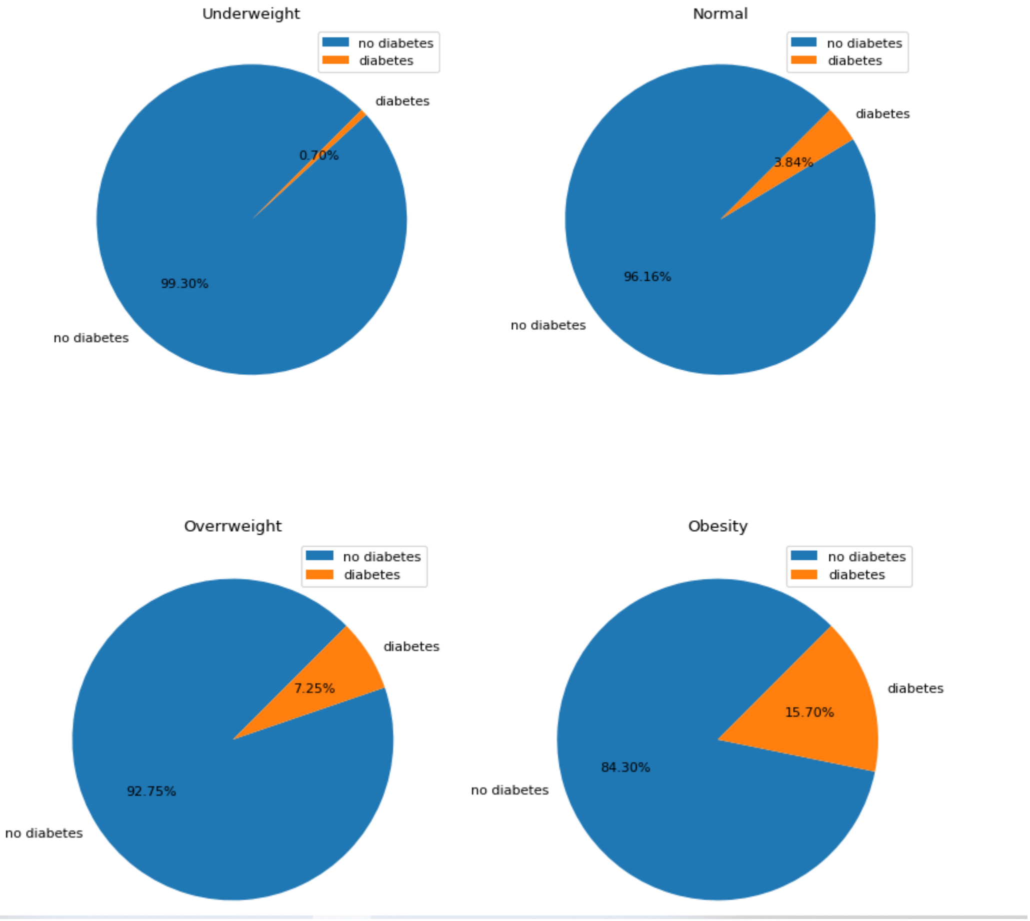
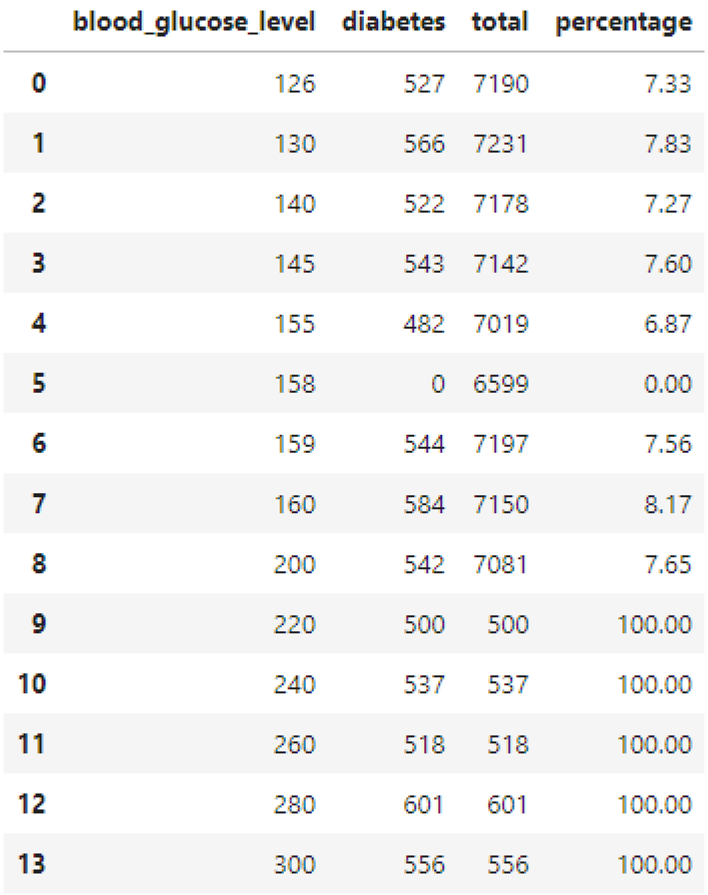


Fig.11: Pie Charts for different ordinal category and diabetes

➢ Blood Glucose Level vs diabetes:

Blood Glucose Level	Category
=< 99	Normal
100 – 125	Prediabetes
>= 126	Diabetes

Table-7: Standard chart for category of patients against blood glucose level in mg/dl



	blood_glucose_level	diabetes	total	percentage
0	126	527	7190	7.33
1	130	566	7231	7.83
2	140	522	7178	7.27
3	145	543	7142	7.60
4	155	482	7019	6.87
5	158	0	6599	0.00
6	159	544	7197	7.56
7	160	584	7150	8.17
8	200	542	7081	7.65
9	220	500	500	100.00
10	240	537	537	100.00
11	260	518	518	100.00
12	280	601	601	100.00
13	300	556	556	100.00

Table-8: Blood Glucose level and number & percentage of diabetic in the dataset

It is evident from the tables 7 and 8 that most of the patients with high blood sugar level have diabetes. Specifically, 1005 of patients with blood sugar level of 220 mg/dl and above are diagnosed with diabetes.

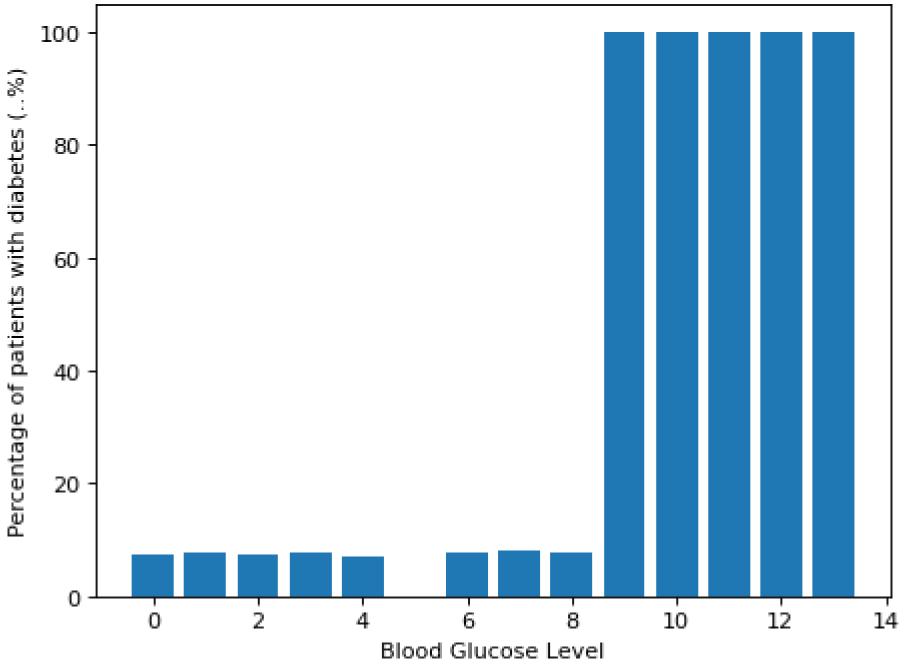


Fig.12: Bar plot for Glucose Level vs percentage of persons with diabetes

Therefore, Blood glucose level is a highly important variable in predicting diabetes.

Applied Methodology & Study Design

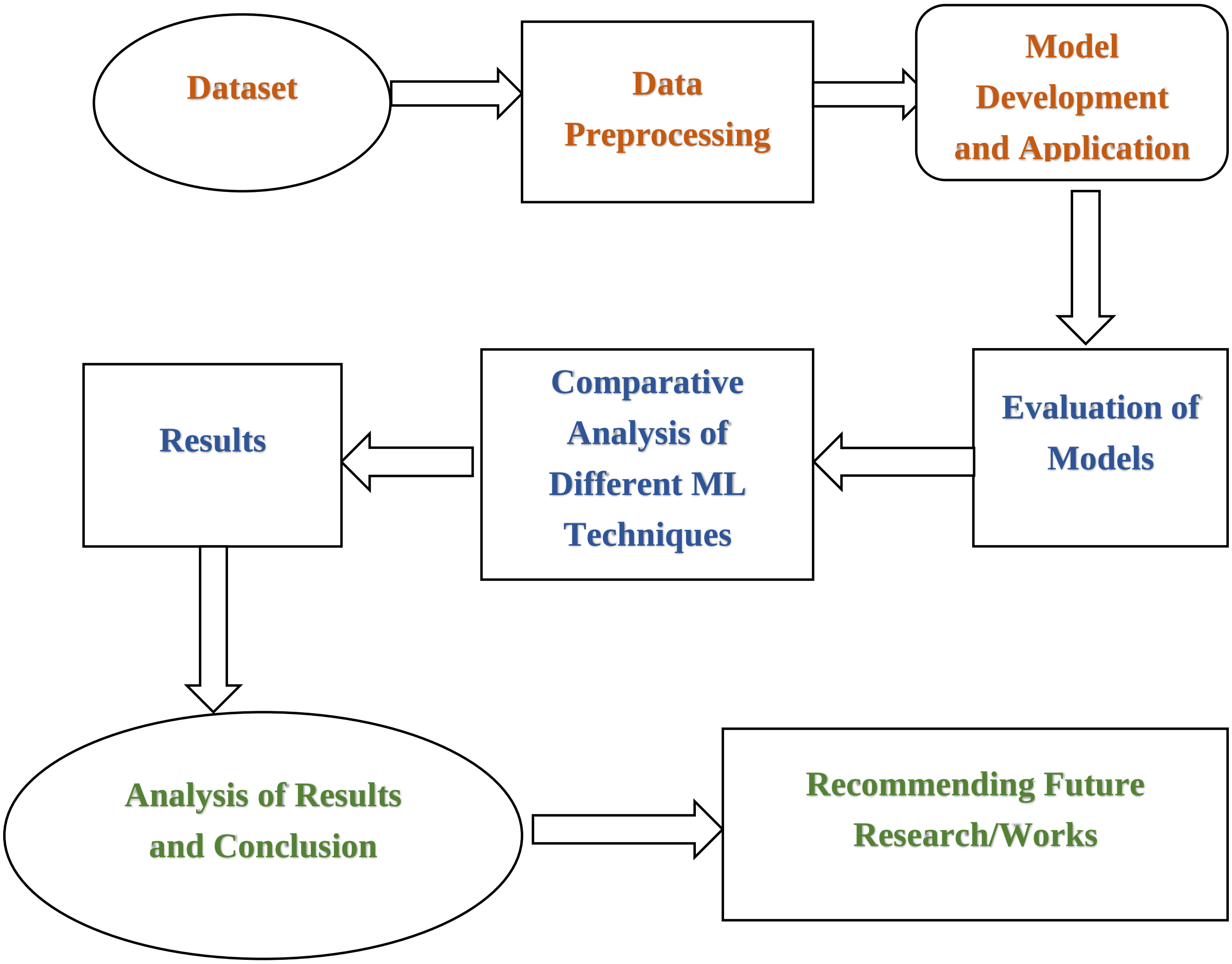

Dataset
Data Preprocessing
Model Development and Application
Evaluation of Models
Comparative Analysis of Different ML Techniques
Results
Analysis of Results and Conclusion
Recommending Future Research/Works

Fig.13: Applied Methodology

The applied methodology has three different stages. Initially, with the raw dataset, *cleaning and preprocessing* had been performed. This process includes removal of data/attributes, transformation or conversion of data, source data correction. and detection of inconsistency within the records. Then further exploratory analysis was performed on the dataset. This step confirms that the data is the consistent and ready for the next stage.

In the next stage of the project, different machine learning algorithms were developed, and models were trained using 80%-20% train-test split. Further scaling on the dataset was performed before applying the five algorithms were: *logistic regression, k-nearest neighbor regression*, *decision tree classifier*, *random forest classifier, support vector modeling*.

After the first iteration of each model, outcomes are evaluated. In order to improve the outcomes, cross-validation was applied on the model.

In the next stage, comparative analysis based on the outcomes was executed and models with best results were identified.

Next, we analyzed the results of the best model and identified the solutions for research questions. Recommendations were also drawn in this stage.

Finally, possibility of future works has been discussed based on the result analysis and limitations of the project.

Results

After the initial iteration of the chosen five machine learning models, confusion matrices were generated and analyzed.

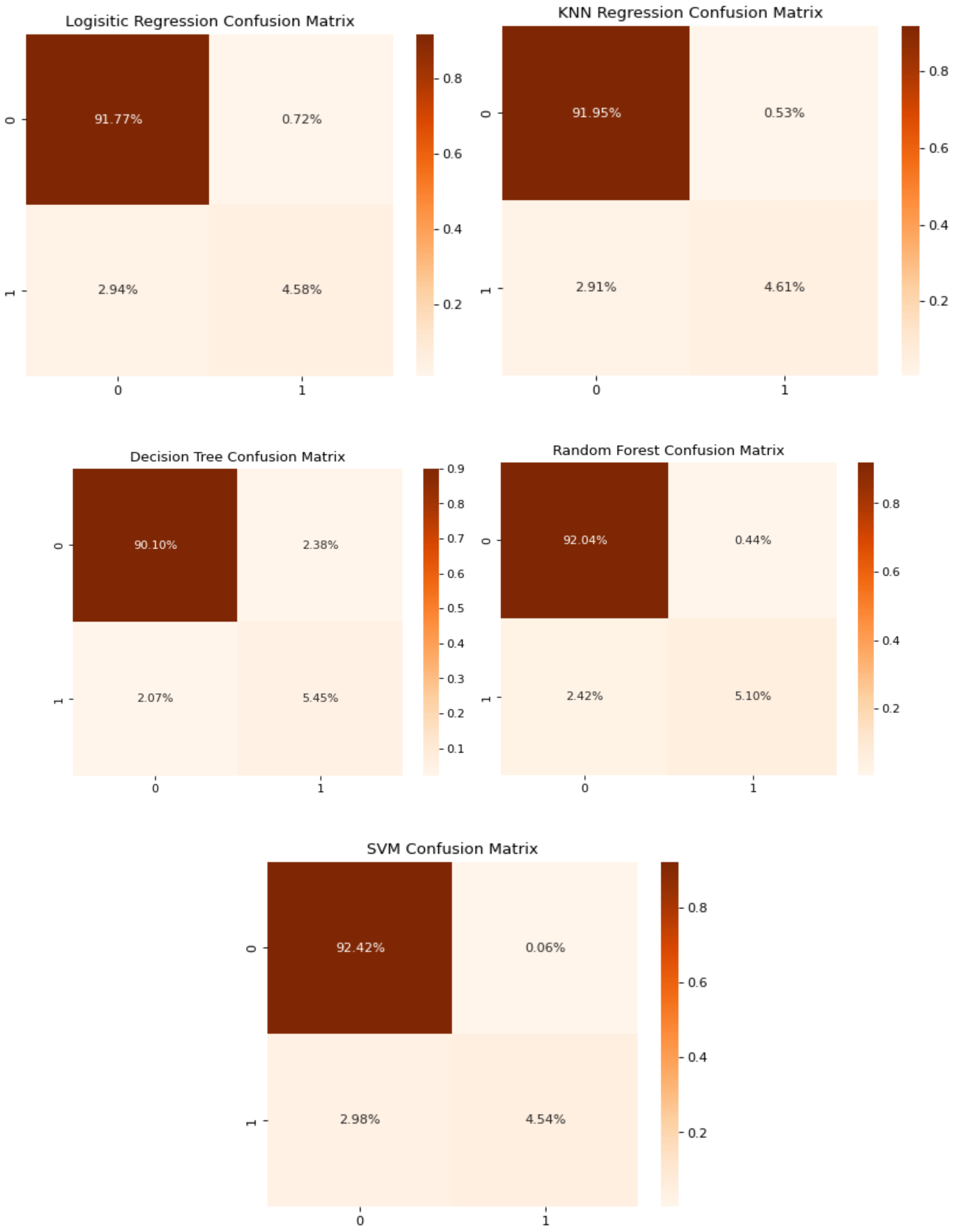


Fig.14: Confusion Matrices after initial iteration.

A review of the evaluation metrics of each algorithm implies that the Random Forest algorithm performs best among the five models with 97.14% accuracy and with an F1 score of 0.780709. Decision tree algorithm performed worst with 95.5% accuracy and an F1 score of 0.709677.

	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.971367	0.920311	0.677881	0.780709
4	Support Vector	0.969537	0.985965	0.603436	0.748668
1	KNN Regression	0.965554	0.896335	0.612742	0.727891
0	Logistic Regression	0.963455	0.864837	0.609162	0.714826
2	Decision Tree	0.955436	0.695533	0.724409	0.709677

Table-8: Evaluation metrics of the five models.

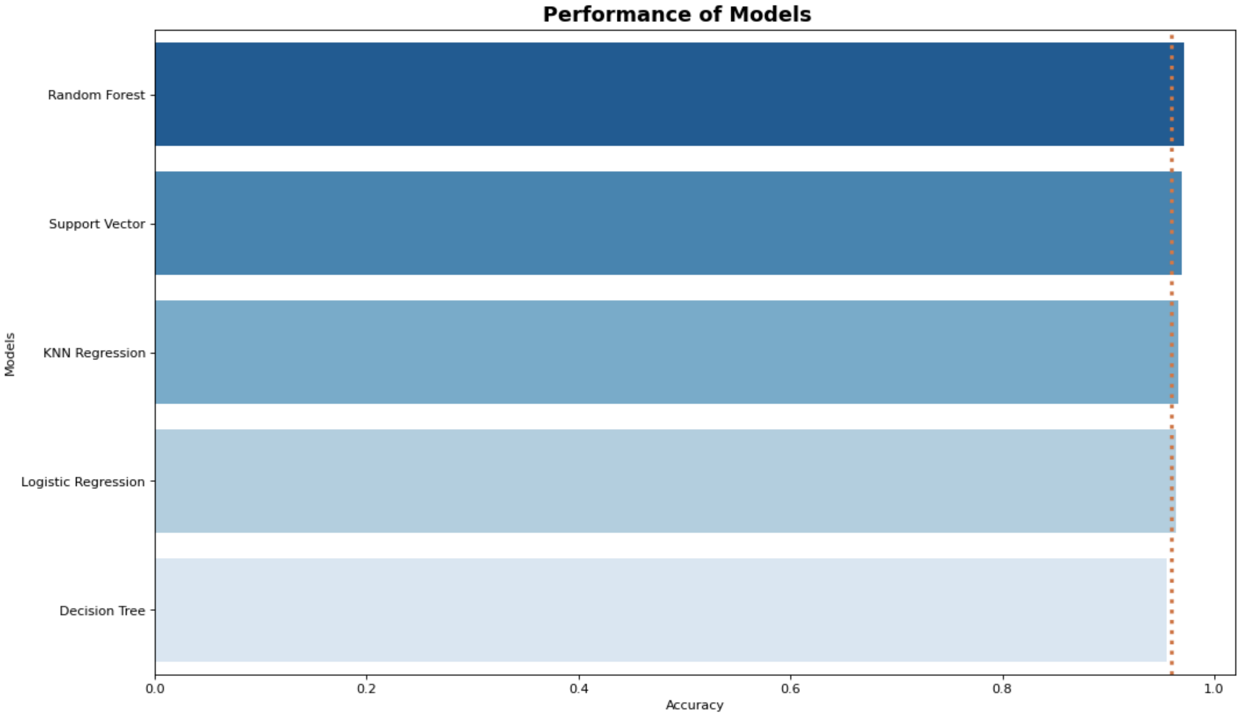


Fig.15: Comparison of algorithms.

Cross Validation

Next, cross validation was applied to the algorithms. Using the K-Fold Cross-Validation method, the consistent dataset (the dataset before train-test split) was used to be split into k number of subsets, where k-1 subsets are used to train the models and the last subset is kept for validation to test the models. The scores of each fold are then averaged to evaluate the overall performance of each model. Cross-validation using 10-folds, where 9 folds were used for training and 1 used for testing, resulted higher accuracy for in most of the algorithms.

Algorithm	Mean Accuracy Score	Standard Deviation
Random Forest	97.16 %	0.18%
KNN	96.05%	0.21 %
Decision Tree	95.73%	0.29%
SVM	95.14%	0.34%

Table-8: Mean Accuracy and Std Deviation after Cross Validation

Random Forest Performance Results

Since Random Forest (RF) has the highest mean accuracy score after cross validation, we returned to the iteration of the RF model to produce a Receiver Operating Characteristic Curve (ROC Curve) graph to visualize the model's performance with respects to their classification threshold levels. The ROC Curve plots the True Positive Rate (recall) against the False Positive Rate (type II error).

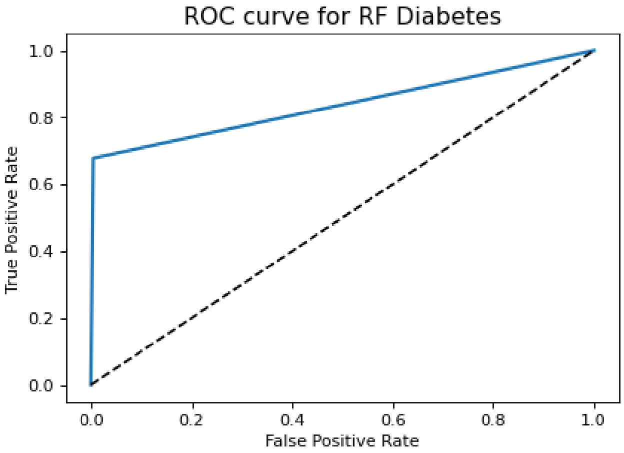


Fig.15: ROC curve for Random Forest (RF) model

Further, area under the curve (ROC AUC) can be calculated which will allow to understand the classifier's performance numerically in comparison with as a perfect classifier which is equal to 1.0. The ROC AUC for RF is 0.8366 and the cross validated ROC AUC is 0.9572 which is consistent with the rest of our evaluation metrics.

Codes for the results section can be accessed directly at: https://github.com/HussainM19/CIND-820-Predicting-Diabetes-using-MAchine-Learning/blob/main/model-building-and-analysis-diabetes-dataset_11102023%20Version.ipynb

Conclusions

Interpretations and Limitations

The analysis pf the result of the project gives us several implications. First of all, the overall project confirms that with machine learning techniques, diabetes can be predicted with higher accuracy. After the iteration of the models, the best performing algorithm was Random Forest classifier with an accuracy of 97.14% and F1 score of 78.07%. However, the other models also got an accuracy between 95-96% which is very close to the percentage of RF classifier. Using cross-validation, the mean accuracy of RF model slightly increased to rose 97.16% with a standard deviation of 0.18%. This indicates the model prediction performance is consistent and accurate to a greater extent.

In terms of processing time, there were no obvious delays while executing the models. This is evident as the dataset consists of 100,000 instances which is not that large. Hence, in terms of run-time evaluations, there was not enough difference among all models to determine a more efficient model in terms of efficiency.

Data preparation had been an essential stage in the methodology. The dataset had no missing values. But we removed outliers from few variables and removed smoking history variable which has approximately 35% values with no information. Eventually, these steps provided a more complete dataset for the models and increased the overall accuracy. The bivariate analysis enables us to identify that HbA1C level and blood glucose highest significance and BMI has a moderately high significance among the variables present in the dataset. These variables eventually have greater impact on predicting diabetes than the rest.

Using both the train-test split for our initial modeling and then cross validating the highest performing models gave higher confidence in final selection of the algorithm best suited for diabetes prediction. ROC AUC value for the RF model was 0.8366 and cross validated ROC AUC value was 0.9572. Therefore, we may conclude that cross validation improved the performance of the model.

Limitations

From the review of previous studies and projects on diabetes prediction by other researchers, in my opinion, this particular dataset lacks some parameters that are important for predicting diabetes. For instances, demographic information such as ethnicity, or genetical information (whether biological parents have diabetes or not), geographical location etc. particularly shapes the risk of having diabetes considerably.

Recommendations for future works

It is evident that the Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life. For further development and future use of this classification model, it would be beneficial to test this model using some other datasets that include few important features mentioned in the limitations. Also, the results of this project can be helpful in adopting other machine learning techniques such as Artificial Neural Network (ANN) models with different numbers of hidden which may give more accurate or better results.

References

Soni, M., & Varma, Dr. S. (2020, October 4). *Diabetes prediction using Machine Learning Techniques*. International Journal of Engineering Research & Technology. https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques

Rani, K. J. (2020, August 30). *Diabetes prediction using machine learning*. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://ijsrcseit.com/CSEIT206463

Khanam, J. J., & Foo, S. Y. (2021, February 20). *A comparison of machine learning algorithms for diabetes prediction*. ICT Express. https://www.sciencedirect.com/science/article/pii/S240595952100020

Oikonomou, E. K., & Khera, R. (2023, September 25). *Machine learning in precision diabetes care and cardiovascular risk prediction - cardiovascular diabetology*. SpringerLink. https://link.springer.com/article/10.1186/s12933-023-01985-3

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018, October 12). *Predicting diabetes mellitus with machine learning techniques*. Frontiers. https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

Tigga, N. P., & Garg, S. (2020, April 16). *Prediction of type 2 diabetes using machine learning classification methods*. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050920308024

Mujumdar, A., & V, Dr. V. (2020, February 27). *Diabetes prediction using machine learning algorithms*. Procedia Computer Science. https://www.sciencedirect.com/science/article/pii/S1877050920300557

Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala AA, Ogurtsova K, Shaw JE, Bright D, Williams R; IDF Diabetes Atlas Committee. *Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas*, 9th edition. Diabetes Res Clin Pract. 2019 Nov;157:107843. doi: 10.1016/j.diabres.2019.107843. Epub 2019 Sep 10. PMID: 31518657.

World Health Organization. (2023, April 5). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

Diabetes in Canada. Diabetes Canada. (n.d.). https://www.diabetes.ca/advocacy---policies/advocacy-reports/national-and-provincial-backgrounders/diabetes-in-canada

U.S. Department of Health and Human Services. (n.d.). *Diabetes, heart disease, & stroke - NIDDK*. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke#:~:text=High%20blood%20glucose%20from%20diabetes,can%20lead%20to%20heart%20disease.&text=People%20with%20diabetes%20tend%20to,age%20than%20people%20without%20diabetes.