HealthCare Analysis - Diabetes Dataset

An in-depth analysis of the Diabetes dataset, exploring various factors that contribute to the prevalence and prediction of diabetes.

INFO 570 - Project Group 8

Section 1

Introduction, Project Goals, Implications

Introduction - Diabetes

What is Diabetes?

Diabetes is a chronic condition characterized by high levels of blood sugar (glucose) due to the body's inability to produce or use insulin effectively.

Importance of Prediction

Early prediction and diagnosis of diabetes can help individuals take proactive steps to manage the condition and prevent or delay the onset of complications.

Diabetes Diagnosis

Accurate diagnosis of diabetes is crucial for effective treatment and management, including blood tests to measure glucose levels and assessing risk factors.

Diabetes Treatment

Effective treatment of diabetes involves a combination of lifestyle changes, medication, and regular monitoring to maintain healthy blood sugar levels and prevent complications.

Predicting, diagnosing, and treating diabetes are essential for improving patient outcomes and reducing the burden of this chronic condition on healthcare systems.

Business Problems

- ## Predicting Diabetes Onset

  Use the dataset to develop predictive models that can identify individuals at risk of developing diabetes based on various health and lifestyle factors.

- ## Assessing Risk Factors

  Analyze the dataset to understand the key risk factors and associated with diabetes, informing prevention and early intervention strategies.

- ## Personalizing Treatment Plans

  Leverage the dataset to create personalized treatment plans and recommendations for individuals with diabetes or those identified as at risk.

- ## Optimizing Healthcare Resource Allocation

  Use insights from the dataset to help healthcare providers and policymakers allocate resources more effectively, such as targeting high-risk populations or focusing on preventive measures.

- ## Improving Patient Outcomes

  Develop models and strategies based on the dataset to help healthcare professionals better monitor, manage, and improve the health outcomes of individuals with diabetes.

Project Goals and Implications

Project Goals

The primary goal of this project is to develop a model that can accurately predict the risk of diabetes based on various medical and laboratory parameters. This model will help healthcare professionals identify individuals at high risk of developing diabetes, enabling early intervention and preventive measures...

Implications

Early identification of individuals at risk of diabetes can lead to proactive lifestyle modifications, regular monitoring, and timely treatment, ultimately reducing the burden of diabetes-related complications and healthcare costs.

Section 2

Dataset

Dataset Description

Dataset Overview

The dataset was collected from the Iraqi society, specifically from the laboratory of Medical City Hospital and the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital. Patient files were used to extract relevant medical information..

Dataset Variables

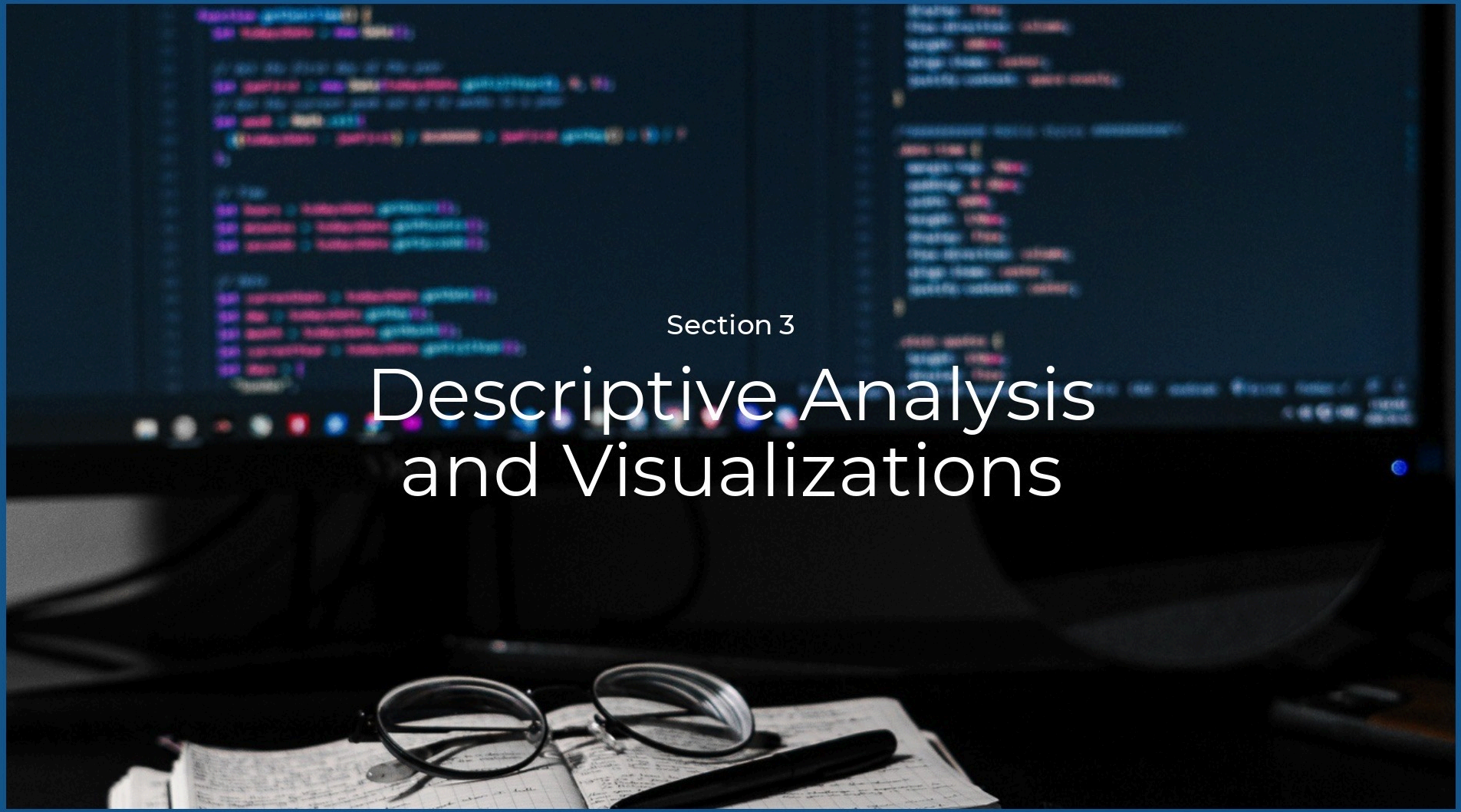
The dataset includes dataset include ID, No. of Patient, Age, Gender, Creatinine ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile (including total, LDL, VLDL, Triglycerides (TG), and HDL Cholesterol), and HBA1C.

Output Variable

The output variable is 'Class', which represents the patient's diabetes disease class (Diabetic, Non-Diabetic, or Predict-Diabetic).

Data Size

The dataset consists of 13995 records



Section 3

Descriptive Analysis and Visualizations

Descriptive Statistics

The dataset consists of 14 columns and 1,000 rows, providing information about various medical and laboratory parameter related to diabetes. There were 5 missing values in the output column "Class", representing .5% of the values in the column. No other values in any other columns were missing.

Descriptive statistics:

	ID	No_Pation	Gender	AGE	Urea
count	1000.000000	1.000000e+03	1000.000000	1000.000000	1000.000000
mean	340.500000	2.705514e+05	0.565000	53.528000	5.124743
std	240.397673	3.380758e+06	0.496005	8.799241	2.935165
min	1.000000	1.230000e+02	0.000000	20.000000	0.500000
25%	125.750000	2.406375e+04	0.000000	51.000000	3.700000
50%	300.500000	3.439550e+04	1.000000	55.000000	4.600000
75%	550.250000	4.538425e+04	1.000000	59.000000	5.700000
max	800.000000	7.543566e+07	1.000000	79.000000	38.900000

	Cr	HbA1c	Chol	TG	HDL
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	68.943000	8.281160	4.862820	2.349610	1.204750
std	59.984747	2.534003	1.301738	1.401176	0.660414
min	6.000000	0.900000	0.000000	0.300000	0.200000
25%	48.000000	6.500000	4.000000	1.500000	0.900000
50%	60.000000	8.000000	4.800000	2.000000	1.100000
75%	73.000000	10.200000	5.600000	2.900000	1.300000
max	800.000000	16.000000	10.300000	13.800000	9.900000

	LDL	VLDL	BMI
count	1000.000000	1000.000000	1000.000000
mean	2.609790	1.854700	29.578020
std	1.115102	3.663599	4.962388
min	0.300000	0.100000	19.000000
25%	1.800000	0.700000	26.000000
50%	2.500000	0.900000	30.000000
75%	3.300000	1.500000	33.000000
max	9.900000	35.000000	47.750000

Missing values percentage:

ID	0.0
No_Pation	0.0
Gender	0.0
AGE	0.0
Urea	0.0
Cr	0.0
HbA1c	0.0
Chol	0.0
TG	0.0
HDL	0.0
LDL	0.0
VLDL	0.0
BMI	0.0
Class	0.5

dtype: float64

Correlation Matrix

The variable 'HbA1c' has the highest correlation with the 'Class' variable, indicating a strong positive relationship. This suggests that 'HbA1c' is likely to be an important predictor of diabetes. Other variables that have a relatively higher correlation with 'Class' include 'AGE' and 'BMI'

Correlation Matrix:

	ID	No_Pation	Gender	AGE	Urea	Cr
ID	1.000000	0.064840	0.016605	-0.060606	-0.094569	-0.103068
No_Pation	0.064840	1.000000	0.051977	-0.088837	-0.018968	0.000732
Gender	0.016605	0.051977	1.000000	0.029524	0.116189	0.155104
AGE	-0.060606	-0.088837	0.029524	1.000000	0.107177	0.055786
Urea	-0.094569	-0.018968	0.116189	0.107177	1.000000	0.624252
Cr	-0.103068	0.000732	0.155104	0.055786	0.624252	1.000000
HbA1c	-0.009334	-0.032621	-0.008212	0.383224	-0.021599	-0.036998
Chol	0.046287	-0.030176	-0.063230	0.034076	0.001807	-0.006978
TG	-0.054960	-0.040026	0.054170	0.145852	0.041747	0.056836
HDL	0.028104	-0.013228	-0.129758	-0.022007	-0.037577	-0.023730
LDL	-0.063631	-0.003092	0.054320	0.016028	-0.007418	0.039512
VLDL	0.144420	0.114998	0.189987	-0.070108	-0.010498	0.010579
BMI	0.051121	0.017640	0.069910	0.386418	0.047118	0.055939
Class	-0.056995	-0.047957	0.105021	0.447322	0.068194	0.038369

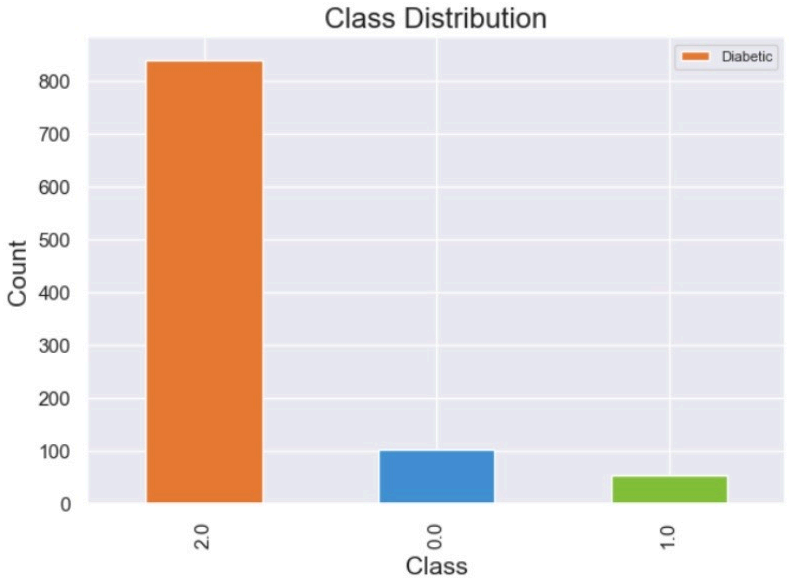
	HbA1c	Chol	TG	HDL	LDL	VLDL
ID	-0.009334	0.046287	-0.054960	0.028104	-0.063631	0.144420
No_Pation	-0.032621	-0.030176	-0.040026	-0.013228	-0.003092	0.114998
Gender	-0.008212	-0.063230	0.054170	-0.129758	0.054320	0.189987
AGE	0.383224	0.034076	0.145852	-0.022007	0.016028	-0.070108
Urea	-0.021599	0.001807	0.041747	-0.037577	-0.007418	-0.010498
Cr	-0.036998	-0.006978	0.056836	-0.023730	0.039512	0.010579
HbA1c	1.000000	0.178072	0.217905	0.029994	0.011817	0.071413
Chol	0.178072	1.000000	0.321233	0.103041	0.417075	0.079197
TG	0.217905	0.321233	1.000000	-0.083333	0.015510	0.149241
HDL	0.029994	0.103041	-0.083333	1.000000	-0.142350	-0.059465
LDL	0.011817	0.417075	0.015510	-0.142350	1.000000	0.064012
VLDL	0.071413	0.079197	0.149241	-0.059465	0.064012	1.000000
BMI	0.414106	0.013944	0.111510	0.072275	-0.068073	0.189284
Class	0.555931	0.168110	0.182597	-0.002399	0.004642	0.098799

	BMI	Class
ID	0.051121	-0.056995
No_Pation	0.017640	-0.047957
Gender	0.069910	0.105021
AGE	0.386418	0.447322
Urea	0.047118	0.068194
Cr	0.055939	0.038369
HbA1c	0.414106	0.555931
Chol	0.013944	0.168110
TG	0.111510	0.182597
HDL	0.072275	-0.002399
LDL	-0.068073	0.004642
VLDL	0.189284	0.098799
BMI	1.000000	0.570376
Class	0.570376	1.000000

Hemoglobin A1c (HbA1c), also known as glycated hemoglobin, is a blood test that measures a person's average blood sugar levels over the past 2–3 months.

BMI - Body mass index is a value derived from the mass and height of a person

Distribution

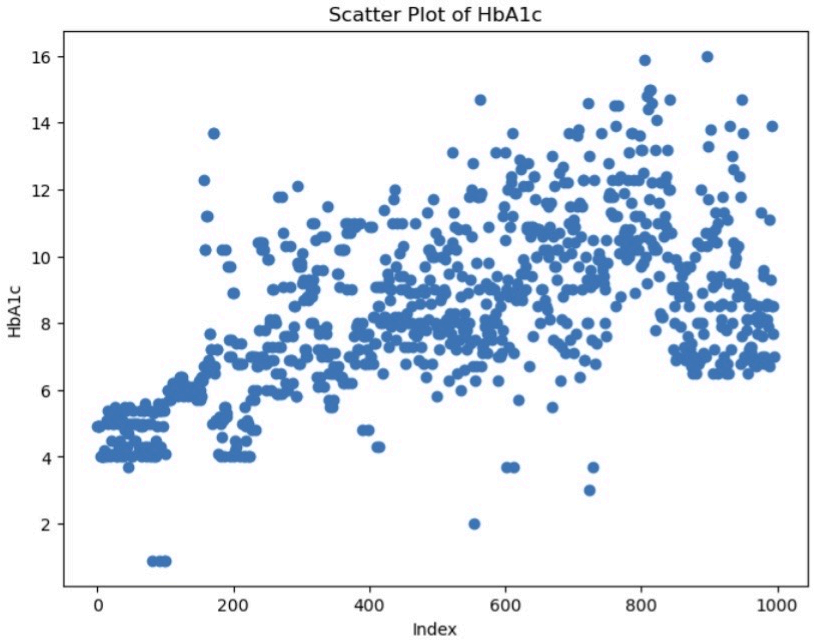



```
Class
2.0    840
0.0    102
1.0     53
Name: count, dtype: int64
```

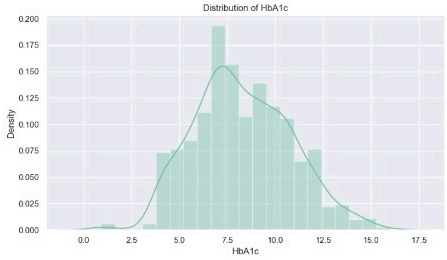
The above visualization indicates that our training dataset is imbalanced. The number of diabetic patients significantly exceeds non diabetes or pre-diabetic patients.

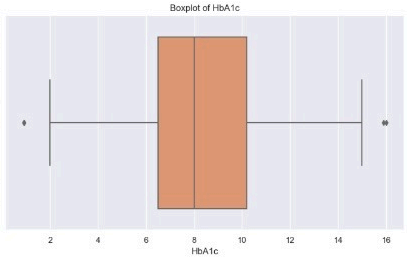
HbA1c Scatter Plot

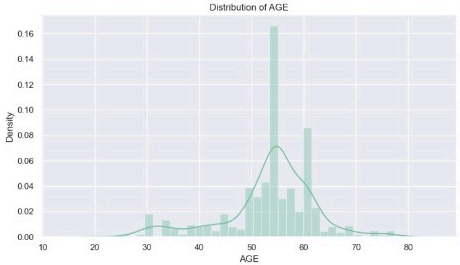
The scatter plot of HbA1c visualizes the distribution of HbA1c values across the dataset. Each point on the plot represents an individual record, with the x-axis representing the index (row number) of the record and the y-axis representing the corresponding HbA1c value. The scatter plot indicates that values are distributed normally.

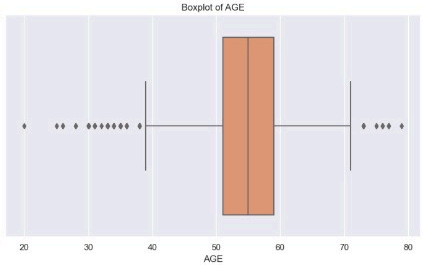


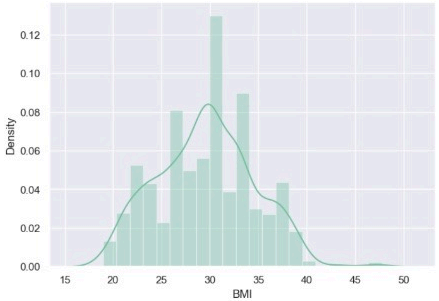
Distribution and Outliers

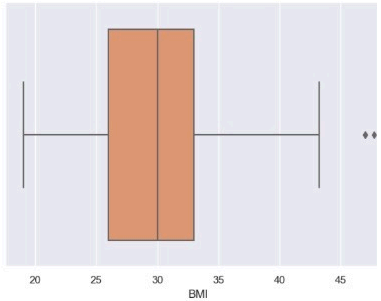








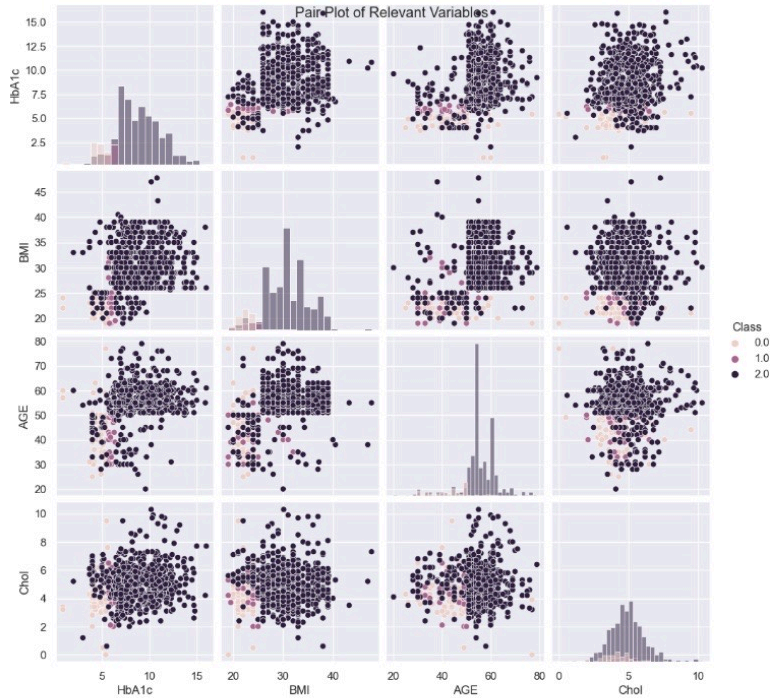




Plot and Box Plot identify the distribution and outliers for the 3 variables with the highest correlation to Class - HbA1c, BMI, and Age

Data Visualization - Pair Plot

The pair plot visualizes the relationships between the selected variables ('HbA1c', 'BMI', 'Age', 'Chol', ') and the target variable 'Class'.



From the pair plot, we can observe the following: HbA1c vs Class: There is a clear separation between the different classes based on the HbA1c values. Higher HbA1c values are associated with the diabetic class (2.0), while lower values are more prevalent in the non-diabetic class (0.0). This confirms the positive correlation between HbA1c and the diabetes class. AGE vs Class: There is a slight trend indicating that older individuals are more likely to be in the diabetic class (2.0). There is a tendency for higher BMI values in the diabetic class (2.0) compared to the non-diabetic class (0.0). The relationship between Chol and Class is not visually apparent in the pair plot.

Overall, the pair plot confirms that HbA1c is the strongest predictor of the diabetes class among the selected variables. It shows a clear separation between the classes based on HbA1c values.

Data Visualization - Heat map

The correlation heatmap provides a visual representation of the pairwise correlations between the variables. It helps identify the strength and direction of the relationships among the variables. The color intensity and the values in each cell indicate the magnitude and direction of the correlation.

Section 4

Data Processing

Data Processing - Procedures, Model Development, Evaluation

Missing values

There were 5 missing values in the dataset in the "Class" column, with rows removed.

X and Y Variable

The target variable 'Class' and the input features were separated into 'X' (input variables) and 'y' (target variable) for model training and evaluation.. ID and The Patient Number Columns were dropped from the model.

Training and Testing Datasets

The dataset was split into training and testing sets using the train_test_split function from scikit-learn, with 20% of the data reserved for testing.

4 Different Models were trained and evaluated

Logistic Regression - used to predict binary outcome

Decision Tree - combines multiple decision trees

Random Forest - combines the output of multiple decision trees to reach a single result.

In addition, we also trained the following models:

KNN - can be effective at capturing complex interactions among variables without having to define a separable statistical mode

Assessment and evaluation

The performance of each model was assessed using various evaluation metrics:

Accuracy: Measures the overall correctness of the model's predictions.

Precision: Indicates the proportion of true positive predictions among all positive predictions.

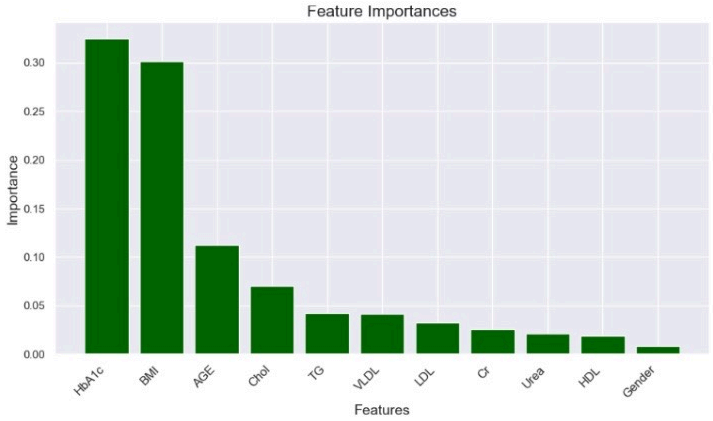
Recall: Represents the proportion of true positive predictions among all actual positive instances.

F1 Score: Provides a balanced measure of precision and recall

Features Importances

The feature importance analysis using Random Forest Classifier shows that 'HbA1c' is the most important feature for predicting the 'Class' variable. This aligns with the findings from the correlation matrix. Other variables that have relatively higher importance scores include 'AGE', 'BMI', 'Chol', and 'TG'. Based on the correlation analysis and feature importance, the variables that are most likely to predict diabetes ('Class') are: HbA1c and BMI

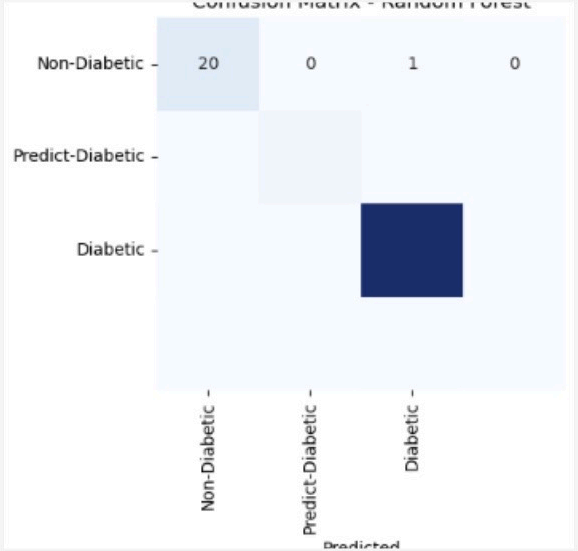
```
Feature Importances:
    Feature  Importance
4    HbA1c    0.324747
10     BMI    0.301125
1      AGE    0.112784
5     Chol    0.070396
6       TG    0.042113
9     VLDL    0.041584
8      LDL    0.032798
3       Cr    0.025508
2     Urea    0.020963
7      HDL    0.019152
0   Gender    0.008830
```



These variables have higher correlations with the 'Class' variable and higher importance scores in the Random Forest Classifier.

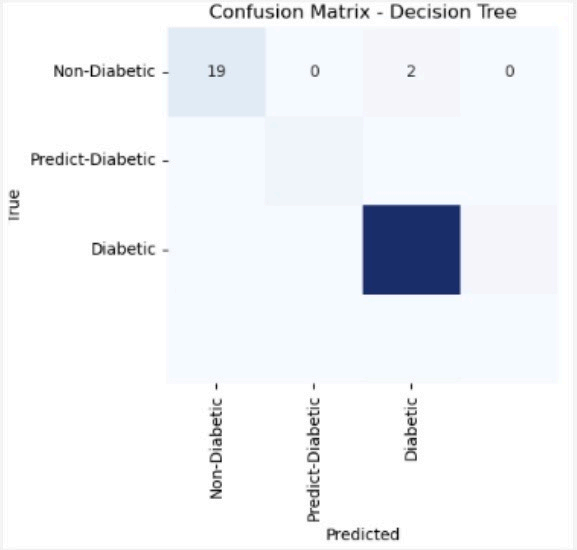
Logistical Regression, Random Forest, Decision Tree - Confusion Matrix

The confusion matrices show the models' performance in correctly classifying instances of each class.



Confusion matrix illustrating the performance of a Random Forest model

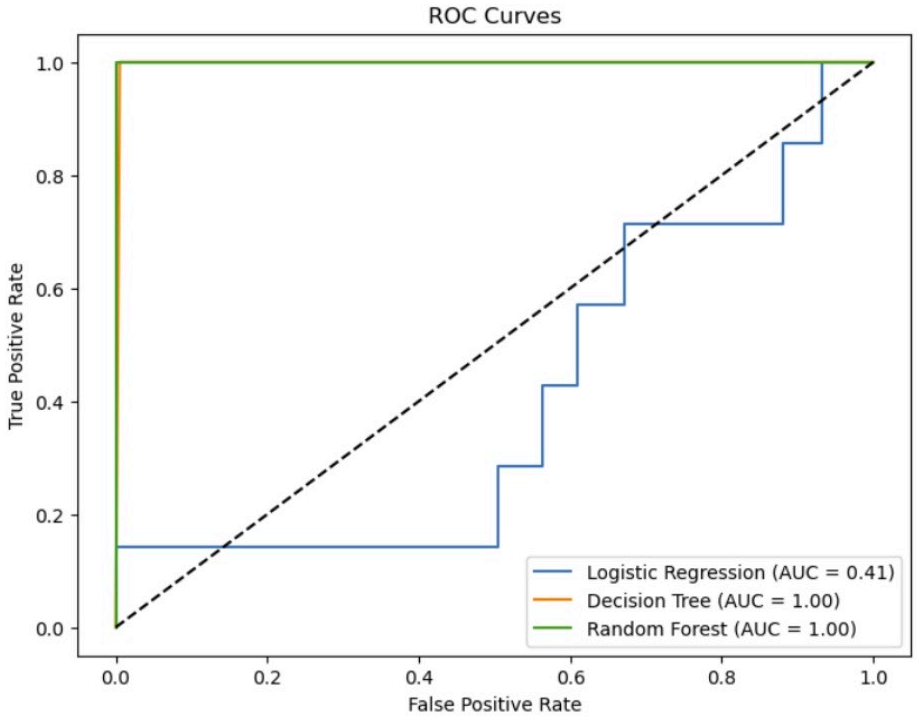
Confusion matrix illustrating the Logistic Regression model



Confusion matrix illustrating the Decision Tree Model

Model - Performance Visualization

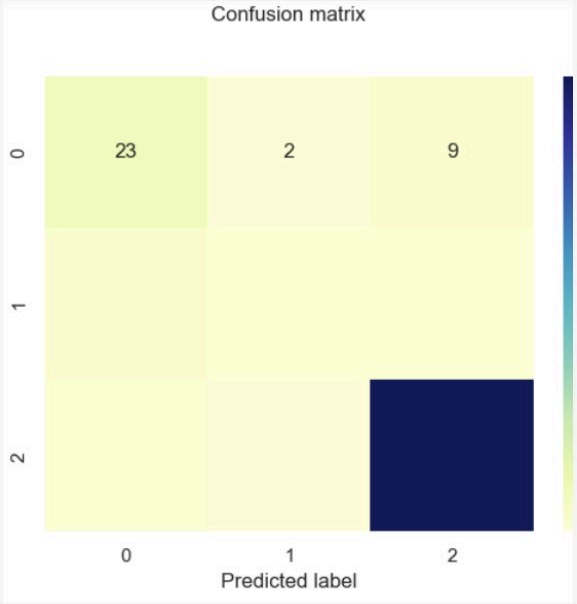
ROC curves - Random Forest, Decision Tree, Logistic Regression



The ROC curves for logistic regression, decision tree, and random forest models provide valuable insights into each model's performance in discriminating between classes at different thresholds. Random forests, however, typically outperform both by combining the strengths of multiple decision trees, resulting in higher overall accuracy and robustness.

KNN Model

The confusion matrices show the models' performance in correctly classifying instances of each class.



Confusion matrix illustrating the performance of a KNN model in predicting True Positives, False Negatives, False

	precision	recall	f1-score	support
0.0	0.62	0.68	0.65	34
1.0	0.50	0.28	0.36	18
2.0	0.95	0.97	0.96	280
accuracy			0.90	332
macro avg	0.69	0.64	0.65	332
weighted avg	0.89	0.90	0.89	332

F1 score for KNN for diabetes prediction is .96

Predicted / True	0.0	1.0	2.0	All
0.0	23	2	9	34
1.0	8	5	5	18
2.0	6	3	271	280
All	37	10	285	332

Predictions using classifier with K neighbors

Evaluation - Logistic Regression, Random Forest, Decision Tree, KNN

Accuracy, Precision, Recall, and F1-Score on the test dataset - The evaluation results showed that the Decision Tree model performed the best among the three models, achieving the highest accuracy, precision, recall, and F1 score. The best-performing model was selected based on the highest F1 score, which in this case was the Decision Tree model. (Best model is based on outcome to 4 decimal places)

Logistic Regression

Accuracy: 0.86

Precision: 0.74

Recall: 0.86

F1 Score: 0.79

Random Forest

Accuracy: 0.99

Precision: 0.99

Recall: 0.99

F1 Score: 0.99

Decision Tree

Accuracy: 0.99

Precision: 0.99

Recall: 0.99

F1 Score: 0.99

KNN

Accuracy: .90

Precision: .95

Recall: .97

F1 Score: .96

Section 5

Insights and Conclusion

Insights

Result

HbA1c is the best predictor for diagnosing the diabities, followed by BMI, with Random Forrest as the best model which offers several advantages over other models. This algorithm demonstrated superior performance metrics, including higher accuracy, precision, recall, and F1-score.

Other Variables

To help identify the risk factor that causes higher levels HbA1c we need access to data that has variables that influence higher levels of glucose - such as stress, gestational diabetes, family history, level of physical activity.

HbA1 and BMI

BMI and HbA1 are highly correlated. BMI and HbA1c are correspond due to their relationship with glucose metabolism and insulin resistance. Higher BMI, often indicative of obesity, is associated with increased insulin resistance, which can lead to elevated blood glucose levels. Over time, consistently high blood glucose levels are reflected in elevated HbA1c values.

Age and BMI

People's body mass index tends to go up with age. As people age, their body composition changes. Typically, there is an increase in body fat and a decrease in muscle mass. This shift contributes to a higher BMI, as the body stores more fat. The trend of increasing BMI with age is influenced by physiological changes, lifestyle factors, and socio-environmental influences.

Conclusion - Diabetes Predictive Models

Machine learning techniques for disease prediction

Powerful methods for analyzing health indicators and lifestyle factors to predict diabetes risk

Approach selection considerations

Dataset size, model interpretability, and complexity of relationships impact the choice of method

Advancements in predictive modeling

Integrating diverse patient datasets and exploring advanced neural networks for improved accuracy

Clinical deployment of predictive models

Enabling early intervention, risk stratification, and tailored treatment plans to improve outcomes

Diabetes predictive models powered by machine learning hold immense potential to enhance healthcare, from early risk identification to personalized treatment plans, ultimately improving patient outcomes.

Citations

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9729599/ Kowsar, R., & Mansouri, A. (2022). Multi-level analysis reveals the association between diabetes, body mass index, and HbA1c in an Iraqi population. *Scientific reports*, *12*(1), 21135. https://doi.org/10.1038/s41598-022-25813-y

- Rashid, Ahlam (2020), “Diabetes Dataset”, Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1