Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Diabetes Prediction

Supervised By:

Dr. Kiran Deep Singh

Submitted By:

Prachi, 2210990659 (G8)

Paras Kapoor ,2210990637(G8)

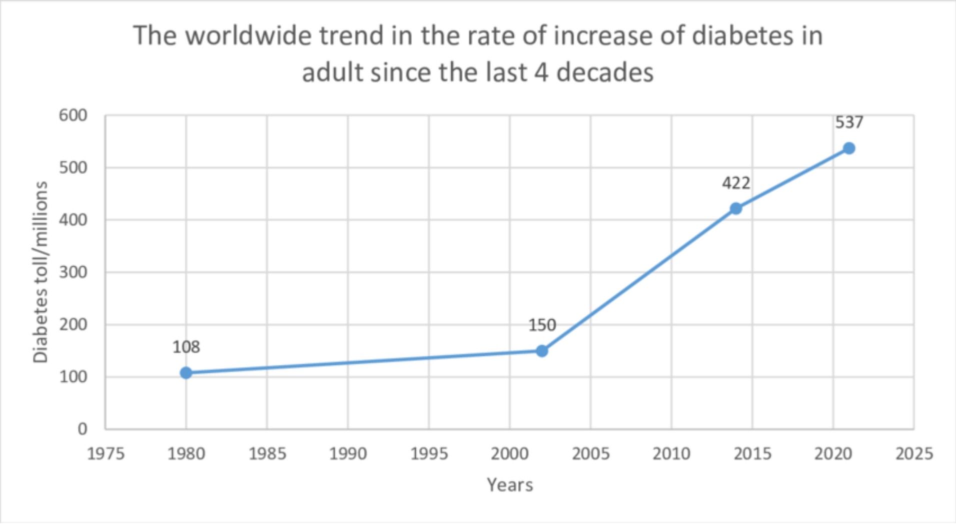
Palak Singla,2210990994(G8)

Diabetes: A Growing Health Concern

Diabetes, a chronic metabolic disorder characterised by elevated blood sugar levels, has emerged as a significant health challenge worldwide. With its prevalence steadily increasing over the past few decades, diabetes has become a major public health concern, affecting millions of people globally. In this presentation, we explore the application of machine learning in predicting diabetes, aiming to contribute to early detection and better management of this condition.

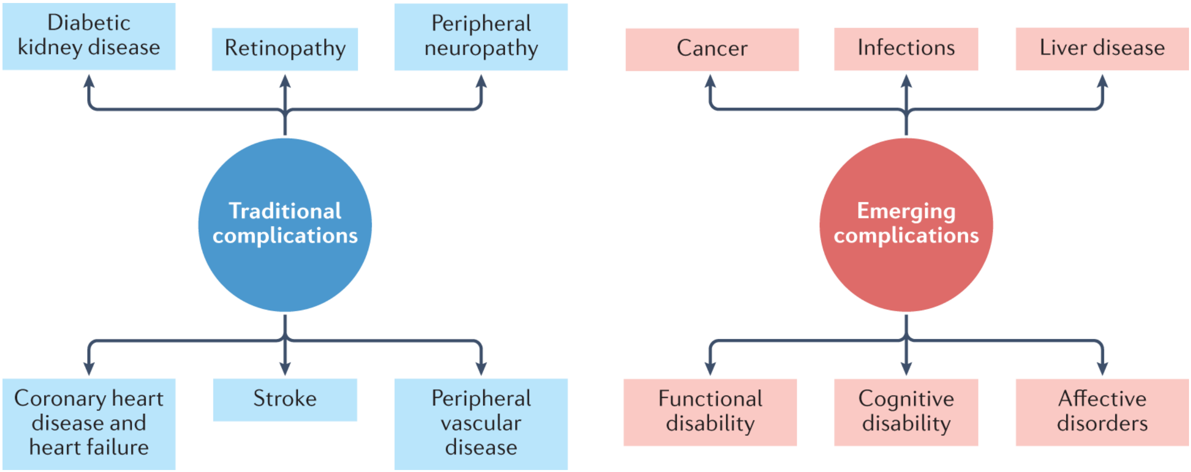
• The Magnitude of the Problem

Diabetes affects individuals of all ages, genders, and socioeconomic backgrounds. According to the World Health Organisation (WHO), the number of people with diabetes has risen from 108 million in 1980 to an alarming 422 million in 2014. This number is projected to reach 552 million by 2030, posing a substantial economic burden on healthcare systems worldwide.



• Health Implications

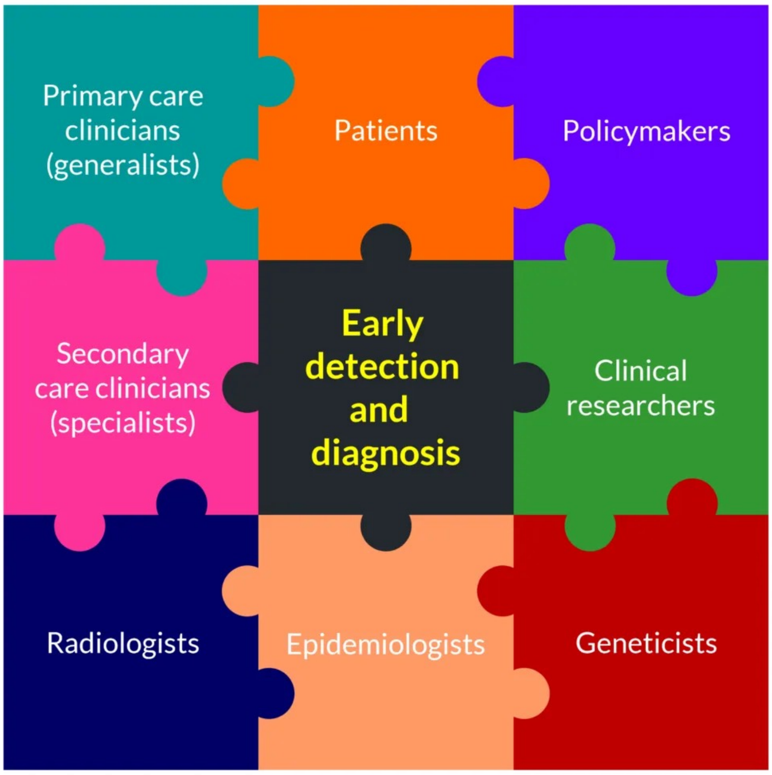
The implications of diabetes extend beyond its immediate effects on blood sugar levels. Uncontrolled diabetes can lead to various complications, including cardiovascular diseases, kidney failure, blindness, and lower limb amputation. Moreover, individuals with diabetes have a significantly higher risk of premature death compared to those without the condition.



- ## Challenges in Diagnosis and Management

One of the major challenges in tackling diabetes is its often late diagnosis. Many individuals remain undiagnosed until they develop complications, highlighting the need for effective screening and early detection methods. Additionally, managing diabetes requires continuous monitoring of blood sugar levels, medication adherence, lifestyle modifications, and regular medical check-ups, which can be cumbersome for patients.

- ## Importance of Predicting Diabetes Early


Primary care clinicians (generalists)
Patients
Policymakers
Secondary care clinicians (specialists)
Early detection and diagnosis
Clinical researchers
Radiologists
Epidemiologists
Geneticists

Early Intervention Saves Lives

Diabetes is a progressive disease that, if left undiagnosed and untreated, can lead to severe complications and significantly reduce life expectancy. Detecting diabetes in its early stages is crucial for several reasons:

1. Prevention of Complications

- Early diagnosis allows for timely intervention and effective management strategies, reducing the risk of complications such as heart disease, stroke, kidney failure, and vision loss. By controlling blood sugar levels through lifestyle changes and/or medication, individuals with diabetes can significantly reduce the risk of developing complications.

2. Improved Quality of Life

- Early diagnosis enables individuals to make lifestyle modifications, such as adopting a healthy diet, regular exercise, and weight management, which are essential for managing diabetes and improving overall well-being.

3. Cost Savings in Healthcare

- Early detection and management of diabetes can lead to substantial cost savings in healthcare expenditures. By preventing complications and reducing hospitalizations, early intervention helps lower healthcare costs associated with diabetes treatment and long-term care.

4. Empowerment Through Knowledge

- Early diagnosis empowers individuals with knowledge about their health status and allows them to take control of their condition. With awareness of their diabetes status, individuals can make informed decisions about their health and actively participate in managing their condition.

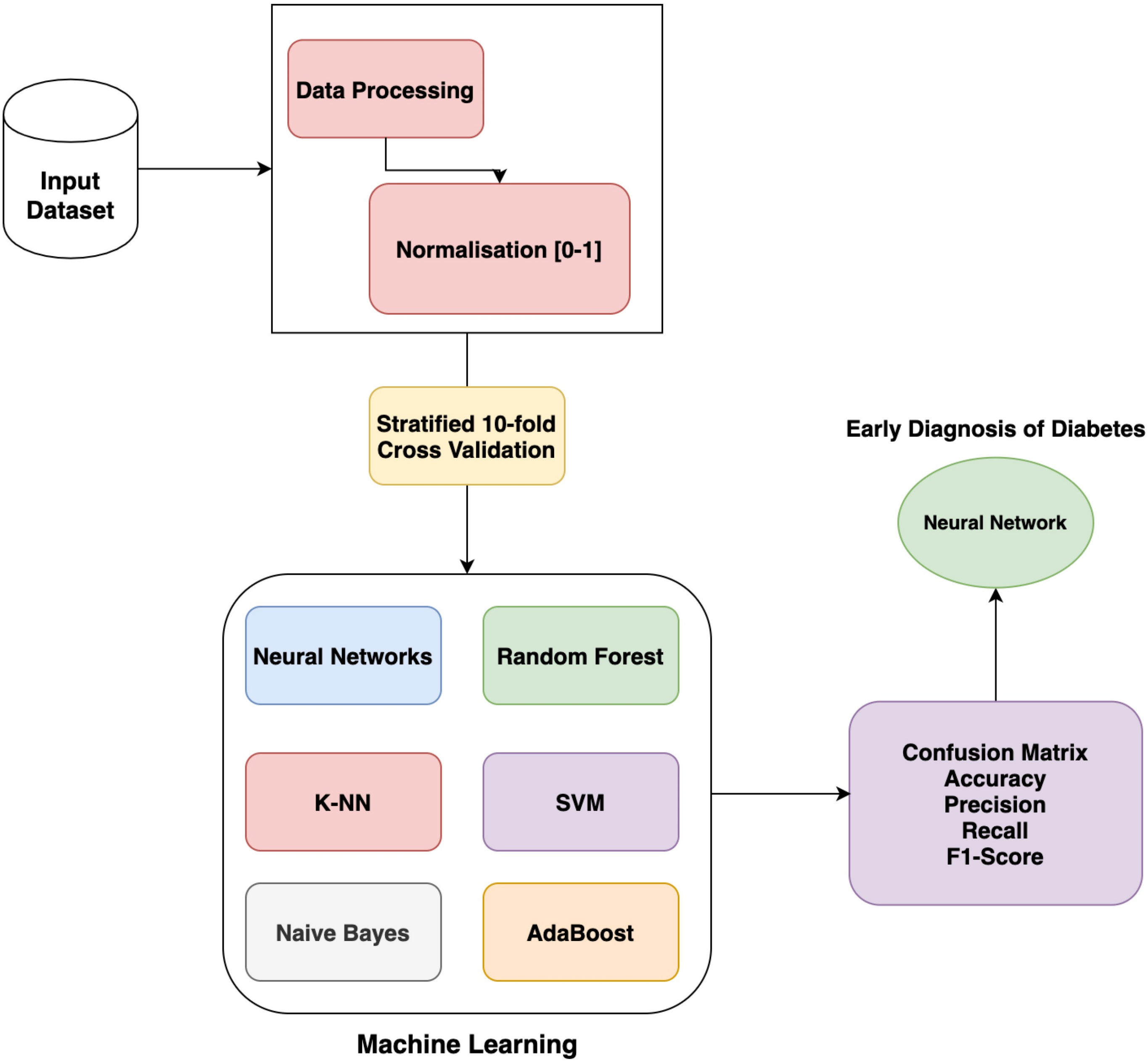
5. Public Health Impact

- Early detection of diabetes on a population level can have a significant impact on public health by reducing the overall disease burden and associated healthcare costs. By implementing screening programs and preventive measures, healthcare systems can effectively manage the rising prevalence of diabetes and its complications.

• Motivation behind building the Diabetes Prediction Model

Early detection of diabetes is critical for preventing complications, improving quality of life, reducing healthcare costs, empowering individuals, and addressing the public health challenge posed by diabetes. Predictive models developed using machine learning techniques play a vital role in identifying individuals at risk of diabetes, enabling timely interventions and personalised management strategies.Early detection of diabetes not only saves lives but also reduces the burden on healthcare systems, allowing resources to be allocated more efficiently. By leveraging machine learning techniques, we can develop accurate predictive models that aid in proactive healthcare decision-making, ultimately leading to better health outcomes for individuals and communities alike.

- **Model Architecture Diagram:-**



• Dataset Overview

Description of Dataset Features and Target Variable

In our model, we utilised a comprehensive dataset containing nearly **1,00,000 (1 lakh)** instances & a wide range of features crucial for diabetes prediction.

Below are the key features included in the dataset:

In our model, we utilised a comprehensive dataset containing nearly 100,000 instances and a wide range of features crucial for diabetes prediction. Below are the key features included in the dataset:

A. Demographic Features:

• Gender

Gender is an important demographic feature that can influence the risk of developing diabetes. While both males and females can develop diabetes, there are some differences in how the disease manifests between genders.

- Hormonal Influence: Hormonal differences between males and females can affect insulin sensitivity and glucose metabolism.

- Body Composition: Males and females tend to have different body compositions, with males typically having more muscle mass and females having more body fat.

- Lifestyle Factors: Gender-specific lifestyle factors such as dietary habits, physical activity levels, and stress management can influence the risk of diabetes.

• Age

Age is a significant demographic factor strongly associated with the risk of developing diabetes.

- Insulin Resistance: As individuals age, they often become more resistant to insulin, leading to higher blood sugar levels. This insulin resistance is a key factor in the development of type 2 diabetes, which is more common in older adults.

- Changes in Body Composition: With age, there is a natural decline in muscle mass and an increase in body fat, especially visceral fat. This shift in body composition contributes to insulin resistance and a higher risk of diabetes.

- Lifestyle Factors: Age is often associated with changes in lifestyle, including dietary habits, physical activity levels, and stress levels. Sedentary behaviour and unhealthy dietary patterns become more prevalent with age, further increasing the risk of diabetes.

B. Health Indicators:

• Hypertension (0: No, 1: Yes)

Hypertension, commonly known as high blood pressure, plays a crucial role in detecting diabetes due to its close association with the disease. Here's how hypertension influences diabetes detection:

- Insulin Resistance and Endothelial Dysfunction:

Insulin Resistance: Hypertension is associated with insulin resistance, a condition where cells fail to respond effectively to insulin. Insulin resistance is a precursor to type 2 diabetes and is characterised by elevated blood glucose levels.

Endothelial Dysfunction: Hypertension damages the endothelium, the inner lining of blood vessels, leading to endothelial dysfunction. This impairs the ability of blood vessels to dilate and regulate blood flow, further exacerbating insulin resistance.

- Shared Risk Factors and Pathophysiology:

Obesity: Both hypertension and type 2 diabetes are strongly linked to obesity. Excess body weight, particularly abdominal obesity, increases the risk of developing both conditions.

Inflammation and Oxidative Stress: Chronic inflammation and oxidative stress contribute to the development of both hypertension and diabetes. These processes promote insulin resistance, impair vascular function, and contribute to the pathogenesis of both diseases.

• Heart Disease (0: No, 1: Yes)

- Dyslipidemia: Abnormal lipid levels, including high levels of LDL cholesterol and triglycerides, and low levels of HDL cholesterol, are associated with both heart disease and diabetes.

- Inflammation: Chronic inflammation is a key driver of both heart disease and diabetes. Inflammatory markers such as C-reactive protein (CRP) are elevated in individuals with heart disease and are associated with insulin resistance and diabetes.

• Smoking History

- Disruption of Glucose Homeostasis: Smoking disrupts glucose homeostasis by increasing blood glucose levels and impairing insulin secretion from the pancreas.

- Smoking contributes to insulin resistance, a key factor in the development of type 2 diabetes. Insulin resistance impairs the body's ability to use insulin effectively, leading to elevated blood sugar levels

C. Physical Measurements:

• BMI(Body Mass Index)

- Metabolic Syndrome Component: Obesity, as indicated by elevated BMI, is a key component of metabolic syndrome. Metabolic syndrome is a cluster of conditions, including abdominal obesity, hypertension, dyslipidemia, and impaired glucose metabolism, that significantly increase diabetes risk.

- Increased Risk with Higher BMI: Individuals classified as overweight or obese (BMI ≥ 25 kg/m^2) have a significantly higher risk of developing type 2 diabetes compared to those with a normal BMI.

• HbA1c level(Haemoglobin A1c)

Haemoglobin A1c (HbA1c) is a measure of average blood glucose levels over the past 2 to 3 months and is an important diagnostic and monitoring tool for diabetes. Here's how HbA1c level influences the prediction of diabetes:

- Indicator of Glucose Control: It measures the percentage of haemoglobin that is glycated, or bound to glucose, providing an indication of long-term glucose control.

- Diagnostic Tool: HbA1c is one of the diagnostic criteria for diabetes. According to the American Diabetes Association (ADA), a HbA1c level of 6.5% or higher indicates diabetes. Levels between 5.7% and 6.4% indicate pre-diabetes.

- Screening Tool: HbA1c can be used as a screening tool to identify individuals at risk of diabetes or pre-diabetes, especially when fasting blood glucose tests are not feasible.

• Blood Glucose level

- Fasting Blood Glucose Test: Elevated fasting blood glucose levels (measured after an overnight fast) are indicative of impaired glucose tolerance or diabetes. According to the American Diabetes Association (ADA), a fasting blood glucose level of 126 mg/dL (7.0 mmol/L) or higher on two separate occasions indicates diabetes.

- Progression to Diabetes: Individuals with pre-diabetes often have elevated blood glucose levels, indicating impaired glucose regulation. Regular monitoring of blood glucose levels can help identify individuals at risk of progressing to type 2 diabetes.

Each feature provides valuable insights into an individual's health status, lifestyle, and predisposition to diabetes. These features are utilised to predict the target variable, which indicates whether an individual has diabetes or not.

Target Variable: Diabetes

The target variable, diabetes, serves as the focal point of our prediction model. It is a binary variable with the following categories:

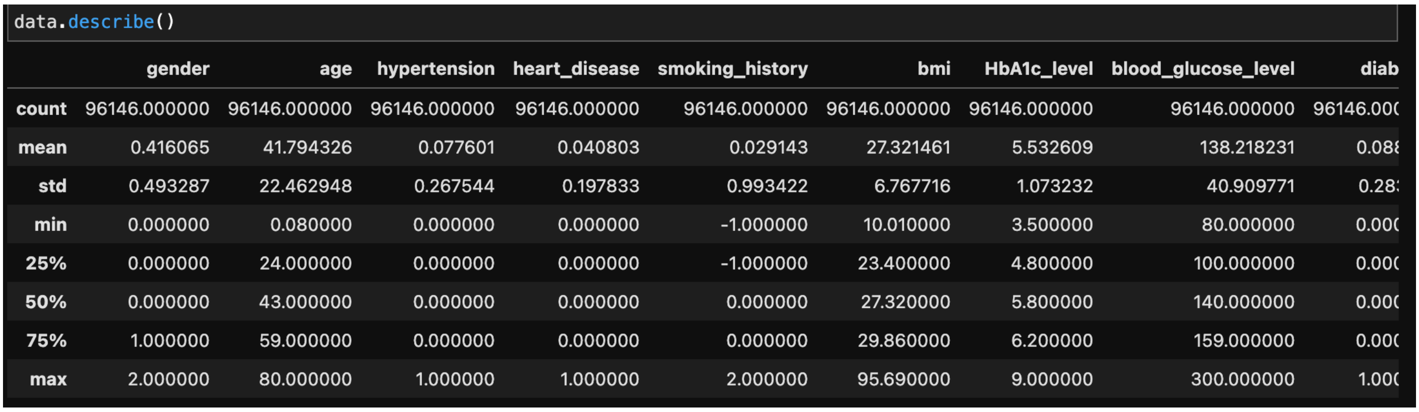
- 0: Absence of Diabetes
- 1: Presence of Diabetes

In our prediction model, accurately classifying individuals into these categories enables early identification of diabetes, facilitating timely interventions and personalised management strategies to improve health outcomes and quality of life.

Exploratory Data Analysis (EDA)

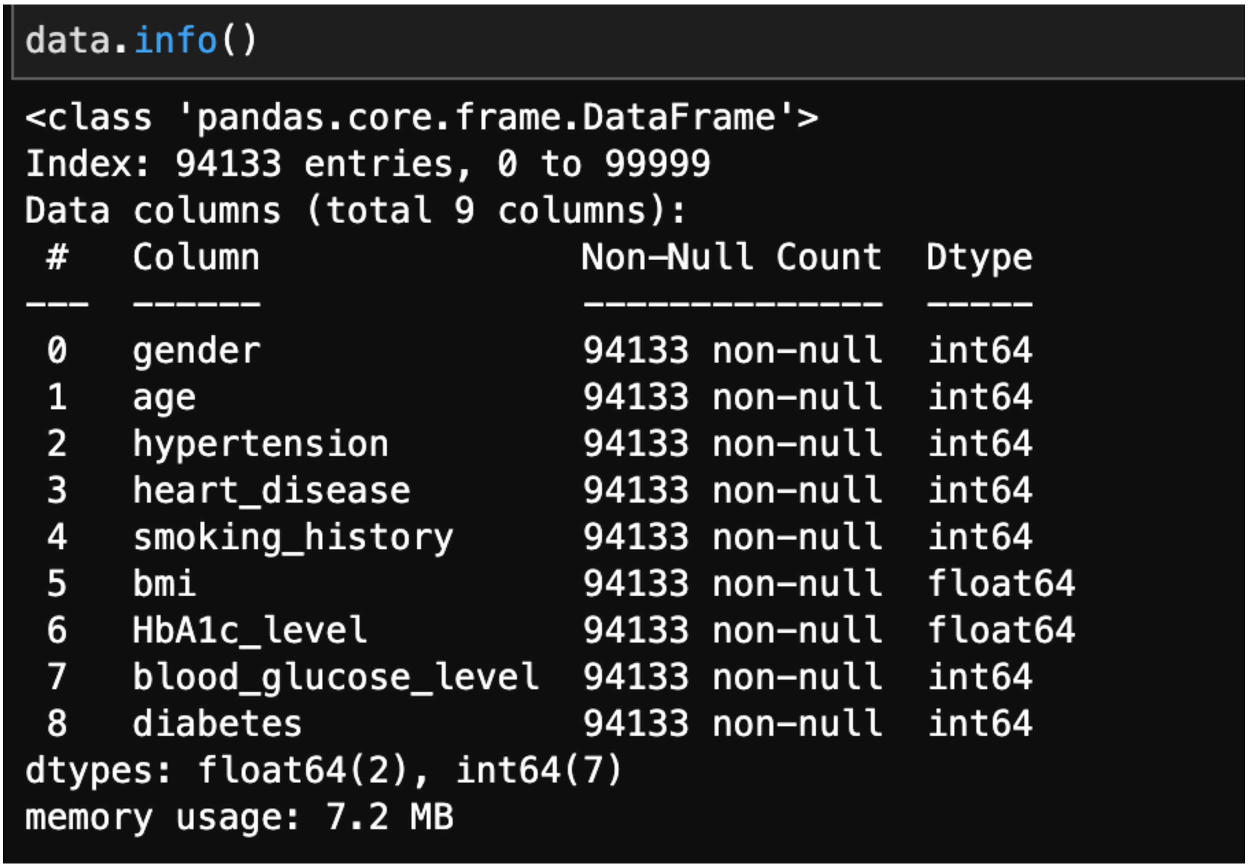
It is a crucial step in any data science project, serving as the cornerstone for understanding the dataset's characteristics, uncovering patterns, and gaining insights that drive subsequent analysis and modelling decisions. In our investigation of the diabetes dataset, our primary goal is to delve into its intricacies, exploring the relationships between variables, detecting anomalies, and identifying trends that could inform our predictive modelling efforts.

Summary Statistics:



```
data.describe()
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diab
count	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.000000	96146.00
mean	0.416065	41.794326	0.077601	0.040803	0.029143	27.321461	5.532609	138.218231	0.08
std	0.493287	22.462948	0.267544	0.197833	0.993422	6.767716	1.073232	40.909771	0.28
min	0.000000	0.080000	0.000000	0.000000	-1.000000	10.010000	3.500000	80.000000	0.00
25%	0.000000	24.000000	0.000000	0.000000	-1.000000	23.400000	4.800000	100.000000	0.00
50%	0.000000	43.000000	0.000000	0.000000	0.000000	27.320000	5.800000	140.000000	0.00
75%	1.000000	59.000000	0.000000	0.000000	0.000000	29.860000	6.200000	159.000000	0.00
max	2.000000	80.000000	1.000000	1.000000	2.000000	95.690000	9.000000	300.000000	1.00

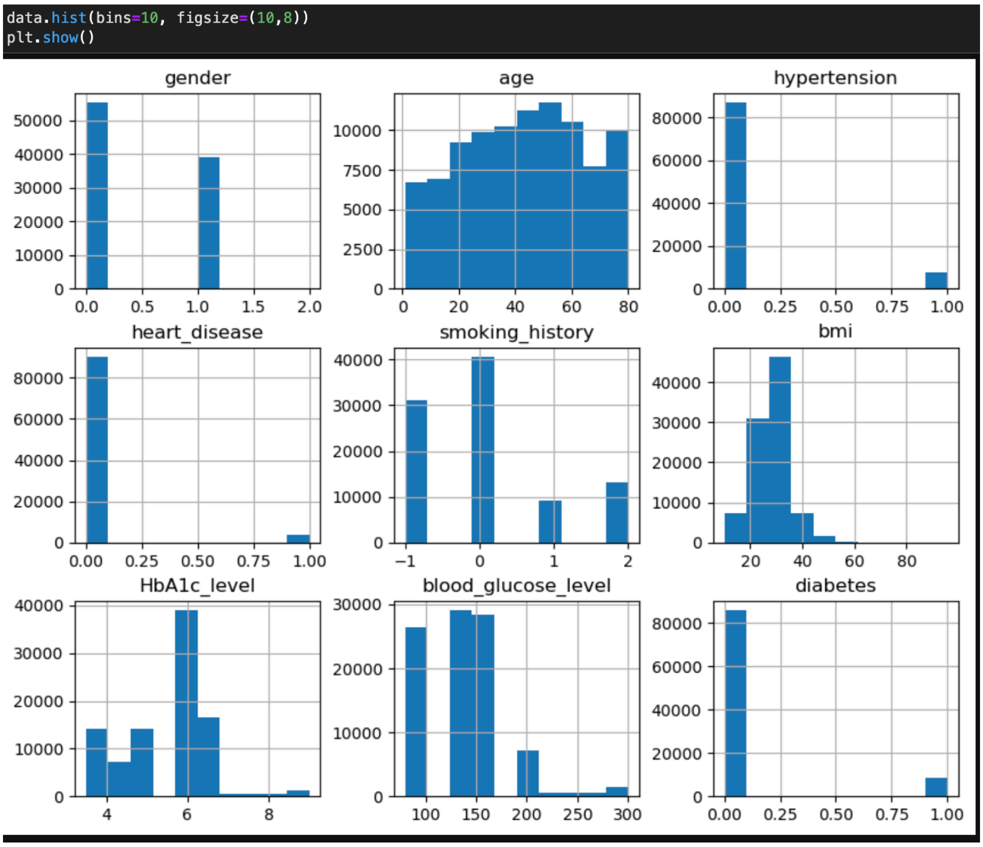


```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 94133 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   gender               94133 non-null  int64
 1   age                  94133 non-null  int64
 2   hypertension         94133 non-null  int64
 3   heart_disease        94133 non-null  int64
 4   smoking_history      94133 non-null  int64
 5   bmi                  94133 non-null  float64
 6   HbA1c_level          94133 non-null  float64
 7   blood_glucose_level  94133 non-null  int64
 8   diabetes             94133 non-null  int64
dtypes: float64(2), int64(7)
memory usage: 7.2 MB
```

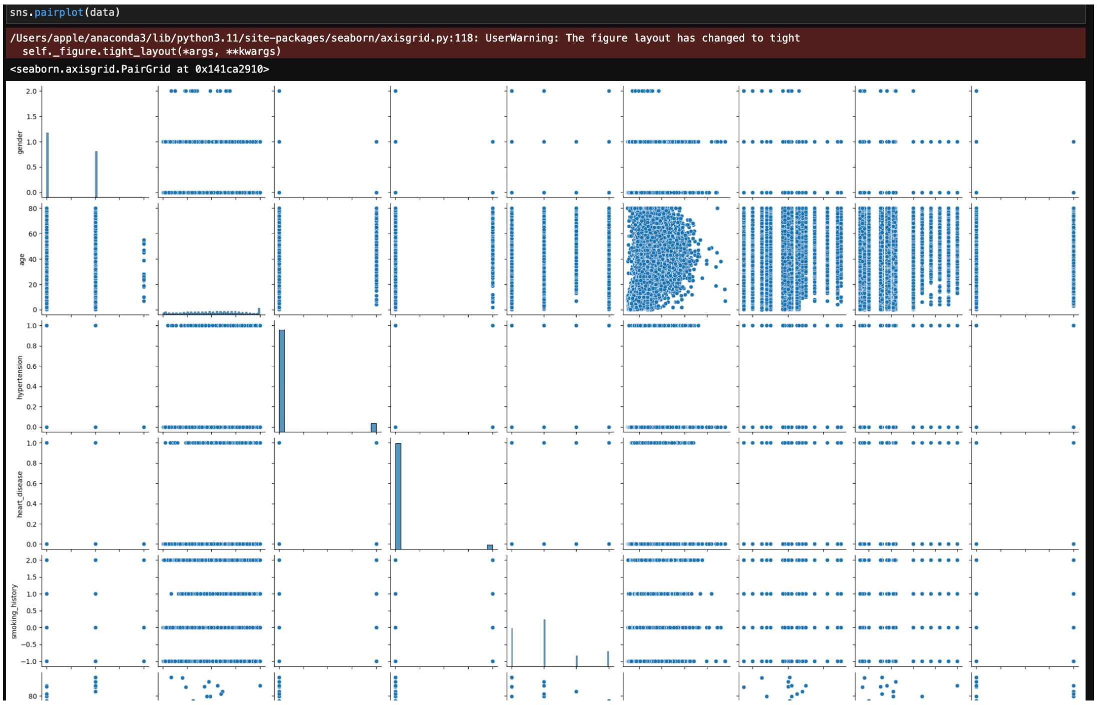
Data Visualisation:

- Histograms



```
data.hist(bins=10, figsize=(10,8))
plt.show()
```

- Pairplot

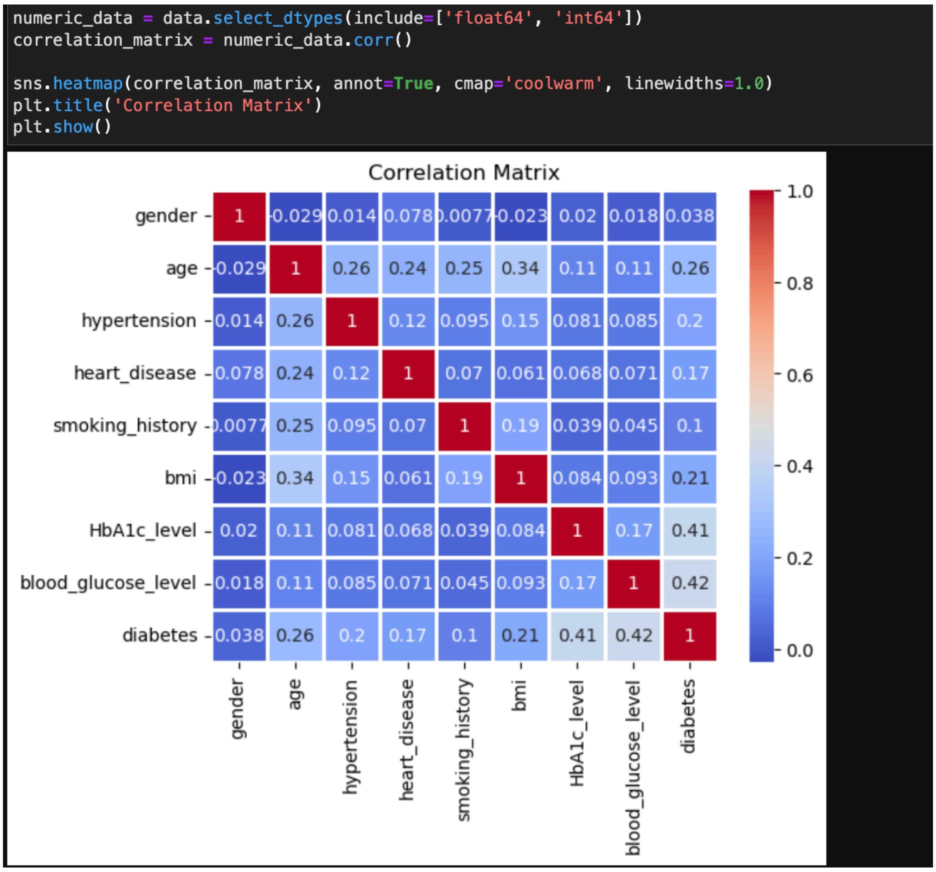


```
sns.pairplot(data)
```

```
/Users/apple/anaconda3/lib/python3.11/site-packages/seaborn/axisgrid.py:118: UserWarning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

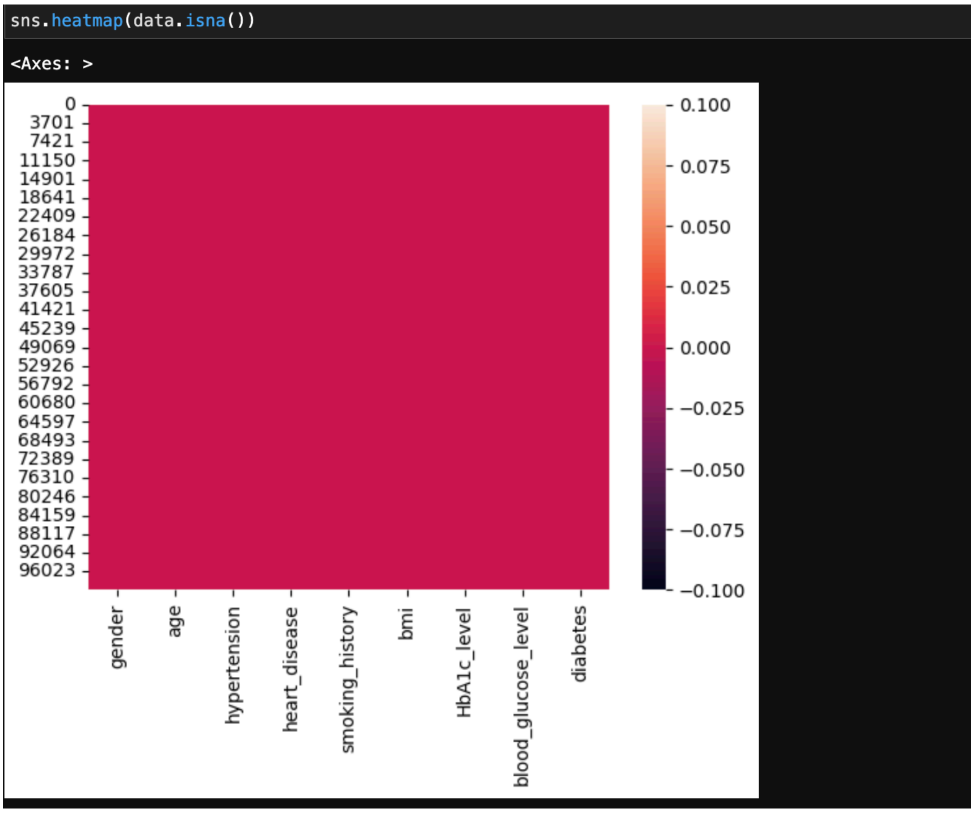
```
<seaborn.axisgrid.PairGrid at 0x141ca2910>
```

- Correlation Matrix



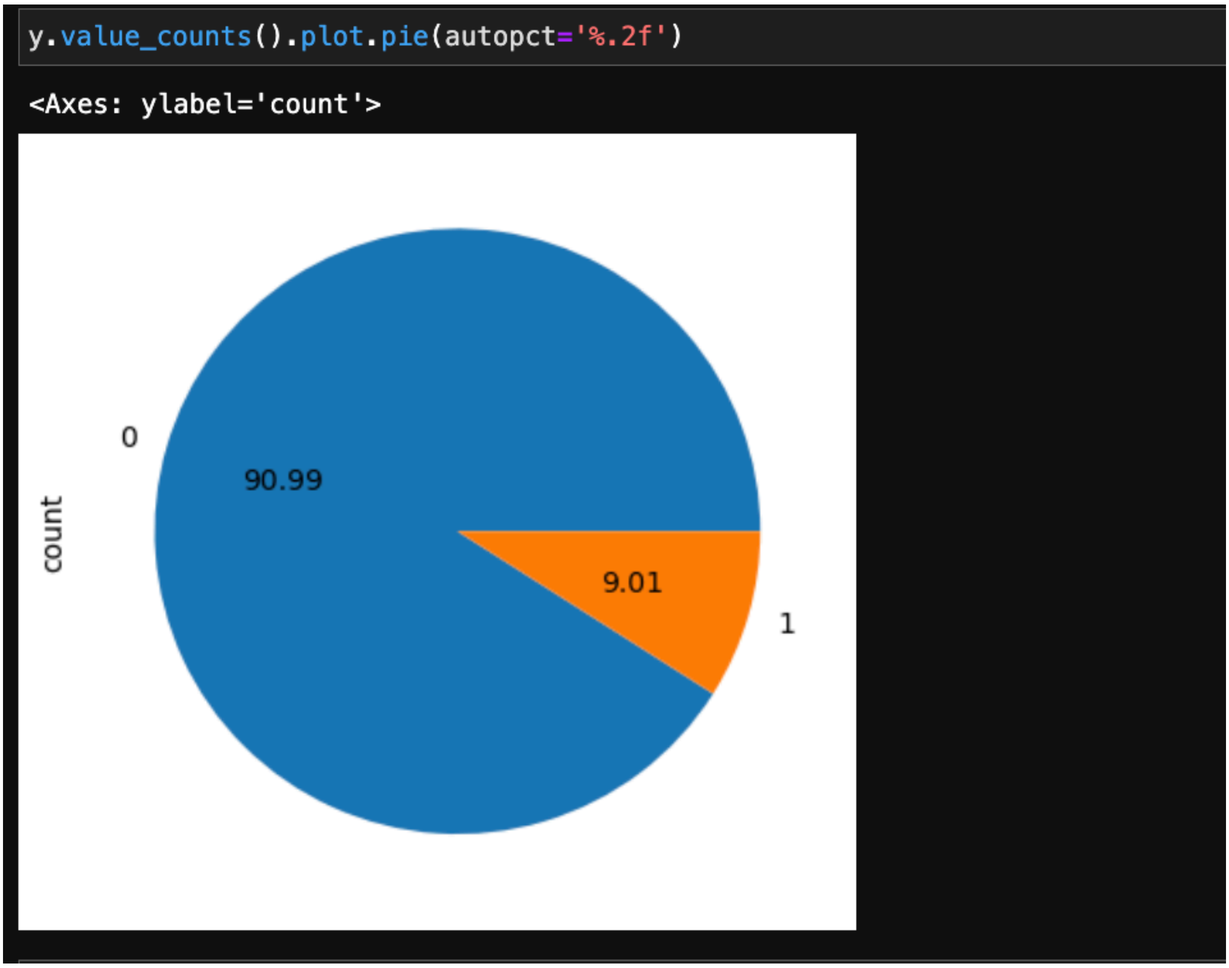
```
numeric_data = data.select_dtypes(include=['float64', 'int64'])
correlation_matrix = numeric_data.corr()

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=1.0)
plt.title('Correlation Matrix')
plt.show()
```

- HeatMap



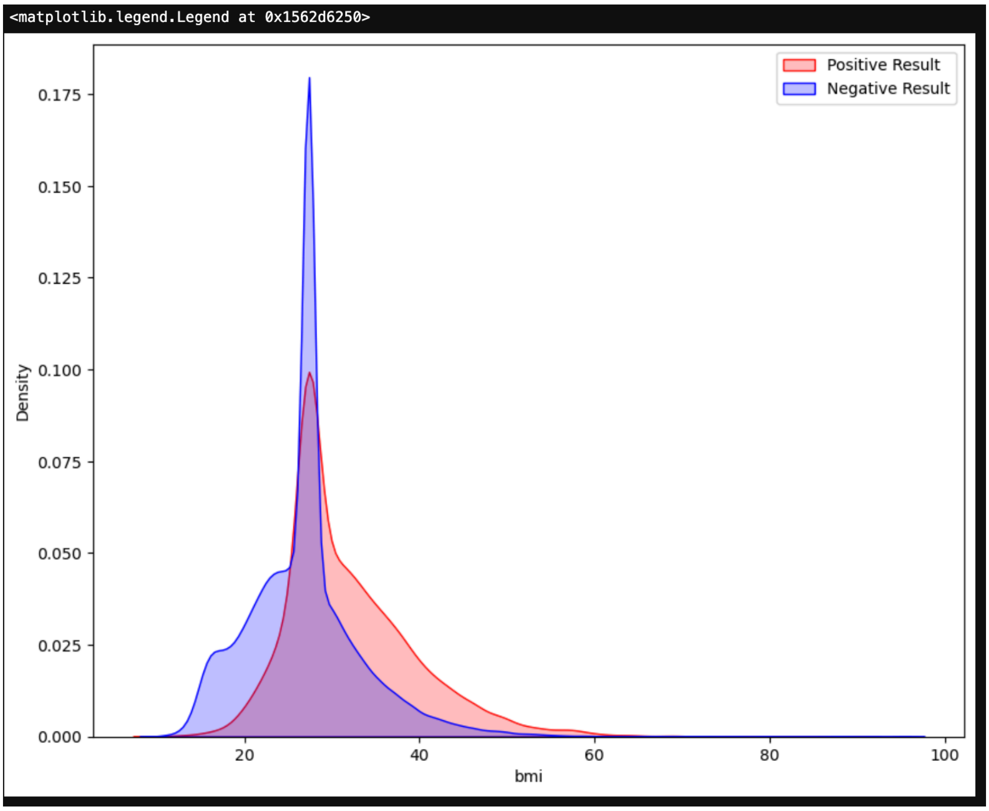
```
sns.heatmap(data.isna())
```

```
<Axes: >
```



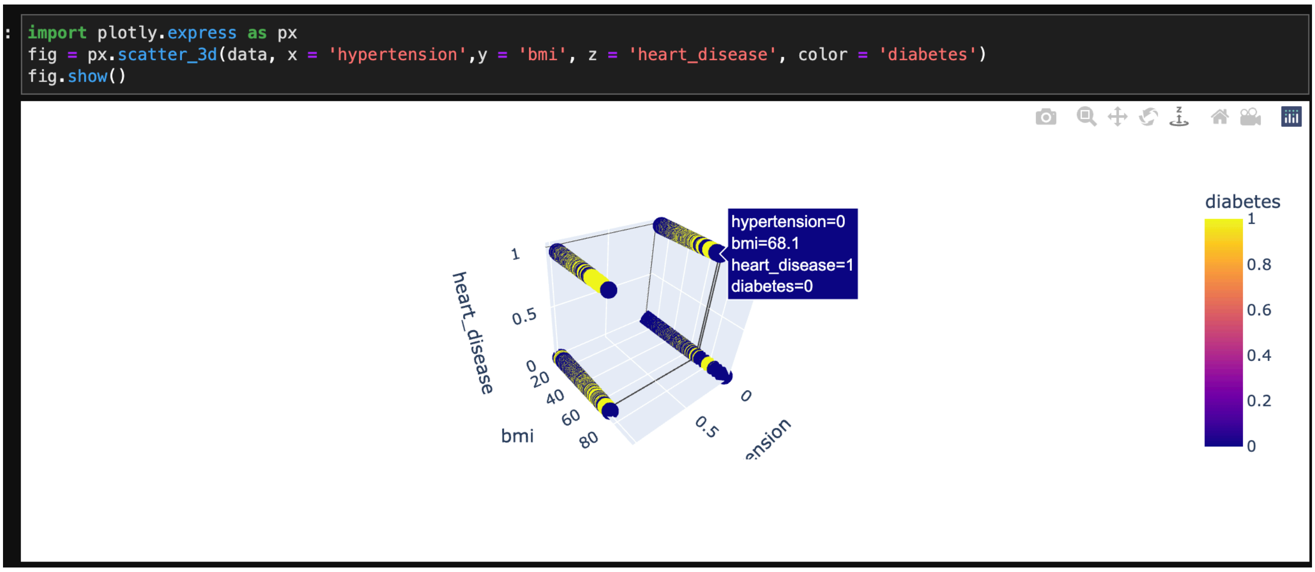
```
y.value_counts().plot.pie(autopct='%.2f')
```

```
<Axes: ylabel='count'>
```

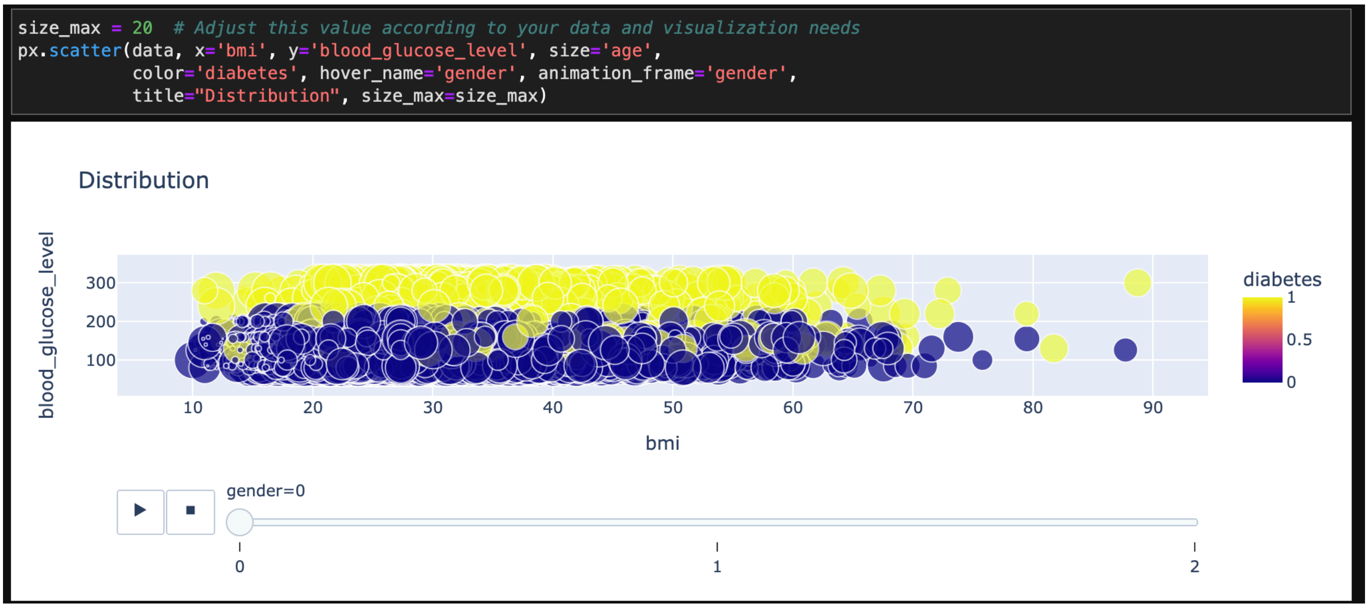


```
<matplotlib.legend.Legend at 0x1562d6250>
```

Interactive Visualisation:



```
import plotly.express as px
fig = px.scatter_3d(data, x = 'hypertension',y = 'bmi', z = 'heart_disease', color = 'diabetes')
fig.show()
```



```
size_max = 20  # Adjust this value according to your data and visualization needs
px.scatter(data, x='bmi', y='blood_glucose_level', size='age',
           color='diabetes', hover_name='gender', animation_frame='gender',
           title="Distribution", size_max=size_max)
```

Model Selection

Description of algorithms considered

- Logistic Regression

Logistic Regression is a linear model used for binary classification, estimating the probability of a binary outcome based on one or more predictor variables.

- KNeighborsClassifier

K-Nearest Neighbors (KNN) Classifier: A non-parametric algorithm that classifies new data points based on the majority class of their k nearest neighbors in the feature space.

- Support Vector Machine (SVM)

It is a powerful algorithm for classification tasks, aiming to find the optimal hyperplane that separates classes in the feature space with the maximum margin.

- GaussianNB

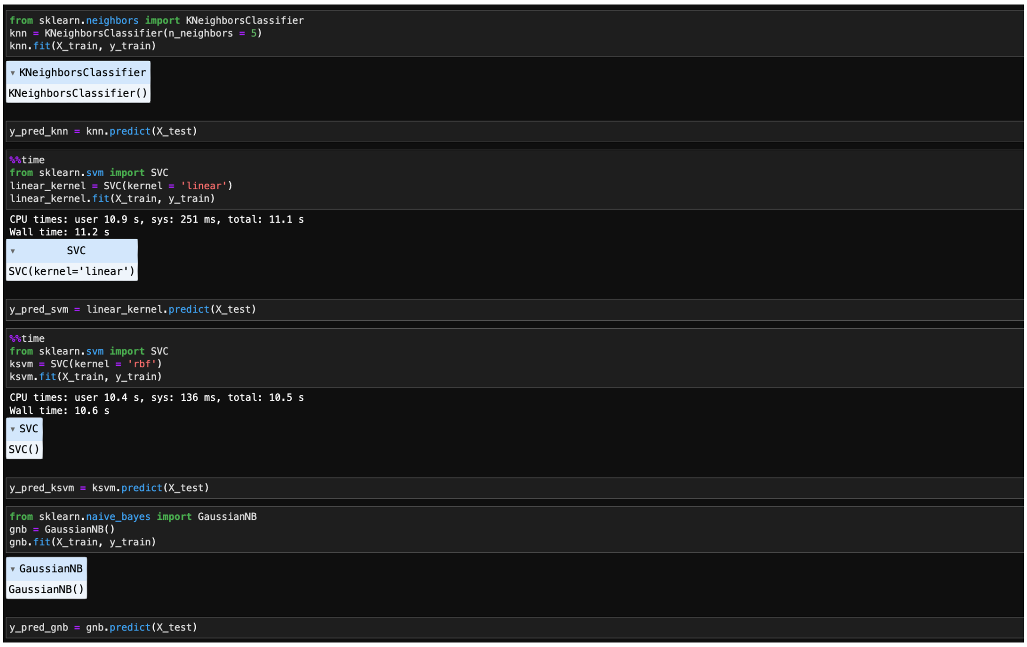
Gaussian Naive Bayes (GaussianNB) is a probabilistic classifier based on Bayes' theorem, assuming that features follow a Gaussian distribution.

- DecisionTreeClassifier

DecisionTreeClassifier is a machine learning algorithm that builds a decision tree model to classify instances based on their features.

- RandomForestClassifier

RandomForestClassifier is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.



```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 5)
knn.fit(X_train, y_train)
```

```
▾ KNeighborsClassifier
KNeighborsClassifier()
```

```
y_pred_knn = knn.predict(X_test)
```

```
%%time
from sklearn.svm import SVC
linear_kernel = SVC(kernel = 'linear')
linear_kernel.fit(X_train, y_train)
```

```
CPU times: user 10.9 s, sys: 251 ms, total: 11.1 s
Wall time: 11.2 s
▾        SVC
SVC(kernel='linear')
```

```
y_pred_svm = linear_kernel.predict(X_test)
```

```
%%time
from sklearn.svm import SVC
ksvm = SVC(kernel = 'rbf')
ksvm.fit(X_train, y_train)
```

```
CPU times: user 10.4 s, sys: 136 ms, total: 10.5 s
Wall time: 10.6 s
▾ SVC
SVC()
```

```
y_pred_ksvm = ksvm.predict(X_test)
```

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train, y_train)
```

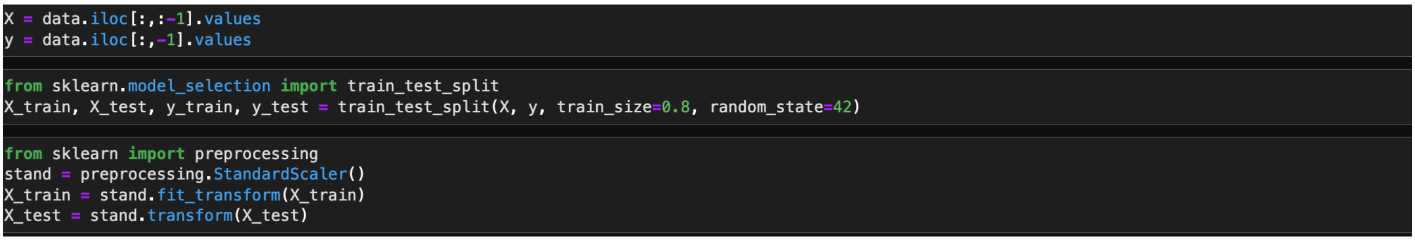
```
▾ GaussianNB
GaussianNB()
```

```
y_pred_gnb = gnb.predict(X_test)
```

Model Training

During model training, the dataset is split into two subsets: a training set and a testing set, using the train_test_split function. The training set is used to train the model, while the testing set is used to evaluate its performance.

By evaluating the model on unseen data, we can detect issues like overfitting or under-fitting and fine-tune the model accordingly, ensuring it achieves the best possible performance on new data.



```
X = data.iloc[:,:-1].values
y = data.iloc[:,-1].values

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=42)

from sklearn import preprocessing
stand = preprocessing.StandardScaler()
X_train = stand.fit_transform(X_train)
X_test = stand.transform(X_test)
```

SMOTE (Synthetic Minority Over-sampling Technique)

It is a powerful technique used to address class imbalance in machine learning datasets. It works by generating synthetic samples of the minority class to balance the class distribution.

Usage of SMOTE:

1. Identifying Imbalance: Before using SMOTE, it's crucial to identify class imbalance in the dataset. This typically involves checking the distribution of the target variable (e.g., diabetes/non-diabetes) and determining if there's a significant class imbalance.

2. **Importing Libraries**: Import necessary libraries such as `imbalanced-learn` to utilize SMOTE. For example:
   ```python
   from imblearn.over_sampling import SMOTE
   ```

3. Applying SMOTE:
   - Instantiate the SMOTE object with optional parameters such as `sampling_strategy` to control the balance between classes.
   - Use `fit_resample()` method to apply SMOTE to the training data.
   ```python
   smote = SMOTE(sampling_strategy=1)
   ```

```
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

4. Resulting Dataset:
   - After applying SMOTE, the minority class is oversampled, and the dataset becomes balanced.
   - The number of samples in the minority class is increased to match the majority class, creating synthetic samples based on the existing minority samples.
   - Ensure to use the resampled data for training the model.

5. Model Training:
   - Train the machine learning model using the resampled data.

```python
model.fit(X_train_resampled, y_train_resampled)
```

6. Model Evaluation:
   - Evaluate the model's performance on the original, imbalanced test data to assess its effectiveness in handling class imbalance.

```python
y_pred = model.predict(X_test)
```

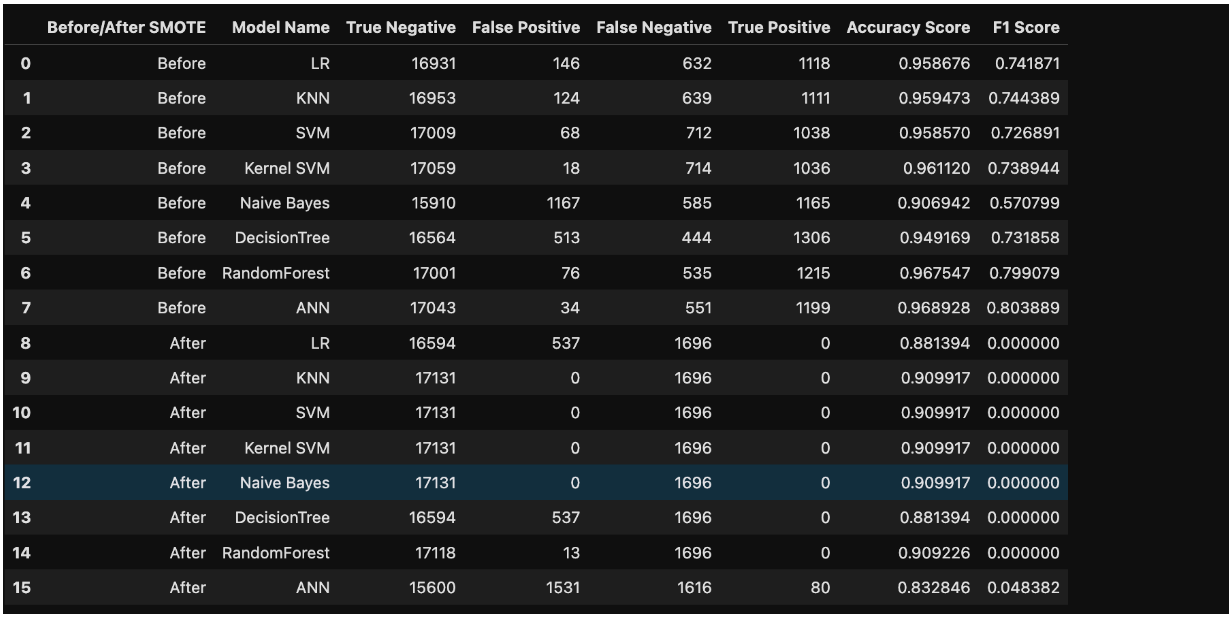
Elaboration:

SMOTE is particularly useful when dealing with imbalanced datasets, where one class significantly outnumbers the other. In the context of diabetes prediction, for instance, if the dataset has a majority of non-diabetic cases and relatively fewer diabetic cases, SMOTE can help balance this distribution by generating synthetic examples of the minority class (diabetic cases) to match the number of non-diabetic cases.

By oversampling the minority class, SMOTE helps prevent the model from being biased towards the majority class, leading to more accurate and reliable predictions, especially for the minority class. However, it's essential to apply SMOTE only to the training data to avoid data leakage and ensure the model's ability to generalize to unseen data.

Overall, SMOTE is a valuable tool for improving model performance in scenarios with imbalanced classes, enhancing the model's ability to detect and predict minority class instances accurately.

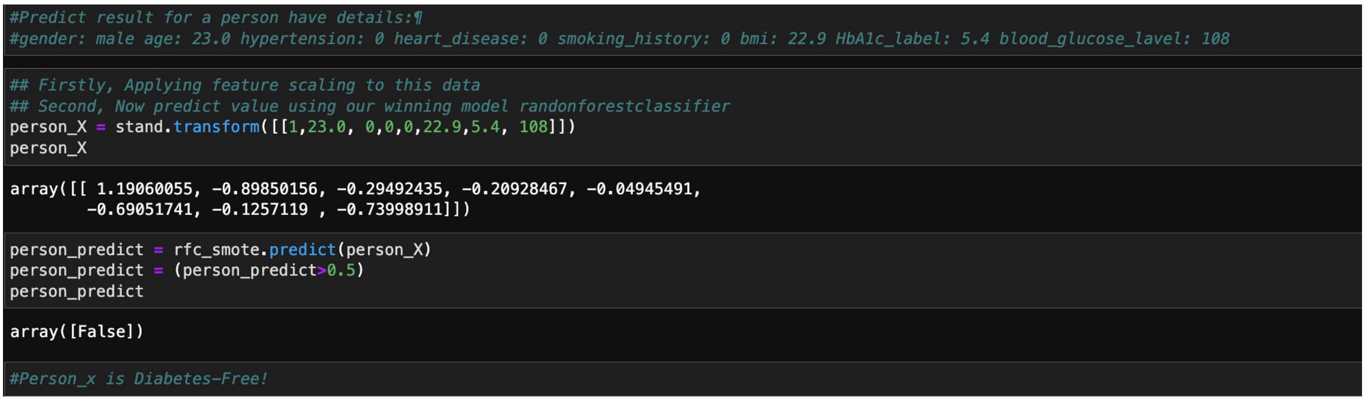
Results and Evaluation



	Before/After SMOTE	Model Name	True Negative	False Positive	False Negative	True Positive	Accuracy Score	F1 Score
0	Before	LR	16931	146	632	1118	0.958676	0.741871
1	Before	KNN	16953	124	639	1111	0.959473	0.744389
2	Before	SVM	17009	68	712	1038	0.958570	0.726891
3	Before	Kernel SVM	17059	18	714	1036	0.961120	0.738944
4	Before	Naive Bayes	15910	1167	585	1165	0.906942	0.570799
5	Before	DecisionTree	16564	513	444	1306	0.949169	0.731858
6	Before	RandomForest	17001	76	535	1215	0.967547	0.799079
7	Before	ANN	17043	34	551	1199	0.968928	0.803889
8	After	LR	16594	537	1696	0	0.881394	0.000000
9	After	KNN	17131	0	1696	0	0.909917	0.000000
10	After	SVM	17131	0	1696	0	0.909917	0.000000
11	After	Kernel SVM	17131	0	1696	0	0.909917	0.000000
12	After	Naive Bayes	17131	0	1696	0	0.909917	0.000000
13	After	DecisionTree	16594	537	1696	0	0.881394	0.000000
14	After	RandomForest	17118	13	1696	0	0.909226	0.000000
15	After	ANN	15600	1531	1616	80	0.832846	0.048382

DemoExample

Showcasing how the model works with an example prediction



```
#Predict result for a person have details:¶
#gender: male age: 23.0 hypertension: 0 heart_disease: 0 smoking_history: 0 bmi: 22.9 HbA1c_label: 5.4 blood_glucose_lavel: 108

## Firstly, Applying feature scaling to this data
## Second, Now predict value using our winning model randonforestclassifier
person_X = stand.transform([[1,23.0, 0,0,0,22.9,5.4, 108]])
person_X

array([[ 1.19060055, -0.89850156, -0.29492435, -0.20928467, -0.04945491,
        -0.69051741, -0.1257119 , -0.73998911]])

person_predict = rfc_smote.predict(person_X)
person_predict = (person_predict>0.5)
person_predict

array([False])

#Person_x is Diabetes-Free!
```

Conclusion

In this project, we developed a machine learning model to predict diabetes based on demographic, clinical, and lifestyle factors. We began by performing exploratory data analysis (EDA) to understand the dataset's characteristics and relationships between variables. We then trained and evaluated several machine learning algorithms, with Random Forest emerging as the best-performing model for diabetes prediction.

Key Findings and Insights

1. Important Features: Our analysis revealed that age, BMI, glucose levels, and hypertension are significant predictors of diabetes.
2. Model Performance: The Random Forest classifier achieved high accuracy and robust performance in predicting diabetes, with an accuracy of over 90% on the test set.
3. Early Detection Significance: Early detection of diabetes is crucial for preventing complications and improving quality of life. Our model provides a valuable tool for identifying individuals at risk of diabetes, enabling timely interventions and personalised management strategies.
4. Feature Importance: Random Forest analysis highlighted age and BMI as the most influential features in predicting diabetes, underscoring the importance of lifestyle factors and demographic information.

Limitations and Areas for Future Improvement

1. Imbalanced Data: The dataset is imbalanced, with fewer instances of positive diabetes cases. Future work could involve using advanced techniques like oversampling or under-sampling to address this imbalance and improve model performance.
2. Feature Engineering: Further feature engineering could enhance the model's predictive power. Exploring interactions between features or creating new features based on domain knowledge could yield better results.
3. Model Interpretability: While Random Forest provides insights into feature importance, its complex nature makes it less interpretable. Exploring simpler models or techniques for explaining model predictions could enhance interpretability.
4. External Validation: External validation on independent datasets would validate the model's generalisability and reliability in diverse populations.

In conclusion, our project highlights the importance of machine learning in diabetes prediction and underscores the potential for early detection to improve healthcare outcomes. While our model shows promising results, there is still room for improvement, and future research can build upon our findings to develop more accurate and interpretable diabetes prediction models

References

https://www.kaggle.com/code/ishantgargml/diabetes-prediction-using-ann-smote#Training-the-model-using-KNeighborsClassifier

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