Data Intensive Computing
Project Phase #1

Unraveling the Determinants of Diabetes: An In-Depth Analysis of Health and Lifestyle Data

Problem Statement:

This study aims to explore and identify the key health and lifestyle factors contributing to the prevalence of diabetes. Through rigorous analysis of demographic, health condition, and lifestyle data, this project seeks to answer pivotal questions:
What are the primary factors influencing diabetes risk?
How do lifestyle choices and pre-existing health conditions correlate with diabetes prevalence?
Understanding these relationships is crucial for developing targeted interventions and preventive measures against diabetes.

Background and Objectives:

Diabetes mellitus has emerged as a global health crisis, with its prevalence increasing at an alarming rate. This chronic disease not only imposes a significant burden on individuals but also strains healthcare systems worldwide.

The objective of this study is to dissect the interplay between these factors and diabetes prevalence, aiming to quantify the impact of demographic variables on diabetes risk, examine the influence of lifestyle choices on diabetes incidence, investigate how pre-existing health conditions affect diabetes development, and determine the role of obesity and blood sugar control in predicting diabetes.

Contribution to the Problem Domain:

This project holds the potential to significantly contribute to the diabetes research domain by providing insights into modifiable risk factors, thus informing public health initiatives and individual lifestyle adjustments. Enhancing understanding of the multifaceted nature of diabetes, paving the way for more effective management and prevention strategies. Supporting healthcare providers in identifying at-risk individuals through data-driven risk assessment tools

Data Sources:

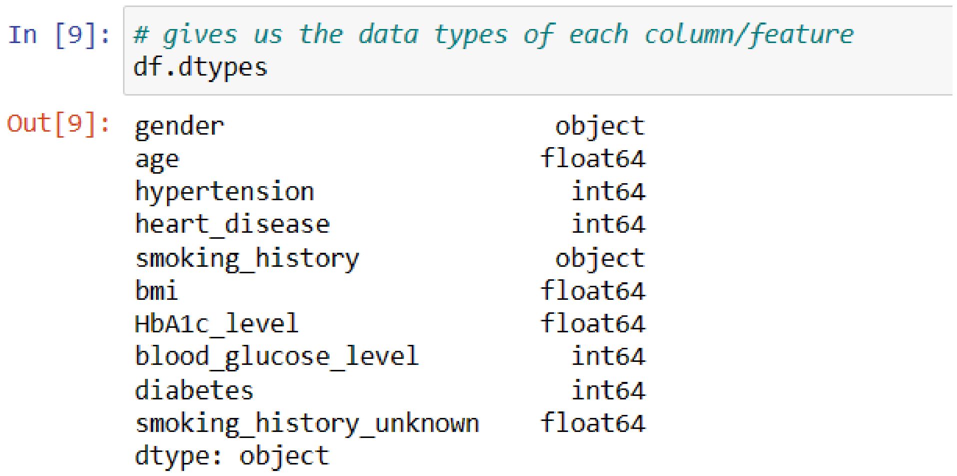
The dataset analyzed in this study comprises patient health records, including demographic information, health conditions, and lifestyle choices. It was sourced from a comprehensive health database [Kaggle], ensuring a broad representation of the population. The dataset contains around 100,019 records and 10 columns, adhering to the requirement for a large enough dataset for significant data analysis.
Diabetes prediction dataset (kaggle.com)
We also augmented the existing dataset for achieving more cleaning/Analysis of data.

Dataset Overview

The dataset includes the following columns, each offering unique insights into the patient's health and lifestyle:

“gender” : Contains the gender of a person which is categorical data.
“age” : Contains the age data of different people.
“hypertension” : tells us whether a person has hypertension or not using ones and zeros.
“heart_disease” : tells us whether a person has heart related disease or not using ones and zeros.
“smoking_history” : Contains the smoking history of a person which is categorical data.
“bmi” : patient's Body Mass Index, with some missing values containing continuous data.
“HbA1c_level”: Continuous variable representing the patient's Hemoglobin A1c level, a key indicator of diabetes management, with some missing values.
“blood_glucose_level” : Continuous variable indicating the patient's blood glucose level.
“diabetes” : has ones or zeros representing whether patient has diabetes or not.
“Smoking_history_unknown” : Binary variable specifically encoding unknown smoking history as a distinct category.

Using the above features together can be used to predict the risk of diabetes.



```
In [9]: # gives us the data types of each column/feature
        df.dtypes
```

```
Out[9]: gender                       object
        age                         float64
        hypertension                  int64
        heart_disease                 int64
        smoking_history              object
        bmi                         float64
        HbA1c_level                 float64
        blood_glucose_level           int64
        diabetes                      int64
        smoking_history_unknown     float64
        dtype: object
```

Data Cleaning/Processing:

The dataset underwent rigorous cleaning and processing to ensure the integrity and reliability of the analysis:

1. Drop Duplicate Rows: Identified and removed duplicate records to prevent skewed analysis.

2. Type Conversion: Converting gender,smoking_history column from object type to String type and float to int data type respectively.

3. Changing to Lowercase : converting all the data in the gender column into lower case.

4. Renaming Column: renaming the column HbA1c_level to ghlevel.

5. Dropping null value containing rows : bmi column has rows containing nan values do we drop them.

6. Dropping unnecessary columns : Dropped smoking_history_unknown column

7. Replacing null values with Median/Fillna() : we replaced the nan/null values in ghlevel column with median and smoking_history with specific string value.

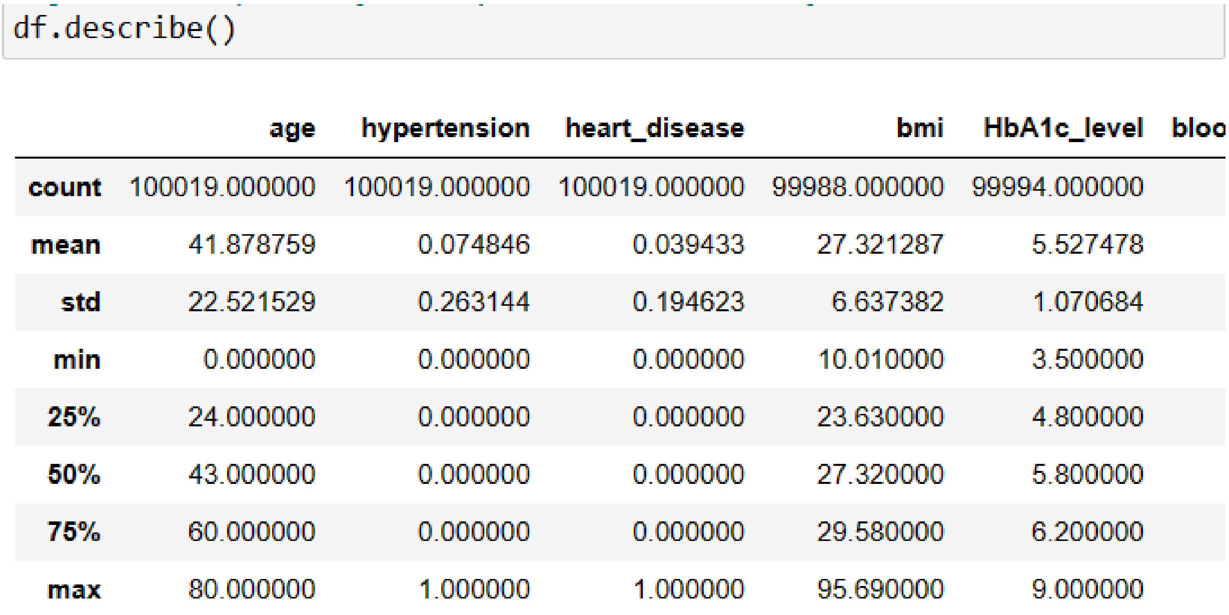
8. Standardization : we standardize [‘age', 'bmi', 'ghlevel', 'blood_glucose_level'] so that their ranges lie on the same plane.

9. Handle Outliers : These data anomalies can skew results, leading to false decisions

10. Encoding categorical data into numerical data. As some data is in categorical form, the model needs to label it to different values so we convert categorical data into numerics.

Exploratory Data Analysis (EDA):

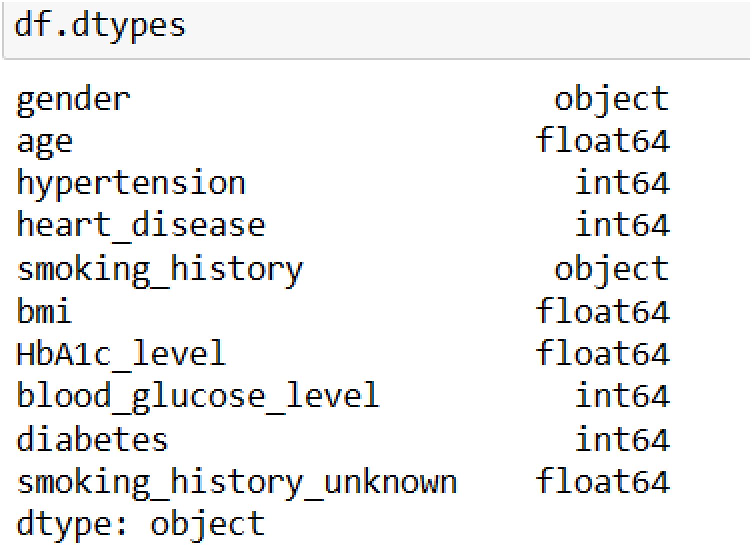
1. The EDA was conducted following NIST guidelines and John Tukey's principles, focusing on uncovering patterns, distributions, and relationships within the dataset:



```
df.describe()
```

	age	hypertension	heart_disease	bmi	HbA1c_level	bloo
count	100019.000000	100019.000000	100019.000000	99988.000000	99994.000000	
mean	41.878759	0.074846	0.039433	27.321287	5.527478	
std	22.521529	0.263144	0.194623	6.637382	1.070684	
min	0.000000	0.000000	0.000000	10.010000	3.500000	
25%	24.000000	0.000000	0.000000	23.630000	4.800000	
50%	43.000000	0.000000	0.000000	27.320000	5.800000	
75%	60.000000	0.000000	0.000000	29.580000	6.200000	
max	80.000000	1.000000	1.000000	95.690000	9.000000	

2. We use the describe method of pandas to print the various statistical measures of different columns/features.

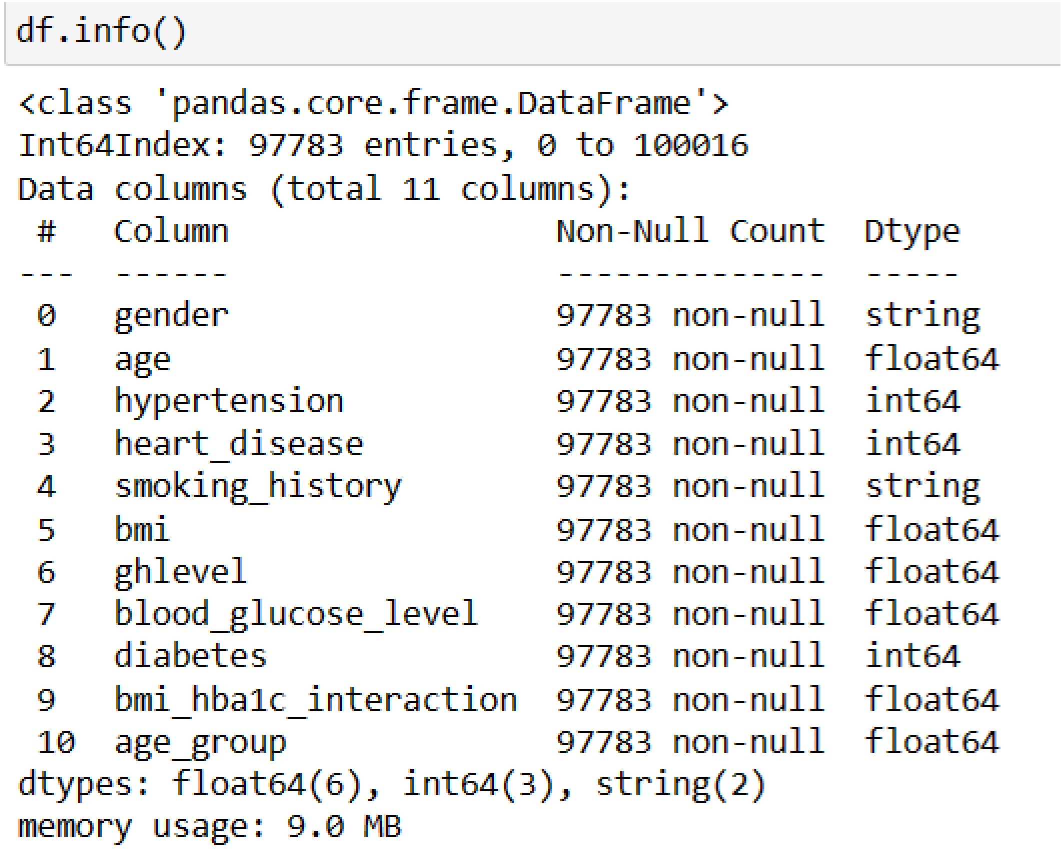


```
df.dtypes
```

```
gender                      object
age                        float64
hypertension                 int64
heart_disease                int64
smoking_history             object
bmi                        float64
HbA1c_level                float64
blood_glucose_level          int64
diabetes                     int64
smoking_history_unknown    float64
dtype: object
```

Initially when we find the data types we find that gender/smoking_history are in Object format. So we convert that into string data type and similarly convert the age from float to int datatype.

3. After standardization we get the following information from the final cleaned data set.



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 97783 entries, 0 to 100016
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   gender                 97783 non-null  string
 1   age                    97783 non-null  float64
 2   hypertension           97783 non-null  int64
 3   heart_disease          97783 non-null  int64
 4   smoking_history        97783 non-null  string
 5   bmi                    97783 non-null  float64
 6   ghlevel                97783 non-null  float64
 7   blood_glucose_level    97783 non-null  float64
 8   diabetes               97783 non-null  int64
 9   bmi_hba1c_interaction  97783 non-null  float64
 10  age_group              97783 non-null  float64
dtypes: float64(6), int64(3), string(2)
memory usage: 9.0 MB
```

Here “bmi_hba1c_interaction” is the new feature added for our better understanding.

4. Distribution Analysis: Examined the distributions of continuous variables like age, bmi, and HbA1c_level to understand the population's health status.

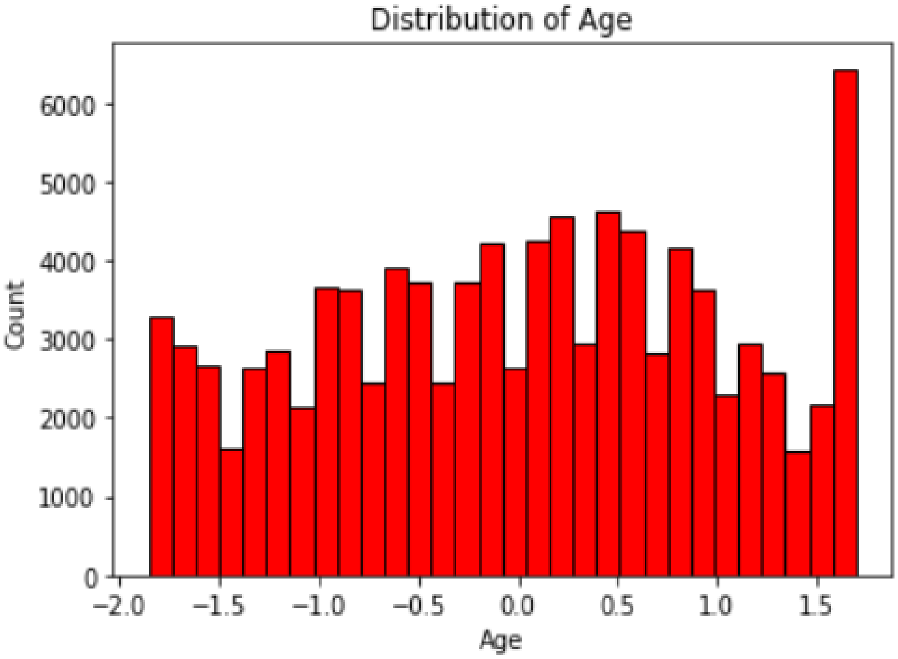


fig.1.1

5. The provided code outputs a histogram that visually represents the distribution of ages in the dataset. This histogram will have 30 bins (or bars). This visualization helps in understanding the spread and central tendencies of age within the data, such as whether most participants are young, middle-aged, or older, and if there are any significant gaps or clusters in the age distribution.

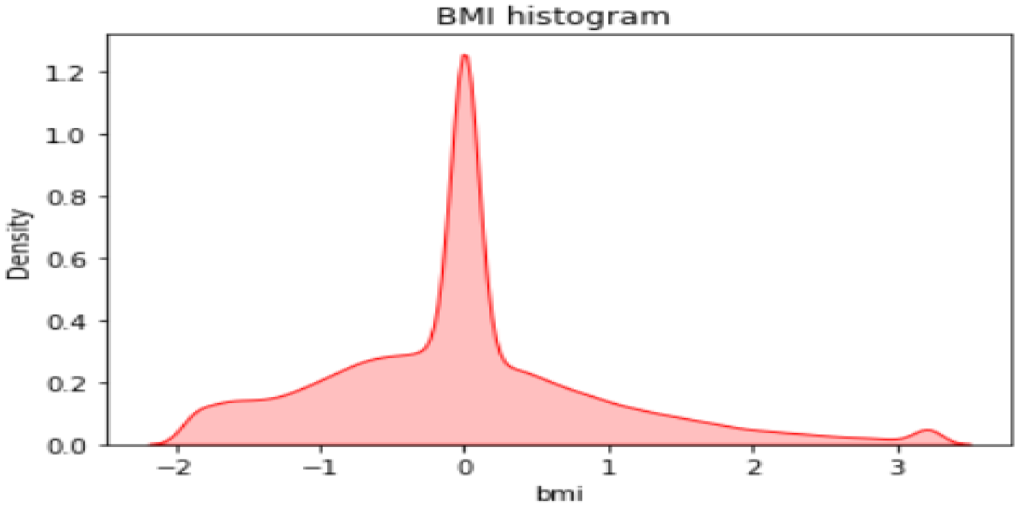


Fig.1.2. KDE BMI

The code outputs a Kernel Density Estimate (KDE) plot for the Body Mass Index (BMI) values from the dataset. A KDE plot is a way to estimate the probability density function of a continuous variable. This plot helps in understanding the distribution of BMI values across the dataset, showing where values are concentrated and how they spread out, indicating common BMI ranges and outliers if any.

6. Categorical Analysis: Analyzed categorical variables such as gender and smoking_history to assess their impact on diabetes prevalence.

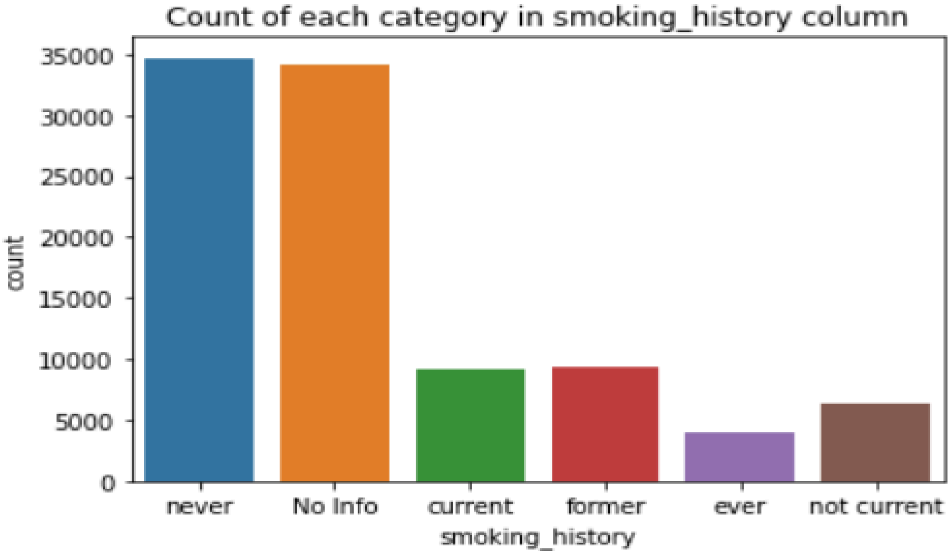
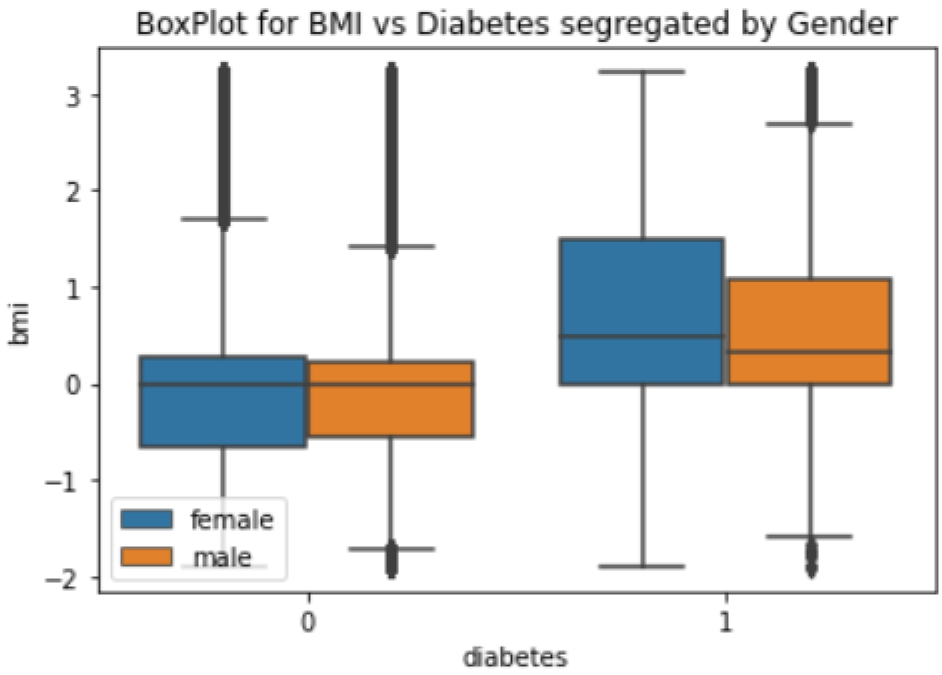


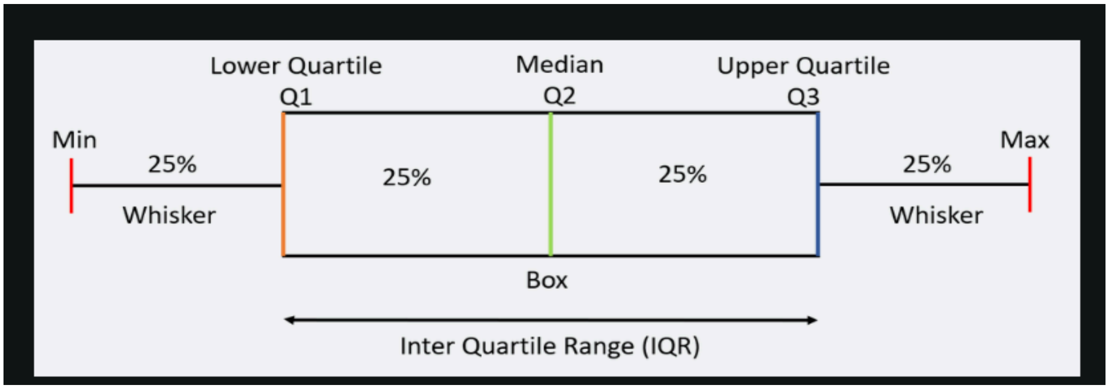
Fig.2.1

Above bar chart showing the distribution of categories within the smoking_history column of a dataset.

7. Each bar represents a different smoking status (e.g., never smoked, current smoker), with the bar height reflecting the number of observations in each category. The chart is titled "Count of each category in smoking_history column."



8. Boxplot gives us the summary of five values as shown in the below snippet between body mass index, gender and diabetes status.



Note: Took above reference from “geeksforgeeks.org”

9. Correlation Analysis: Utilized heatmaps to visualize correlations between variables, identifying potential predictors of diabetes.

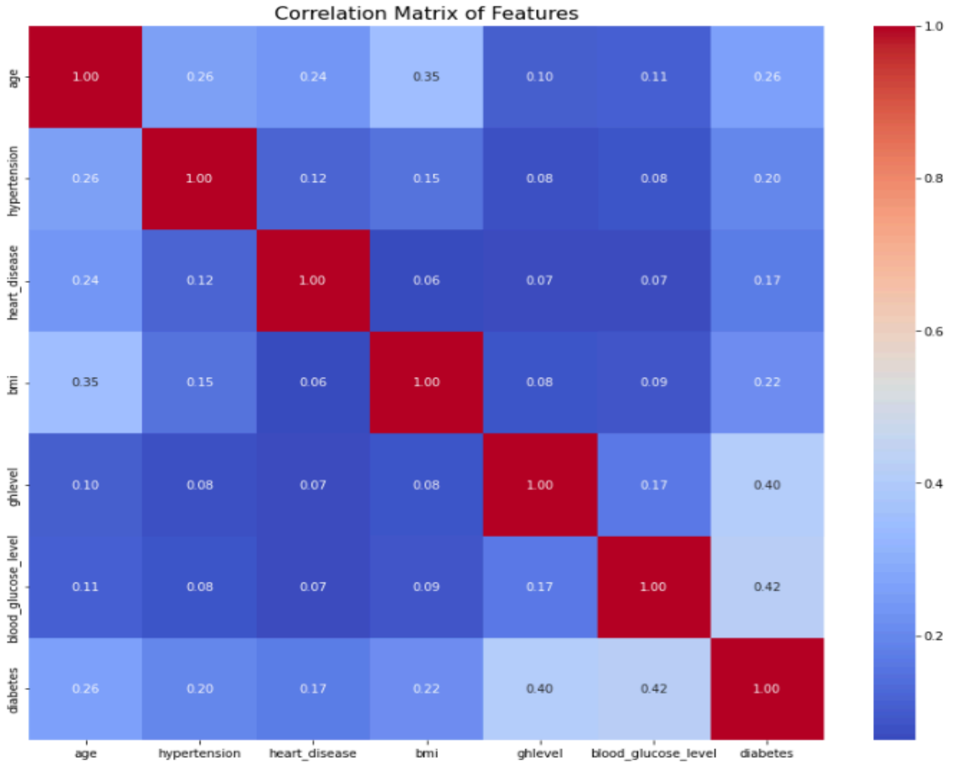
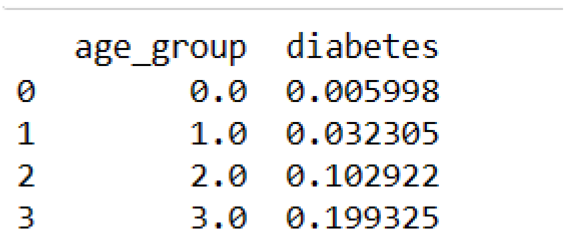


Fig.3.1

Above fig.3.1 heatmap visualizes the correlation matrix of the dataset's features(age,bmi,blood_glucose_level,ghlevel,diabetes,heart_disease,hypertension), displaying pairwise correlation coefficients with values between -1 and 1.

The exploratory data analysis (EDA) of our dataset revealed several critical outcomes that enhance our understanding of diabetes and its associated risk factors. Notably, the analysis led to the identification of significant patterns and relationships within the dataset, which are instrumental in shaping our approach to feature selection, column management, and feature engineering.

10. we are calculating the prevalence of diabetes for different age_groups



```
   age_group  diabetes
0        0.0  0.005998
1        1.0  0.032305
2        2.0  0.102922
3        3.0  0.199325
```

Influential Features on Diabetes Risk:

- Age, BMI, HbA1c levels, and blood glucose levels were identified as strong predictors of diabetes, indicating their importance in the analysis. Based on the above features we finally predict whether diabetes occurs or not using some predictive models.
- Hypertension and heart disease emerged as significant health conditions associated with an increased risk of diabetes.

Impact of Lifestyle Choices:

- A clear relationship between smoking history and diabetes prevalence was observed, underscoring the impact of lifestyle factors on health.

Data Quality and Relevance:

- The presence of missing values in critical columns like smoking_history and BMI highlighted the need for careful data imputation strategies to preserve the dataset's integrity.

Learnings:

Through the process of conducting EDA, several key learnings were acquired:

- **Data munging Importance:** The necessity of thorough data cleaning and preprocessing became evident, especially in handling missing values and outliers to maintain the accuracy of the analysis.

- **Inclusion of Features:** The analysis underscored the importance of selecting features that significantly impact the target variable, diabetes, for inclusion in modeling efforts.

- Feature Engineering: The potential to enhance model performance through feature engineering, such as creating interaction terms or new variables that capture the key values of the data more effectively, was recognized.

Application of Learnings

These insights will be directly applied to the subsequent phases of the project, particularly in model development and feature engineering:

Refined Feature Selection:

- The identified key predictors will be prioritized in the model development phase to ensure that the most influential factors are considered in predicting diabetes risk.

Data Cleaning and Preprocessing:

- A systematic approach will be adopted to address missing values and outliers, based on the insights gained, to improve data quality and analysis reliability.

Feature Engineering:

- New features will be engineered, such as interaction terms between BMI and age or HbA1c levels, to capture their combined effect on diabetes risk more accurately.
- Redundant or less informative columns, identified during the EDA, will be dropped to streamline the dataset and focus on the most relevant predictors.

Informed Modeling Decisions:

- The insights gained will guide the selection of modeling techniques and the formulation of hypotheses regarding diabetes risk factors, enhancing the predictive accuracy and interpretability of the models developed.

Conclusion:

The outcomes and learnings from the EDA phase are invaluable in advancing our understanding of diabetes and its risk factors. They provide a solid foundation for informed decision-making in feature selection, data processing, and model development. By applying these insights, we aim to improve the accurate prediction of diabetes risk and contribute to more effective prevention and management strategies.

Peer Evaluation Form for Final Group Work

CSE 487/587A

- Please write the names of your group members.

Group member 1: Harshavardhan Reddy Nadedi

Group member 2: Likhith Kongara

Group member 3: Sai Sohan Kosaraju

- Rate each groupmate on a scale of 5 on the following points, with 5 being HIGHEST and 1 being LOWEST.

Evaluation Criteria	**Group member 1**	**Group member 2**	**Group member 3**
How effectively did your group mate work with you?	5	5	5
Contribution in writing the report	5	5	5
Demonstrates a cooperative and supportive attitude.	5	5	5
Contributes significantly to the success of the project .	5	5	5
TOTAL	20	20	20

Also please state the overall contribution of your teammate in percentage below, with total of all the three members accounting for 100% (33.33+33.33+33.33 ~ 100%):

Group member 1: 33.3

Group member 2: 33.3

Group member 3: 33.4