Identifying Key Health Indicators and Sociodemographic Factors Associated with Diabetes Risk

Description of the questions:
The aim of this project is to identify key factors related to diabetes risk by analyzing health indicators from both diabetic and non-diabetic individuals. The goal is to develop a machine learning model that uses these health metrics to predict the likelihood of diabetes occurrence, while also exploring health disparities across different groups. In addition, the project will examine how these factors influence diabetes diagnosis and the healthcare needs of patients, particularly in groups with poorer health conditions.

Description of the variables:

Diabetes : This binary variable indicates whether an individual has diabetes (0 means no diabetes and 1 means diabetes), which is the target we want to predict.
HbA1c_level: Hemoglobin A1C (HbA1C) level represents the average blood sugar (glucose) level over the past two to three months.
Blood_glucose_level: current level of glucose in the blood, provides a snapshot of the person's blood sugar level at that particular point.
Bmi: The Body Mass Index (BMI) is a calculation based on an individual's weight and height,
Hypertension: A condition characterized by consistently high blood pressure, which increases the risk of heart disease, stroke, and kidney problems.
Heart_disease: Include various heart conditions, including coronary artery disease, heart attacks, heart failure etc.
Age and Race: Demographic factors at the data collection time.

Description of the variable methods:
Firstly, we inspect the data and don't find any missing values. Then, we find the correlations between these variables but it's hard to identify the contributing factors and the importance of each variable. Instead, we use the ANOVA test to evaluate the significance of continuous variables like BMI, HbA1c_level, and blood_glucose_level across the two groups (diabetes = 0 and diabetes = 1), which helps us select important predictors. We will subsequently apply models such as logistic regression, SVM, and KNN to tackle the problem, and we will also visualize the overall distribution as well as the model visualization.

Q1: Can we predict the diabetes status based on a person's medical record, such as hypertension, heart disease and age?
Q2: What's the best model and hyperparameters to predict diabetes from hbA1c level, blood glucose level and bmi separately?

Data:

Diabetes Health & Demographics Dataset (100K Individuals) by ZIYA
Source: Diabetes Clinical Dataset
Key Variables: Numerical: age, BMI, hbA1c_level, blood_glucose_lovel; Categorical: gender, race, hypertension, heart disease,

Code:

```
df = pd.read_csv('diabetes_dataset_with_notes.csv', index_col=None)
X = df[['bmi', 'hbA1c_level', 'blood_glucose_level','age','hypertension','heart_disease']]
y = df['diabetes']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42,stratify=y)
```