README for Diabetes Analysis

Project Overview

This project aims to analyze a healthcare diabetes dataset, sourced from Kaggle, to uncover patterns and correlations that may help in understanding diabetes. Various exploratory data analysis (EDA) techniques are used to visualize and better understand the dataset, including descriptive statistics, distribution plots, and correlation matrices.

Dataset Information

- **Source**: Kaggle, uploaded by Abhinandan Sharma19
- **Description**: The dataset contains several medical predictor variables and one target variable. The predictor variables include features like the number of pregnancies, insulin levels, age, glucose levels, etc., while the target variable indicates whether the patient has diabetes (Outcome: 1) or not (Outcome: 0).

Dataset Columns:

1. **Pregnancies**: Number of times the patient has been pregnant.
2. **Glucose**: Plasma glucose concentration (mg/dL).
3. **Blood Pressure**: Diastolic blood pressure (mm Hg).
4. **Skin Thickness**: Skinfold thickness (mm).
5. **Insulin**: 2-Hour serum insulin (mu U/ml).
6. **BMI**: Body Mass Index (weight in kg/ (height in m) ^2).
7. **Diabetes Pedigree Function**: A function that represents a patient's diabetes history (a higher value indicates a stronger genetic predisposition).
8. **Age**: Patient's age (years).
9. **Outcome**: Class variable (0: Non-diabetic, 1: Diabetic).

Dependencies

The following libraries are required to run the code:

- **pandas:** For loading and manipulating the dataset.
- **matplotlib:** For creating static visualizations.
- **seaborn:** For creating advanced and attractive statistical plots.

You can install the required dependencies using the following command:

```
pip install pandas matplotlib seaborn
```

Code Explanation

1. Loading the Dataset

```
data = pd.read_csv(r'C:\Users\tutha\OneDrive\Desktop\diabetes.csv')
```

This line loads the dataset from the specified path into a Pandas DataFrame.

2. Displaying the First Few Rows

```
data.head()
```

This function displays the first few rows of the dataset to get an overview of the data structure and values.

3. Checking for Missing Values

```
data.isnull().sum()
```

This line checks for missing values in the dataset. In this dataset, there are no missing values, so the result shows zero for all columns.

4. Descriptive Statistics

```
data.describe()
```

This function provides summary statistics such as count, mean, standard deviation, and percentiles for each numeric column in the dataset.

5. Visualizing Age Distribution

```
sns.histplot(data['Age'], bins=20, kde=True)

plt.title('Age Distribution of Patients')

plt.show()
```

This code generates a histogram to visualize the age distribution of patients in the dataset. A kernel density estimate (KDE) is also plotted to show the data's probability density.

6. Glucose Levels by Outcome

```
sns.boxplot(x='Outcome', y='Glucose', data=data)

plt.title('Glucose Levels vs Diabetes Outcome')

plt.show()
```

A boxplot is created to visualize the glucose levels for patients who have diabetes (Outcome=1) and those who do not (Outcome=0). This plot helps in identifying how glucose levels are distributed among diabetic and non-diabetic patients.

7. Correlation Matrix

```
corr = data.corr()
```

```
sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.title('Correlation Between Variables')

plt.show()
```

The correlation matrix shows the relationships between all pairs of features. A heatmap is generated with the correlation values annotated on the cells, providing a clear view of how the different variables are related to each other.

Conclusion

This project is a basic analysis of the diabetes dataset using common data exploration techniques. The visualizations help to understand the distribution of data points and the relationships between features, potentially aiding in understanding which variables are more significant in predicting diabetes outcomes.

Future Work

- **Feature Engineering**: Transforming or creating new features for better predictive power.
- **Machine Learning Models**: Implementing models like Logistic Regression, Decision Trees, or Random Forest to predict the likelihood of diabetes.