Cardiometabolic Insight Project

Project Proposal: Exploratory Data Analysis on Heart Attacks and Diabetes Risk Factors

1. Introduction:

This project aims to conduct an exploratory data analysis on risk factors for heart attacks and diabetes. The dataset will be created by merging two relevant datasets on heart attacks and diabetes, allowing us to examine the relationship between various risk factors and the occurrence of these conditions. The analysis will involve statistical analysis, visualization, and potentially building predictive models. The project will be implemented using Python and its libraries, with the results documented in a Jupyter Notebook.

2. Dataset Selection and Merging:

Two datasets will be selected, one specifically focused on heart attacks and another on diabetes. The datasets should contain variables related to risk factors such as age, sex, cholesterol levels, blood pressure, diabetes, family history, smoking, obesity, alcohol consumption, exercise hours per week, previous heart problems, hypertension, heart disease, BMI (body mass index), HbA1c level, and blood glucose level. After selecting the datasets, we will merge them into a single dataset to facilitate analysis and exploration.

3. Project Objectives:

The primary objectives of this project are as follows:

- Perform data cleaning and preprocessing on the merged dataset, addressing missing values, inconsistencies, and merging data from the separate datasets.
- Conduct a statistical analysis to identify the most contributing risk factors for heart attacks and diabetes. Explore the correlation or association between the risk factors and the occurrence of each condition.
- Compare the distribution and characteristics of risk factors between individuals with heart attacks and those with diabetes, identifying any overlapping or unique factors associated with each condition.
- Analyze the relationship between obesity and the occurrence of heart attacks or diabetes, considering other interacting variables such as age, sex, and lifestyle factors.
- Calculate the prevalence rates of heart attacks and diabetes in the dataset and assess their distribution across different demographic and risk factor categories.
- Explore the possibility of building predictive models to identify individuals at high risk of heart attacks or diabetes. Evaluate different algorithms and assess the importance of each variable in the prediction process.

- Investigate potential interactions or synergistic effects between risk factors for heart attacks and diabetes, determining if certain combinations of risk factors pose a higher risk for developing both conditions concurrently.

4. Methodology:

The project will follow a structured approach:
a) Data Acquisition and Merging: Select two suitable datasets related to heart attacks and diabetes, respectively. Merge the datasets into a single dataset for further analysis.
b) Data Cleaning and Preprocessing: Address missing values, inconsistencies, and merge data from the separate datasets. Document the cleaning process and any modifications made.
c) Statistical Analysis and Visualization: Perform statistical analysis to identify the most contributing risk factors for heart attacks and diabetes. Compare risk factor distributions between the two conditions. Analyze the relationship between obesity and the occurrence of heart attacks or diabetes. Visualize the findings using appropriate plots and parameter choices.
d) Prevalence Calculation and Assessment: Calculate the prevalence rates of heart attacks and diabetes in the dataset, considering different demographic and risk factor categories. Assess the distribution and patterns observed.
e) Predictive Modeling (Optional): Explore the possibility of building predictive models to identify high-risk individuals for heart attacks or diabetes. Evaluate the performance of different algorithms and assess variable importance. [out of Scope]
f) Documentation and Submission: Create a well-structured Jupyter Notebook documenting the entire workflow, including data cleaning, analysis, and visualization. Submit the merged dataset, the Jupyter Notebook, and any additional required materials.

5. Evaluation and Assessment:

The project will be evaluated based on the following criteria:

- Code functionality: Ensure the code is functional, error-free, and follows good coding practices.
- Quality of analysis: Clearly state research questions and address them comprehensively in the analysis.
- Data cleaning and merging: Document the cleaning process and merging of datasets, ensuring transparency and reproducibility.
- Statistical analysis and visualization: Perform appropriate statistical tests and visualize the results effectively.
- Prevalence calculation and assessment: Calculate prevalence rates and analyze their distribution across different categories.
- Predictive modeling (if applicable): Evaluate the performance of predictive models and assess variable importance.
- Documentation and submission: Submit a well-documented Jupyter Notebook, the merged dataset, and any additional required materials.

6. Results

Heart Attack Risk Factors

The most significant risk factors for heart attack, in decreasing order of importance, were found to be Sleep Hours Per Day, Diabetes, Obesity, and Cholesterol. These factors showed the highest correlation with the occurrence of heart attacks.

Diabetes Risk Factors

The risk factors most associated with diabetes were Heart Attack Risk, Cholesterol, Previous Heart Problems, and Sleep Hours Per Day. These factors showed the highest correlation with the occurrence of diabetes.

Differences in Risk Factors Between Heart Attack and Diabetes

The analysis revealed significant differences in the distributions of Obesity and Cholesterol between individuals with heart attacks and those with diabetes. The mean level of Obesity was lower in the heart attack group than in the diabetes group, while the mean level of Cholesterol was higher in the heart attack group.

Relationship Between Obesity and Heart Attacks or Diabetes

The correlation analysis showed a significant positive relationship between Obesity and the occurrence of heart attacks and diabetes. This suggests that individuals with higher levels of obesity are more likely to experience heart attacks or diabetes.

Prevalence of Heart Attacks and Diabetes

The overall prevalence of heart attacks in the dataset was approximately 35.4%, while the prevalence of diabetes was higher at approximately 65.6%.

Interactions Between Risk Factors

There were significant interactions observed between various risk factors for heart attacks and diabetes. Certain combinations of risk factors, such as high levels of Obesity and Cholesterol, were associated with a higher risk of developing both conditions concurrently.

Geographical Distribution of Heart Attacks and Diabetes

The geographical analysis showed varying prevalence rates of heart attacks and diabetes across different countries.

Conclusion:

This project aims to explore the risk factors for heart attacks and diabetes through an extensive analysis of merged datasets. By conducting statistical analysis, visualization, and potentially building predictive models, we will gain insights into the most contributing factors for each condition. The

project will provide valuable experience in data cleaning, exploratory data analysis, and visualization using Python, enabling effective communication of findings related to heart attacks and diabetes risk factors.

In conclusion, this analysis provides valuable insights into the risk factors associated with heart attacks and diabetes, and could inform targeted prevention and intervention strategies. However, further research is needed to understand the causal pathways underlying these associations and to develop effective predictive models for these conditions.

Appendix A

Brain Storming Questions

I. What are the most contributing risk factors for heart attack? Perform a statistical analysis to identify the factors that have the highest correlation or association with heart attacks. Consider variables such as age, sex, cholesterol levels, blood pressure, diabetes, family history, smoking, obesity, alcohol consumption, exercise hours per week, and previous heart problems.

II. What are the most contributing risk factors for diabetes? Conduct a statistical analysis to determine the factors that have the strongest relationship with diabetes. Explore variables such as gender, age, hypertension, heart disease, smoking history, BMI (body mass index), HbA1c level, and blood glucose level.

III. Are there any significant differences in risk factors between individuals with heart attacks and those with diabetes? Compare the distribution and characteristics of risk factors in these two groups using appropriate statistical tests. Identify any overlapping risk factors or unique factors associated with each condition.

IV. Is there a relationship between obesity and the occurrence of heart attacks or diabetes? Analyze the correlation between obesity and the presence of heart attacks or diabetes. Consider other variables that may interact with obesity, such as age, sex, and lifestyle factors.

V. What is the overall prevalence of heart attacks and diabetes in the dataset? Calculate the prevalence rates for each condition and assess their distribution across different demographic and risk factor categories.

VI. Can a predictive model be built to identify individuals at high risk of heart attacks or diabetes? Explore the possibility of developing machine learning models to predict the likelihood of heart attacks or diabetes based on the available risk factors. Evaluate the performance of different algorithms and assess the importance of each variable in the prediction.[Out of Scope]

VII. Are there any interactions or synergistic effects between risk factors for heart attacks and diabetes? Investigate potential interactions between variables to determine if certain combinations of risk factors pose a higher risk for developing both conditions concurrently.