CS 699: Revised Project Proposal

(23M0808) & (23M0811)

September 30, 2023

Project Title

Strokes Uncovered: Data Analysis, Visualization, and Predictive Insights

Note to the Professor: **In response to your feedback and guidance, We have made revisions to the project proposal. The most significant change includes the integration of HTML using the Python's Flask web framework to develop a web application. The *Methods and Tools* section outline this update and provide a clearer overview of the HTML integration.**

Introduction

The project, titled "Strokes Uncovered: Data Analysis, Visualization, and Predictive Insights," is dedicated to conducting exploratory data analysis (EDA), which encompasses various techniques such as histograms, scatter plots, bar charts, and heatmaps. Additionally, it involves data visualization and predictive modeling using a publicly available dataset. Strokes represent a significant global health concern, accounting for approximately 11% of worldwide deaths, as reported by the World Health Organization (WHO). The primary objective of this project is to gain a deeper understanding of the risk factors that influence stroke and to facilitate more effective preventive measures and early interventions.

About Dataset:

The dataset consists of more than 5000 data points and has 10 input features such as (age, hypertension, heart_disease, martial_status, work_type, residence_type, average_glucose_level, bmi, smoking_status, gender).

Objectives

1. **Data Analysis:** The project will start with a comprehensive data analysis to uncover insights into attribute distributions and relationships. It will address specific questions and hypotheses using statistical methods such as descriptive statistics and hypothesis testing:
    - What is the gender distribution in the dataset, and does it impact stroke likelihood?
    - Is there a correlation between patient age and stroke risk?
    - Does residence type significantly influence stroke risk?
    - Are married individuals more or less likely to experience strokes compared to unmarried individuals in the dataset?
2. **Data Visualization:** The project will create informative visualizations using Python's Matplotlib and Plotly libraries to effectively convey dataset characteristics, relationships, and trends:
    - Age, glucose levels, and BMI will be depicted through histograms and density plots.
    - Gender distribution and marital status will be visualized using bar charts.
    - Scatter plots will explore attribute relationships with stroke risk.
    - An interactive Plotly visualization will offer dynamic dataset exploration.

3. **Stroke Prediction Model:** In this phase of the project, we will build a predictive model employing machine learning algorithms. The model's purpose is to discern individuals at higher risk of stroke by analyzing the attributes within our dataset.

Methods and Tools

To fulfill project requirements, the following tools and technologies will be employed:

- **Python:** Python and its various libraries will be used for data analysis, visualization, and model construction.
- **HTML Integration with Python (Flask):** We will integrate HTML with Python using the Flask web framework. In this integration, HTML will serve as the front-end interface, while Flask will function as the back-end framework, enabling the creation of an interactive web application. Users can engage with our project through this interface.
- **LaTeX Integration:** LaTeX will play a crucial role in creating a comprehensive and structured report to document our findings and project details.
- **Pyplot:** Pyplot will be used for creating a wide array of data visualizations, including bar charts, line plots, scatter plots, and histograms, to effectively describe our insights and analysis results.
- **PostgreSQL (Optional):** PostgreSQL will be employed for data storage and retrieval, particularly if the dataset size or database management complexity requires it.

Project Documentation

Thorough documentation, including code comments, explanations, dataset sources, data pre-processing details, and model evaluation results, will be carefully drafted.

Conclusion and Impact

The project's primary objective is to offer valuable key insights into stroke risk factors by finding underlying trends and patterns in the data. By performing data analysis, visualization, and predictive modeling, we can help healthcare professionals and policymakers use these insights to develop targeted prevention strategies, promote healthier lifestyles, and allocate resources more effectively to reduce the burden of strokes on society.

Note

This proposal outlines the initial project scope. Further refinements, modifications, details, and data information may be incorporated as the project progresses.