Data Mining Final Project Topic Proposal

Team 6 Collaborators: Abhradeep Das, Disha Kacha, Neelima Puthanveetil, Devarsh Sheth

Introduction:

Stroke stands as the second leading cause of death globally, responsible for approximately 11% of total mortalities, according to the World Health Organization (WHO). The debilitating impact of strokes on individuals, families, and healthcare systems necessitates proactive measures to predict and prevent its occurrence.

This project delves into the prediction of stroke occurrences based on various demographic and health-related factors. The dataset at hand offers a wealth of information, including patient demographics (age, gender), health conditions (hypertension, heart disease), lifestyle factors (BMI, smoking status), and other relevant parameters.

SMART Questions:

1. **What is the impact of lifestyle choices, particularly BMI and smoking status, on the probability of having a stroke? Are there significant differences among different groups?**
    - Evaluate the relationship between lifestyle factors (BMI, smoking status) and stroke occurrences, considering variations across different groups like gender, age, or residential areas.

2. **How do the levels of hypertension and heart disease individually, or in combination, impact the probability of stroke among different work types?**
    - Investigate the relationship between hypertension, heart disease, and their interaction among various work types to discern if certain professions are more susceptible to strokes due to these health factors.

3. **Is it possible to assess the influence of residence type, occupation, and smoking habits on stroke frequency?**
    - Evaluate the potential influence of smoking patterns and occupation on stroke risk among urban and rural residents.
4. **Is there a significant correlation between marital status, gender, and age in this analysis?**
    - Analyze marital status data to identify patterns within specific age groups among the married, and assess gender-based stroke frequency differences.

Abstract:

This project's primary objective is to construct a predictive model that determines the likelihood of an individual experiencing a stroke based on diverse parameters. By employing machine learning algorithms and statistical analyses on the provided dataset, the aim is to discern patterns and correlations between demographic attributes, lifestyle factors, and existing health conditions with the occurrence of strokes.

Insights derived from this analysis can not only aid in accurate stroke predictions but also contribute to informing public health initiatives, personalized healthcare strategies, and interventions. Ultimately, the envisioned outcome is the reduction of stroke incidences through targeted preventive measures, fostering healthier lifestyles, and offering timely medical interventions to mitigate the impact of this life-threatening condition.

Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

Github Repository: https://github.com/DishaKacha7/Data_Mining_Project