**Datathon, Week #3 Final Report**

**Team**: Hongyan Chen,Zhi Lin Zhou,Kayla Esser **Team Number**: 1

**Date**: 2023-09-26

**Introduction:**

For this datathon, we choose to explore the ability to early prediction and detection of heart diseases using individual factors presenting in the “*Mortality Dataset for Cardiovascular Disease Complications”* dataset. Driven by genetic, biological, lifestyle, and environmental factors, there has been a significant higher number of mortality and morbidity in In specific populations(1). Given the significant personal and public health burden of this condition, we sought to predict the risk of a patient developing a stroke. To do this, we choose logistic regression as the model to predict the likelihood of an individual experiencing a stroke. The application of logistic regression model in stroke prediction can aid healthcare professions in effectively detecting potential and high-risk patients, also mitigating the risk of stroke occurrence by implementing prevent measures.

**Data Engineering Process:**

Using the “*Mortality Dataset for Cardiovascular Disease Complications”* dataset, we sought to explore the ability of a combination of demographic variables (age, gender), anthropomorphic variables (BMI,average glucose level), medical history factors (hypertension,heart disease), and lifestyle factors (marital status, work type, residence type and smoking status) to identify the risk of an individual experiencing a stroke.We began by performing descriptive analyses on data to ensure sufficient data quality for machine learning analysis.We identified any Null values within variables(BMI) and explored continuous data for potential outliers(BMI, average glucose level). Moreover, we explored the distribution of each input feature, using a boxplot for continuous data and a histogram for discrete data. Also, a pair plot is used to Displaying the relationships and distributions between variables. Our goal here was to ensure that there was no significant bias within the dataset that might impact our classification results.

Next, we pre-processed the data to ensure that the variables were in the proper format for logistic regression modeling. We encoded the categorical variable of gender,marital status, residence type, work type and smoking status into a binary format. We also restrict 'Children' variables from work type and drop missing NaN values. Finally, we normalized our data using scaler transformation from the scikit library in order to ensure all variables were normalized with a standard deviation (SD) of 1 and a mean of 0. This prevents our LR model from being disproportionately impacted by variables with different scales.

**Analysis:**

We used a Logistic Regression Model to predict the risk of patients experiencing a stroke on input variables including age, gender, BMI, average glucose level, hypertension,heart disease, marital status, work type, residence type and smoking status. Then, wu used the trained LR model to predict the test dataset (X\_test) and assess the model's performance with a confusion matrix and heat map. Lastly, we Perform a statistical summary analysis of the LR model and visualize the relationships between different features and stroke.

**Findings:**

Within the dataset, there were 5110 observations.BMI is only variable that has missing values and there are some outliers for average glucose level and BMI. However, the outliers were included in model development given their biologically plausible values. Also, work status is unevenly distributed especially for those who don't work. Based on these findings, this dataset is likely representative of the diversity seen among the specific population. We found that there exists a significant positive correlation between age, hypertension, and average blood glucose level with stroke. Lastly, accuracy of our model is 17%.

**Conclusion:**

Factors such as age, hypertension, and average blood glucose level may have a significant impact on the occurrence of stroke. Our work shows that risk of stroke can be predicted with some accuracy based on our input variables, though further work is required to fine-tune our model to obtain an accuracy that would be clinically relevant.

**Code Link:**

**Presentation Link:**

**Individual Contributions:**

HC, ZLZ, and KE all contributed to project design. All members contributed to coding, writing the report and presentation.

**References:**

1. P, Manuel D G, Leung M, et al. Risk factors for cardiovascular disease in Canada[J]. Canadian Journal of Cardiology, 2003, 19(11): 1249-1260.