CLASS ASSIGNMENT I

Instructions

The objective of this assignment is to gain hands-on experience in fitting a logistic regression model to real-world healthcare data and perform feature selection based on the model's results. In this assignment, you will work with a dataset containing information about individuals and their risk factors for stroke. **You will apply logistic regression to predict the likelihood of stroke and then identify the most important features through feature selection techniques**.

Dataset:

You will use the "stroke prediction" dataset present on the class shared google drive folder, which contains the following attributes:

1. **‘id’**: Unique identifier for each individual
2. **‘gender’**: Gender of the individual (Male/Female/Other)
3. **‘age’**: Age of the individual (numeric)
4. **‘hypertension’**: Whether the individual has hypertension (0 for No, 1 for Yes)
5. **‘heart_disease’**: Whether the individual has heart disease (0 for No, 1 for Yes)
6. **‘ever_married’**: Whether the individual is ever married (Yes/No)
7. **‘work_type’**: Type of work (Private, Self-employed, Govt_job, children, Never_worked)
8. **‘Residence_type’**: Residence type (Urban/Rural)
9. **‘avg_glucose_level’**: Average glucose level (numeric)
10. **‘bmi’**: Body Mass Index (numeric)
11. **‘smoking_status’**: Smoking status (formerly smoked, never smoked, smokes, unknown)
12. **‘stroke’**: Whether the individual had a stroke (2 for No, 1 for Yes) [Target Variable]

Follow through the 5-step process you went through in the Logistic regression class notes to complete the assignment.

STEP 1: THE DATA

Get the data set from the class shared google drive: “stroke prediction.csv” and load it into R (or Python).

STEP 2: EXPLORING AND PREPARING THE DATA (10 Marks)

1. Explore the structure of the data.
2. Handle missing values if any.
3. Conduct summary statistics for the data.
4. Run a plot for every predictor variable with our response variable, STROKE.
5. Draw a histogram of the response variable.
6. Conduct a correlation analysis of the variables.
7. Visualize the relationship between the features using a scatterplot matrix function, or any other appropriate function.
8. Encode the categorical variables using appropriate methods. Check that the target variable meets the specifications of the logistic model.

STEP 3: TRAINING THE MODEL (10 Marks)

1. Split the dataset into the training set (70%) and the testing set (30%).
2. Fit a logistic regression model on the data.
3. Interpret the coefficients of the model.

STEP 4: EVALUATING MODEL PERFORMANCE (10 Marks)

1. Evaluate the model's performance on the testing data using appropriate metrics (e.g., accuracy, precision, recall, F1-score, ROC-AUC).

STEP 5: IMPROVING MODEL PERFROMANCE (10 Marks)

1. Rank the features based on their significance using an appropriate method (L1 Regularization/Lasso Regression).
2. Fit a logistic regression model to this new dataset.
3. Evaluate the performance of the model using the selected features and compare it to the original model.
4. Discuss the impact of feature selection on model performance.

SUBMISSION INSTRUCTIONS

- This is an **individual assignment**.
- The assignment can be done in RMarkdown or Python/Jupyter Notebooks.
- Provide clear visualizations to support your findings.
- Ensure your R code is organized and follows best practices.
- Submit a document that includes code chunks, comments, and explanations for each code block and analysis. Any comments or information should be commented on your code appropriately. (Any explanation required of you from the questions SHOULD NOT be done on a separate document.)
- It should be **submitted as a PDF file**, either run under R Markdown (you can also knit to word, then save as PDF) or copy your code to Notebooks and submit it.
- The work should be submitted through this **FORM.**

The deadline for this assignment is EOD October 20th 2023.

Feel free to seek clarification or assistance from the lecturer where needed.