### **Assignment 3 Report**

**Introduction**

The aim of the assignment was to go through some simple regression and classification exercises. This assignment used the stroke data set, with 'stroke' as the target variable referring to whether a patient experienced a stroke: 1 for those who did and 0 for those who didn't. By just looking at the binary nature of this target variable, one would know that it is a classification task.

**Data Preprocessing**

1. Importing Data: The dataset was loaded using pandas.

2. Handling Missing Values: Missing values in the column `bmi` were replaced with its median value.

3. Encoding Categorical Variables: The categorical variables were converted into numeric format using one-hot encoding.

4. Feature Scaling: All of the features were scaled using StandardScaler so that all of these features are on a similar scale.

**Model Selection and Evaluation**

For this task, the following three classification models have been used.

1. Logistic Regression

2. Random Forest Classifier

3. Support Vector Machine (SVM)

GridSearchCV has been used to tune the model and is further evaluated using accuracy, precision, recall, and F1-score for the performance of the model. The data was split into two parts: a training set and test set in an 80-20 ratio. SMOTE has been used to handle class imbalance, and class weights have been balanced.

**Results**

Here are the results of the evaluation for each model:

| **Model** | **Metric** | **Class 0** | **Class 1** | **Accuracy** | **Macro Avg** | **Weighted Avg** |
| --- | --- | --- | --- | --- | --- | --- |
| Logistic Regression | Precision | 0.98 | 0.17 | 0.755 | 0.58 | 0.93 |
|  | Recall | 0.75 | 0.81 |  | 0.78 | 0.76 |
|  | F1-Score | 0.85 | 0.29 |  | 0.76 | 0.82 |
|  | Support | 960 | 62 | 1022 | 1022 | 1022 |
| Random Forest | Precision | 0.94 | 0.15 | 0.919 | 0.55 | 0.89 |
|  | Recall | 0.97 | 0.08 |  | 0.53 | 0.92 |
|  | F1-Score | 0.96 | 0.11 |  | 0.53 | 0.9 |
|  | Support | 960 | 62 | 1022 | 1022 | 1022 |
| SVM | Precision | 0.95 | 0.14 | 0.891 | 0.55 | 0.9 |
|  | Recall | 0.94 | 0.16 |  | 0.55 | 0.89 |
|  | F1-Score | 0.94 | 0.15 |  | 0.55 | 0.89 |
|  | Support | 960 | 62 | 1022 | 1022 | 1022 |

**Conclusion**

For all variables, the Random Forest model came out to be most accurate, but its recall on the minority class of stroke cases was not very good. The Logistic Regression was slightly better in terms of recall of the minority class, though at a lower accuracy as compared to the former. The SVM model maintained a relatively balanced performance but still did not perform well on the minority class.

**Submission**

- GitHub Repository: All code, data preprocessing steps, model tuning, and evaluation results were documented and submitted to the GitHub repository as required.

**Link:** <https://github.com/mukeshmugadi/Practical-Data-Analytics/blob/main/Assignment_3___Classification_of_Stroke_Data_Using_diffrent_models.ipynb>