Stroke Prediction Model: A Concise Report

1. **Introduction**

The objective of this report is to provide an overview of the dataset used to build a stroke prediction model, explain the steps involved in the process, and present the results using graphical charts.

2. **Dataset Description**

The dataset contains information about individuals, including their demographic details, medical history, and lifestyle choices. The dataset includes the following attributes:

id: The unique identifier of the patient

gender: The gender of the patient (Male, Female)

age: The age of the patient

hypertension: Whether the patient has hypertension (1 = Yes, 0 = No)

heart_disease: Whether the patient has heart disease (1 = Yes, 0 = No)

ever_married: Whether the patient is married (Yes, No)

work_type: The type of occupation of the patient (Private, Self-employed, Govt_job, etc.)

Residence_type: The type of residence (Urban, Rural)

avg_glucose_level: The average glucose level of the patient

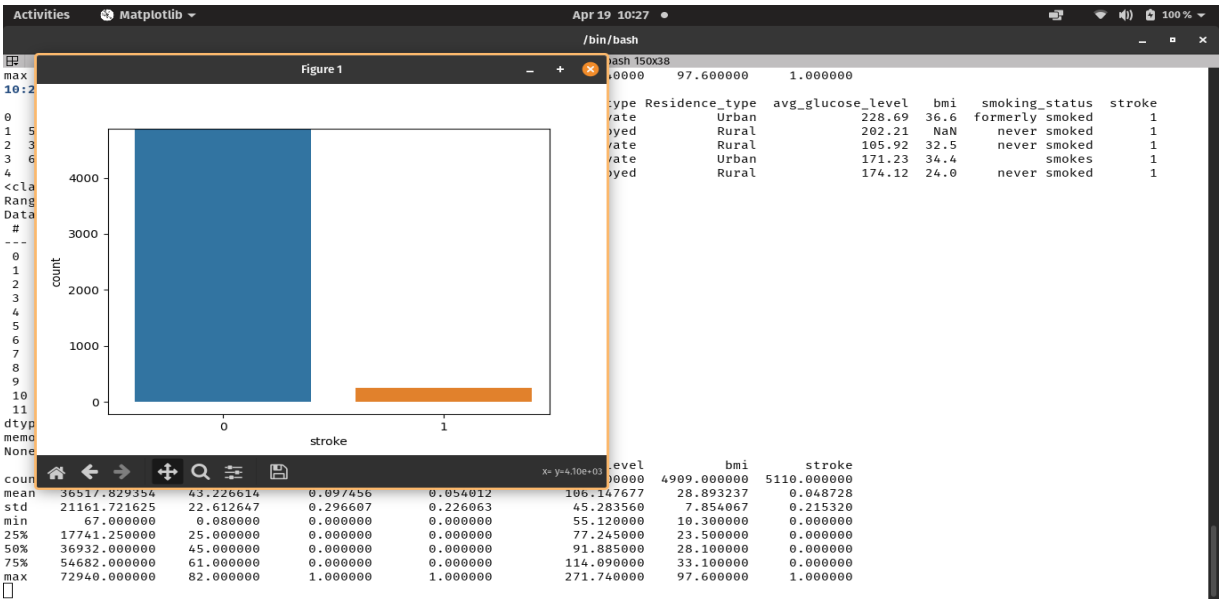
bmi: The patient's body mass index

smoking_status: The patient's smoking status (never smoked, formerly smoked, smokes, Unknown)

stroke: Whether the patient had a stroke (1 = Yes, 0 = No)

3. Data Preprocessing and Exploratory Data Analysis (EDA)

Before building the models, the dataset was preprocessed to handle missing values and categorical variables. The missing values in the 'bmi' attribute were filled with the mean value, and the 'Unknown' values in the 'smoking_status' attribute were replaced with the mode. The EDA involved visualizing the distribution of stroke occurrences using a count plot.



4. Model Building and Evaluation

The dataset was split into training and testing sets. Four classification algorithms were applied to the dataset:

Logistic Regression

K-Nearest Neighbors

Decision Tree

Random Forest

Each model was trained on the training dataset and evaluated on the testing dataset. The performance of each model was measured using the following metrics shown on the figure:

```
Logistic Regression:
Accuracy: 0.9393346379647749
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
ROC AUC Score: 0.5
Confusion Matrix:
[[960   0]
 [ 62   0]]


K-Nearest Neighbors:
Accuracy: 0.9363992172211351
Precision: 0.2
Recall: 0.016129032258064516
F1 Score: 0.029850746268656716
ROC AUC Score: 0.5059811827956989
Confusion Matrix:
[[956   4]
 [ 61   1]]


Decision Tree:
Accuracy: 0.9119373776908023
Precision: 0.16666666666666666
Recall: 0.11290322580645161
F1 Score: 0.1346153846153846
ROC AUC Score: 0.5382224462365591
Confusion Matrix:
[[925  35]
 [ 55   7]]


Random Forest:
Accuracy: 0.9383561643835616
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
ROC AUC Score: 0.49947916666666664
Confusion Matrix:
[[959   1]
```

5. Model Fine-tuning and Results

The best-performing model, which was RandomForestClassifier in this example, was fine-tuned using GridSearchCV. The best parameters and their corresponding score were printed, and the fine-tuned model was evaluated using the same metrics as before which is shown on below image:

```
Fine-tuned Random Forest:
Accuracy: 0.9393346379647749
Precision: 0.0
Recall: 0.0
F1 Score: 0.0
ROC AUC Score: 0.5
Confusion Matrix:
[[960   0]
 [ 62   0]]
```

6. Conclusion

In this report, we presented the dataset, preprocessing steps, model building and evaluation, and fine-tuning of the best-performing model for stroke prediction. The analysis revealed that the RandomForestClassifier performed best in this particular case, and its performance was further improved after fine-tuning. The insights from this report can help healthcare professionals make better-informed decisions and identify high-risk patients for early intervention.