Summary

Data

The data consists of multiple variables related to individual health and lifestyle factors that could influence the likelihood of having a stroke. These variables include age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status. The primary focus of the analysis is to predict the occurrence of a stroke, marked as the target variable in the dataset.

In the process of data preparation, multiple steps were undertaken. These included handling missing data, removing redundant information, managing categorical variables, addressing outliers, and transforming data for machine learning analysis. A key decision was made regarding the BMI column, which had missing values in both the training and testing sets. To avoid potential bias from imputing these values, rows with missing BMI data were omitted. This decision, while necessary to maintain data integrity, could impact the representativeness and size of the dataset.

Further, non-contributory columns like 'Unnamed: 0' and 'id' were identified and removed, ensuring that the analysis remained focused on variables directly relevant to stroke prediction. The 'smoking_status' variable, which included an 'Unknown' category, was another area of focus. Rather than imputing these unknown values or omitting them, they were retained as a separate category. This approach acknowledges the limitations inherent in the data while avoiding the introduction of bias that might come from imputation.

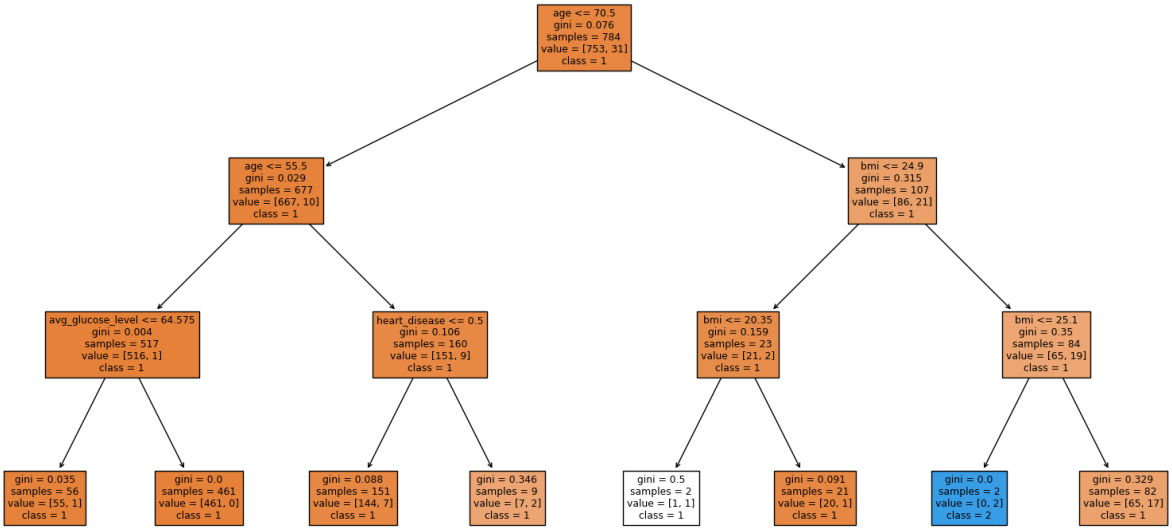
Outliers in variables such as 'bmi' and 'avg_glucose_level' were managed using the Windsorizing technique. This method involves adjusting extreme data points to reduce the impact of anomalies, ensuring a more accurate and representative dataset for analysis.

Lastly, the preparation of the dataset for machine learning model analysis involved the transformation of categorical variables into dummy variables. This conversion was applied to several variables including 'smoking_status', 'Residence_type', 'gender', 'ever_married', and 'work_type'. This transformation is vital as it allows these categorical data points to be effectively used in machine learning models, which typically require numerical input.

However, these data cleaning and preparation steps presented certain challenges. The exclusion of rows with missing BMI values, while avoiding imputation bias, might affect the dataset's overall representativeness. The use of Windsorizing to manage outliers requires careful execution to prevent loss of valuable information or distortion of data distribution. Additionally, the dataset's imbalance, with a significantly higher number of non-stroke cases compared to stroke cases (973 vs. 50), poses challenges for accurately predicting and training models.

Results:

Figure 1: Decision Tree Classifier



After concluding that a Decision Tree Regressor is not suitable for a binary outcome due to negative R^2 and RMSE scores, we utilized a Decision Tree Classifier with a maximum depth of 3. The decision tree's visualization highlighted the following rules for classification:

- Individuals ≤70.5 years old with average glucose levels ≤64.575 and without heart disease are most likely to be classified as not having a stroke (class 1).
- Those with a BMI ≤24.9 are also likely to be classified as class 1, irrespective of age.

The tree indicates that age is the most determining feature, followed by average glucose level, heart disease, and BMI.

The Decision Tree Classifier achieved an overall accuracy of 95.43%, successfully predicting non-stroke cases (class 0) in 188 out of 197 instances. However, it failed to correctly identify any true stroke events (class 1), as evidenced by the 7 false negatives and the absence of true positives in the confusion matrix. While the classifier's accuracy appears high, it is somewhat misleading due to the dataset's imbalance, with a greater number of non-stroke cases. This skew leads to a high number of true negatives, which inflates accuracy but does not reflect the classifier's ability to identify actual stroke events. The absence of true positives for stroke occurrences is a significant concern, indicating the model's limited predictive capability for the minority class.

Figure 2: kNN
Figure 3: kMC
Figure 4: linear regression

Conclusion