Group 19 - Project 2 Report

Summary

Our primary goal was to construct a predictive model that predicted the likelihood of an individual experiencing a stroke, utilizing a dataset with various health-related and lifestyle variables. The key predictor variables we chose to focus on were age, average glucose level, body mass index, marriage status, gender, heart disease status, hypertension, residence type, smoking status, and work type. The response variable was if the person suffered a stroke in the data sampling period. Before creating the models, rigorous data-cleaning procedures were implemented. Notably, missing values in the 'bmi' variable were replaced with its mean, and winsorization was applied to BMI and average glucose level to address outliers in both the training and test sets.

Following data preparation, we employed various visualization techniques to elucidate the intricate relationships between stroke and other variables, such as work type, age, gender, and smoking status. Additionally, we explored correlations between different variables and stroke occurrences. Surprisingly, most variables exhibited a weak association with the likelihood of stroke. Subsequently, we crafted three distinct models: a decision tree model, a linear regression model incorporating polynomial features, and a k-nearest Neighbors (KNN) model. The decision tree emphasized age as the most influential feature, closely followed by average glucose level. However, the classification model's performance metrics, as depicted by the confusion matrix, revealed a significant imbalance. While achieving an overall accuracy of 95.17%, the model excelled primarily in predicting the more prevalent non-stroke cases, demonstrating limitations in identifying actual stroke events. The linear regression model produced an R-squared value of 0.06635, indicating that it explains only around 6.6% of the variability in the dependent variable ('stroke'). The RMSE value of 0.2078 highlighted the model's precision in capturing underlying patterns. The KNN model, with an R-squared value of 0.04 and an RMSE of 0.211, suggested that unaccounted variables likely contribute to stroke occurrences beyond those considered in our analysis. In essence, our models provide valuable insights into stroke prediction, emphasizing the complexities and multifaceted nature of the factors influencing such health events.

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 the training and testing sets.

To avoid potential bias and lack of robust data by maintaining sample size, we imputed the averages of the ‘bmi’ column in both the testing and training data. 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 focused on variables directly relevant to stroke prediction. The 'smoking_status' variable, which included an 'Unknown' category, was another area of focus. Rather than replacing these unknown values or omitting them, they were retained as a separate category since incidences of ‘Unknown’ were far more extensive than other levels in ‘smoking status’. 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.

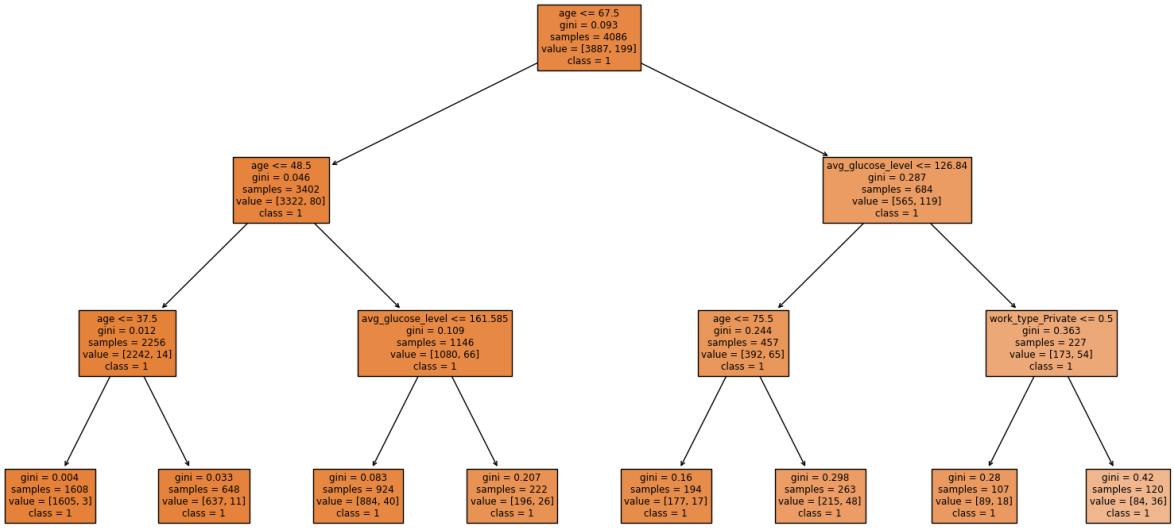
A step in the cleaning process we had to make unexpectedly was to remove the “other” observation in the gender variable. This observation was present only once in the training set and was not present at all in the test set. When using the training and test sets for the learning methods, we ran into problems where the same factors were not present in both the training and test sets. Since it was only 1 observation in the “other” category, we felt comfortable removing it so we could proceed with the methods.

However, these data cleaning and preparation steps presented certain challenges. The inclusion of rows with ‘Unknown’ smoking_status 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

Decision Tree Classifier

Figure 1: Decision Tree Classifier Depth 3



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:

- The most significant variables at this depth are age, average glucose level, and work type.
- Individuals over 67.5 years and with an average glucose level greater than 126.84 and who work primarily in the private sector are most at risk for stroke as seen by the highest gini value for that classification.
- Individuals less than 37.5 years old are least likely to be categorized as having a stroke.

The tree indicates that age is the most determining feature followed by average glucose level, then private work type. The performance metrics of the classification model, reflected by the confusion matrix, show a substantial imbalance. While the model has a high overall accuracy of 95.17%, this figure is primarily due to its ability to predict the more prevalent non-stroke class correctly. This is evidenced by the 973 correct predictions of non-stroke cases (true negatives) and the complete absence of stroke predictions, resulting in 50 stroke cases (false negatives) that the model failed to identify. There were no instances of true positive predictions for stroke, indicating that the model has a significant limitation in identifying actual stroke events. In a health context, this is especially harmful as false negatives could result in not getting the proper treatment and worsening conditions. The RMSE for each depth of this tree ranged from 0.22 to

0.3. As seen further in the results, this is the highest RMSE of all three methods and indicates that this method is not suitable for predicting the likelihood of stroke.

Linear Regression:

```
0.0663549633394519
0.2078679318645726
0.08417775915375991
0.20670024019739772
```

Due to the response variable (stroke) being a dummy variable, we didn't use linear regression models very extensively here. However, we did experiment with a couple of models: one in which only the numeric variables (age, avg_glucose_lvl, bmi) were considered, and one in which all the variables were used. Both of these models did not appear to be very predictive at first glance, given that the R-squared values for both were below 0.10 (the former being slightly lower than the latter at 0.6 and 0.8 respectively). The RMSE for the model with just numerical variables was 0.207. The RMSE for the model for all variables was 0.206. This means that out of these two models, we would prefer the model with all variables, as it has the lowest RMSE value and highest R-squared value. What is interesting here is that the RMSE is actually quite similar to our K-nearest neighbors model - at 0.2, it implies that the average error was within 20% of any given stroke value.

<u>K-nearest Neighbors</u>

Figure 2.1:

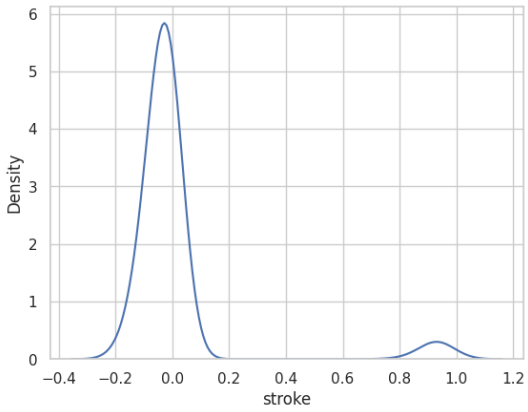
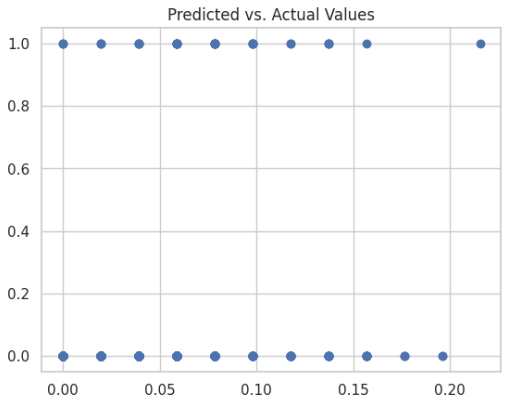


Figure 2.2:



The low R^2 indicates that there are likely very many other variables, not included in our regression/part of the error term, which contribute to stroke occurrences outside of

```
R2: 0.037308058829968904
RMSE:  0.21154832766265794
```

just the variables we included. This is not surprising, as someone's health is a complex function of many, many factors (diet, exercise, location, genetic predispositions, etc). The RMSE is approximately 0.211, indicating that the model's predictions are on average within 20% of the actual stroke values in our maxmin normalized scale. Given that kNN is a non-parametric, instance-based learning method, this level of accuracy suggests that our choice of features, the number of neighbors (k), and our distance metric are reasonably well-tuned for capturing the underlying patterns in the data. We did winsorize the data, so it is likely not outliers that are preventing the model from doing a better job. It is also possible that there are irrelevant features we included which need to be revised to improve the accuracy of the model. Lastly, perhaps there is a better method of dealing with our missing values, particularly in the BMI category (we replace missings with the mean), which if handled better would increase the performance of our model.

Conclusion:

Our analyses explored how different factors affect the likelihood of an individual experiencing a stroke. During our work, we took into account several different factors, including BMI, age, gender, and glucose levels. We ended up experimenting with three different models throughout the project, including Decision Trees, Linear Regression, and K-nearest neighbor. Overall, we found that decision tree classification has the highest RMSE, followed by KNN, and linear regression models had the lowest RMSE. However, from the decision tree, we did feel that we gleaned some useful information about factors that go into predicting strokes, most notably age, average glucose level, and work type. Based on a holistic view of RMSE and R2 values, we found that linear regression with all variables does the best job of predicting the likelihood of stroke. However, since regression is more useful for numerical response variables, we believe that KNN is the best method. It had the second-lowest RMSE of the three methods we tried.

One specific criticism we could face is the fact that we removed outliers in BMI and average glucose level. This ensures that the methods can predict the outcome. However, in real life, very high BMIs and average glucose levels could have more instances of stroke than strokes within the winsorized range. It does decrease the validity of predicting strokes in a real-life context, but it is still important to remove outliers to test the methods we used.

For future studies, we would recommend researching more categories to find ones that work better with our end goal (predicting the possibility of a stroke). Some possible options could include alcohol use, previous strokes, income, and other variables that would appear to have stronger correlations. We would also be interested in investigating a response variable that is related to strokes but is not a dummy variable - one potential route of investigation would be to look at how *many* strokes/other heart events a person experienced during their lifetime, and then go from there. That could potentially make it easier to use a K-nearest neighbor or Linear Regression model. Overall, however, we do feel like certain parts of our analyses (such as the decision tree splits) gave useful information regarding factors that affect the likelihood of

experiencing a stroke, and we believe that further research would further illuminate this important topic.