Predicting Stroke Events

Adam Brenner

The purpose of this project was in seeing which machine learning algorithm, decision trees, k-nearest neighbors, or categorical naïve bayes is best at predicting a stroke event based on a few life features. The data was available via Kaggle here, <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>, and consisted of the following feature variables and target variable. This is important due to the fact that every 45 seconds someone in the United States experiences a stroke. Thus, the accurate prediction of such stroke events could be vital for recovery and survival, as well as reducing healthcare costs.

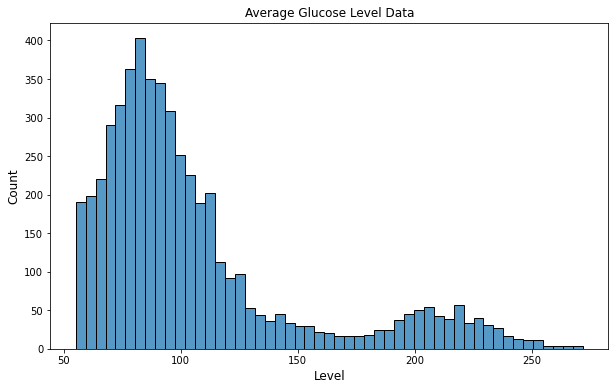
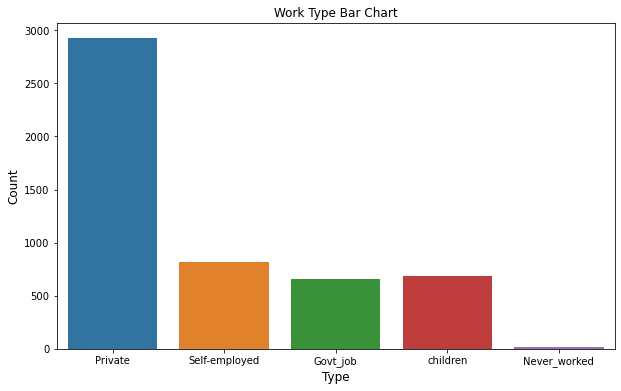
Features:

* ID: unique identifier
* Gender: male, female, or other
* Age: patient age
* Hypertension: 0 if no hypertension, 1 if hypertension
* Heart Disease: 0 if no heart diseases, 1 if heart diseases
* Ever Married: no or yes
* Work Type: children, govt. job, never worked, private, or self-employed
* Residence Type: rural or urban
* Avg Glucose Level: average glucose level in blood
* BMI: body mass index
* Smoking Status: formerly smoked, never smoked, smokes, or unknown

Target Variable:

* Stroke: 0 if no stroke, 1 if stroke

The dataset is comprised of 5110 rows and 11 columns. The feature work type is shown in a bar chart below. This was done for each categorical feature above to gain some understanding of the data distributions. Another feature, average glucose level, is also shown below in a histogram chart. Again, this was done for each numerical feature.

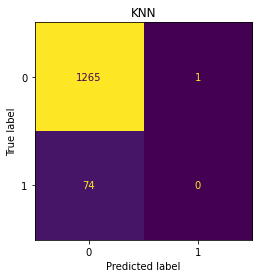
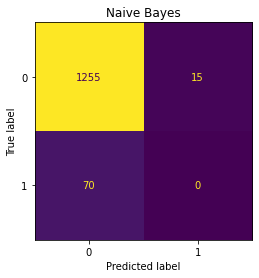
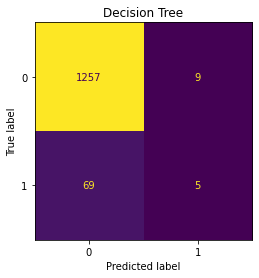


Data preprocessing involved changing the categorical features, gender, ever married, work type, residence type, and smoking status to numerical. This was done by simply changing the data type to a category type. Then, using pandas changed each unique value in each specific column to a numerical value from 0 to n-1 classes. The categorical columns were then dropped from the data frame, as well as the id column. Outliers were then checked using descriptive statistics for the numerical features. This did reveal some unusual minimums, particularly in the age and bmi columns, where each had a minimum value of .08. Upon further investigation, the age column had a lot of very young values. Subsequently, all rows where the age was at or below 14 were dropped from the dataset. This was mainly due to ages 14 and below having had only 1 stroke event and 644 non stroke events. Which would hopefully help the imbalance in the dataset, which will be discussed during model evaluation.

A feature and target set were created. The feature set comprised of all columns except stroke, which comprised the target. A standard scaler and min-max scaler were then applied to the feature set, creating two sets of features, one that had been standard scaled, and one that had been min-max scaled. The reasoning behind the min-max scaler, is that the standard scaler could introduce negative values, which the categorical naïve bayes cannot handle, while the min-max only scales between 0 and 1. Likewise, training and testing sets were created, one set comprised of the standard scaled features (standard features), and one comprised of the min-max scaled features (min-max features). A decision tree, categorical naïve bayes, and k-nearest neighbor model, were trained and fitted to the data. After making predictions, each of the models produced the accuracy scores below.

* Decision Tree Accuracy: 0.9126865671641791
* Naïve Bayes Accuracy: 0.9365671641791045
* KNN Accuracy: 0.9417910447761194

While this may seem impressive, as stated above, the dataset is heavily imbalanced, with a total of 4,217 no stroke events or 0’s, and only 248 stroke events or 1’s. With this imbalance, some more digging was required. This involved creating a confusion matrix for each model, shown below.



From these individual confusion matrices, the decision tree model actually performed the best out of the three, even though it had the worst accuracy above. It correctly identified 5 actual stroke events while the other two identified 0. For this reason, the decision tree was picked as the best model for stroke predictions, but far from being actually able to predict real stroke events. To test this theory, a for loop was used and a decision tree was trained and fitted on differently split data each time. The results are shown below, but they do cement the fact that the decision tree is the best model out of the three. It is fair to note, that the decision tree was allowed to grow to full depth.

