Programming Lab #2: Predictive Stroke Algorithm

DS 3001: Foundations of Machine Learning

University of Virginia

Eric Nguyen, Elaine Zhang, Hieu Vu, Cheryl Bai, Ashley Luk,

Adam Chow, Bryant Chow

Fall 2023

SUMMARY (~350 words)

Question

For our second project, we are exploring different tools to create a predictive algorithm to predict the likelihood of a person having a stroke. Our training and testing datasets leverage patient different data such as age, blood sugar levels, body mass index (BMI), marital status, gender, heart disease and hypertension presence, residence type, smoking status, and employment type.

Methods

First, we began the data exploration with a heatmap to examine relationships between variables, finding no significant correlations among the independent variables. One of our major findings was the increased likelihood of strokes in older individuals, and a slight increase in BMI among those who had experienced strokes.

We also explored various bar plots and histograms to assess the distribution of our data values for these continuous variables and categorical variables. Our scatter plots, on the other hand, showed strokes against glucose levels, age, and BMI, but did not reveal a strong relationship. However, we observed that individuals with higher glucose levels might have a higher stroke probability, suggesting a correlation that warrants further investigation.

We then focused our predictive models on variable correlations and their indirect effects on stroke occurrence. We experimented with linear models, decision trees, and kNN, adjusting hyperparameters for optimal results. Although continuous predictions were made using regressors, we recognize that classifiers would be more suitable for discrete predictions, aligning better with the binary nature of the stroke variable. This approach highlights the importance of aligning model selection with data characteristics and project goals.

Results

Overall, we noticed that our best linear model was the simple linear model with a R2 Score of 0.082 and a RMSE of 0.206. When we ran our decision tree model, our best model was at a max depth of 2 with an RMSE of 0.207 and a R2 Score of 0.079. Both of these models are very similar to the given default code. For our third KNN model we decided to attempt a classification approach rather than the regressor method. This change ultimately made all of the difference, resulting in our best model by far with a RMSE with best k: 0.037 and a R2 Score with best k: 0.97 at a k value of 5. This is our best model of predicting the likelihood of a stroke.

DATA

Both the training and testing datasets provide an expansive list of variablesthat are all weaved into our overall models. Here, we have a brief summary and description of each of our variables used in our stroke predictive models:

Age (Patient age, numeric): Age is a critical factor in stroke risk, with the likelihood often increasing as people get older.

Avg_glucose_level (Blood sugar levels, numeric): High blood sugar levels can be indicative of diabetes or other health conditions, which are known risk factors for stroke.

BMI (Body mass index, numeric): BMI can indicate obesity or being overweight, both of which are significant risk factors for stroke due to their association with high blood pressure and cardiovascular disease.

Ever_married (Ever married, dummy/character - Yes, No): Marital status could indirectly indicate social support systems, stress levels, and lifestyle choices, all of which can impact stroke risk.

Gender (Male, Female, or Other, character): Gender can be significant as stroke risk and symptoms can vary between males and females, potentially due to hormonal differences or lifestyle factors.

Heart_disease (Has heart disease, dummy): The presence of heart disease is a strong predictor of stroke risk as it is directly related to the health of the cardiovascular system.

Hypertension (Has hypertension, dummy): Hypertension or high blood pressure is one of the leading causes of stroke, making this a key variable for prediction.

ID (Study identification number): While not directly related to stroke prediction, the ID is crucial for data management and ensuring the integrity of the analysis.

Residence_type (Type of residence, dummy/character - Urban, Rural): The type of residence can be a proxy for environmental factors, lifestyle, and access to healthcare, all of which can influence stroke risk.

Smoking_status (Former, never, or current smoker, categorical): Smoking is a well-known risk factor for stroke due to its effects on blood vessels and the cardiovascular system.

Work_type (Employment type - Never worked, Homemaker, Public sector employment, Private sector employment, Self-employed): Employment type can reflect lifestyle, stress levels, and socioeconomic status, which are all relevant to health and stroke risk.

Stroke (Suffered a stroke in the sample period): This is the dependent variable we are trying to predict, indicating whether or not the individual has had a stroke.

Our initial step included using the functions `.columns` and `.shape` to verify the successful import of data into Colab, ensuring all necessary columns were present. Notably, the training data was approximately four times larger than the testing data, a beneficial aspect as it allowed our models to learn from a diverse set of patterns, enhancing their generalization ability and providing a comprehensive representation of the data distribution.

Next, we employed the '.info' function to assess the variables, producing the number of non-null values and data types of each variable. This step ended up being extremely crucial as it revealed that the BMI variable contained missing values – 159 cases in the training data and 42 in the testing data. Given the importance of BMI in stroke prediction, we decided to impute these missing values with the median BMI of the dataset. This approach, we believe, was more advantageous than removing these observations as it preserved the dataset's integrity and provided a more accurate representation of BMI's impact on stroke risk. The median was chosen over the mean due to its resistance against skewness and outliers, often present in real-world medical data.

After identifying and imputing the missing values, we examined the different distributions of our data through histograms. We observed that the average glucose level was right-skewed, while BMI displayed a relatively normal distribution centered around 25, and age was evenly distributed without a specific pattern. Furthermore, we decided to drop the "Unnamed: 0" and "id" columns, as they were irrelevant to our predictive models.

While running our `.info` function, this line of code indicated that certain columns, specifically hypertension, stroke, and heart disease, were classified as `int64` data types. However, their nature in our dataset classify as dummy variables representing binary (Yes/No) responses. Thus, they needed to be reclassified as such. This adjustment was integral for our analysis, as it aligned the data types with the variables' true nature, ensuring accurate model interpretation and results, particularly in the classification models.

RESULTS

Data Exploration Findings

To begin our data exploration, we used a heatmap and found there were no significant relationships between the independent variables of the dataset.

Our first major finding was that older people were much more likely to have a stroke, as seen on the plots of figures 1 and 2. Also, figure 3 shows that people who have had strokes have a slightly higher bmi.

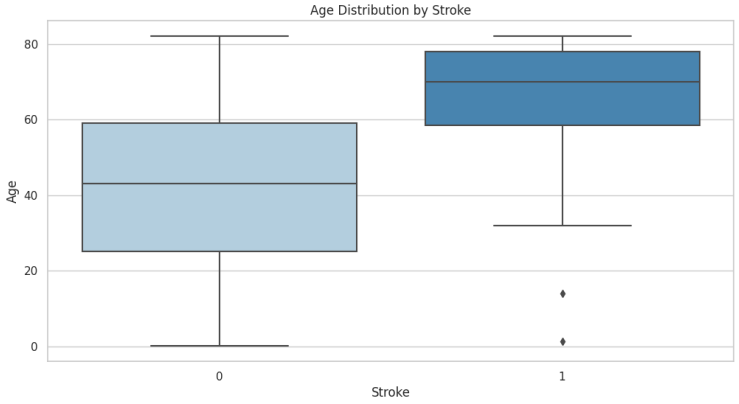

Figure 1. Age Distribution Boxplot by Stroke

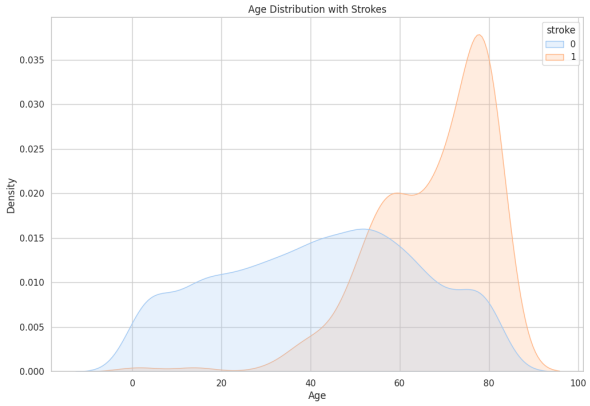

Figure 2. Age Distribution by Stroke

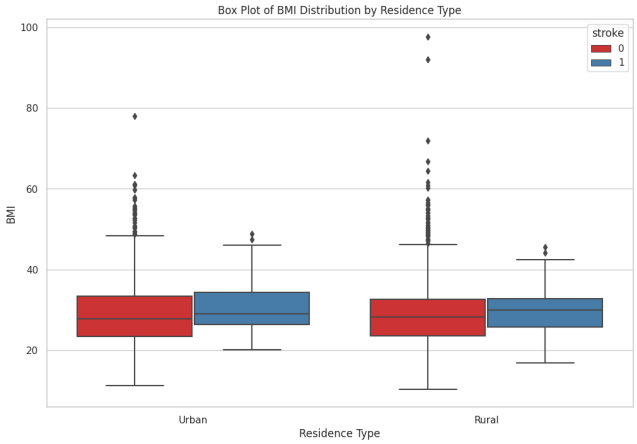

Figure 3. BMI Distribution Boxplot by Stroke and Residence Type

We also believe that it's important to note the potential correlation between glucose level and stroke chance. Figures 4 and 5 show strokes by glucose level plotted against age and bmi respectively. On each plot, there doesn't seem to be any major relationship between glucose level and stroke as the people with strokes seem to be evenly distributed across glucose level. However, of the overall population, the majority of the subjects have lower glucose levels; therefore, since the distribution of strokes across glucose level is somewhat even, one could argue that a greater percentage of those with high glucose have strokes than those with low glucose. We believe with more data this correlation could be investigated further.

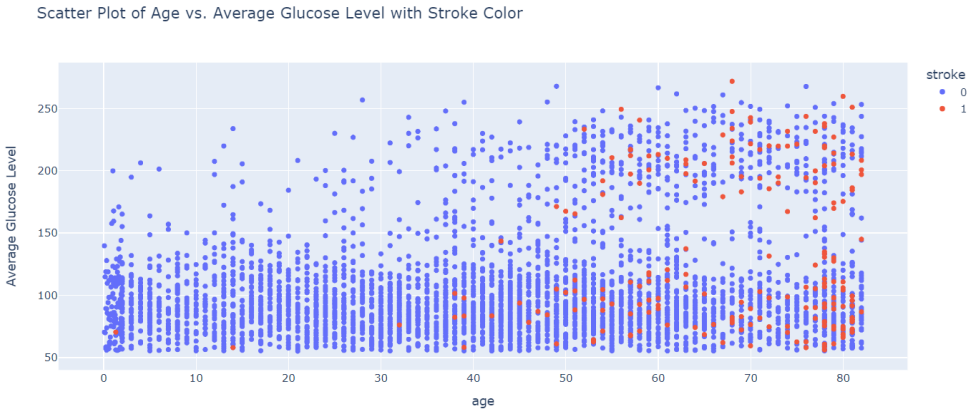

Figure 4. Glucose Level vs Age vs Stroke

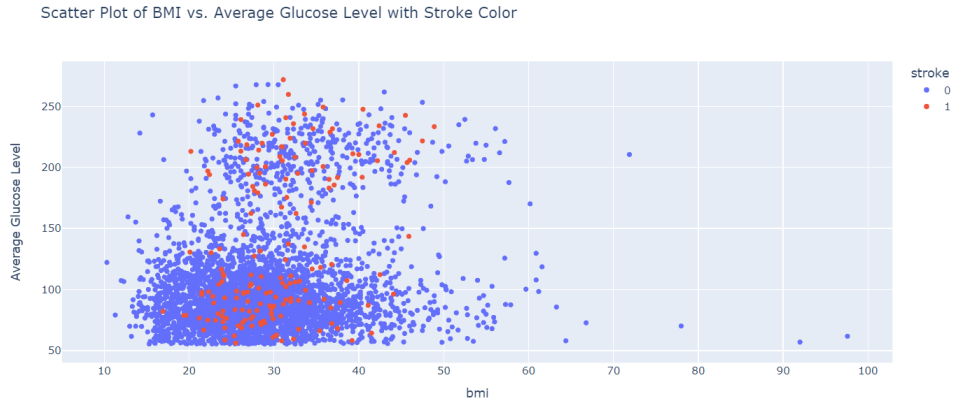

Figure 5. Glucose Level vs BMI vs Stroke

There isn't a visibly strong correlation between smoking and strokes when directly plotted against each other. Although, due to the fact that old age is correlated with stroke likelyhood, it's important to note that smokers and former smokers are found to be slightly older as seen in figure 6.

Figure 6. Age by Smoking Status

Model Performance

In the simple regression model, the yield is an R^2 of 0.0819 and an RMSE of 0.206. Then, the linear model with degree 2 polynomial features slightly decreased in performance (R^2 = 0.0523, RMSE = 0.2099), while degree 3 polynomials led to extremely poor results, indicating overfitting.

Next, The decision tree with a best fitted max depth of 2 resulted in a similar performance to the initial simple linear model (R^2 = 0.0793, RMSE = 0.2069).

Lastly, our initial runs of the KNN model produced a fitted k value around 300, also outputting a similar performance to the previous decision tree model and simple linear model. However, our best model to where the optimal k value is 5. This K value gave us our best R^2 = 0.9706 and RMSE = 0.037. In order to predict the likelihood of a stroke in a person, we should use this KNN model.

Methodology

Our group did well with data exploration as we investigated possible variables that could have correlations to strokes. Furthermore, we looked at inter-variable correlations to see how the relationships between variables could indirectly influence strokes.

Part of the challenge of this assignment was to use the models we learned in class (linear models, decision trees, and kNN) to reach the lowest possible RMSE. Our group tried several different hyperparameters to meet optimal results for each of the models leading to decent results.

However, if it wasn't for the challenge of using the models listed, we would have taken a different approach. The stroke variable of the dataset is a categorical variable that contains binary values representing whether or not a subject has had a stroke. In this assignment, we used regressors; therefore, our numerical predictions were continuous and not binary. In order to match with the stroke variable and better simulate a categorical guess, we would instead use classifiers in order to create discrete predictions. This change would also allow us to report accuracy.

CONCLUSION

In conclusion, our group was able to properly investigate variable correlations and train and test predictive models. We found that age and bmi had a positive correlation with strokes. Also, since smoking status had a positive correlation with age, it could have an indirect effect on strokes. Last, we believe that the patterns in the relationship between glucose levels and stroke warrant further investigation.

After some tuning, all of our models were able to present moderately accurate results. The optimal k nearest neighbor by far performed the best, with decision trees and linear regression

having similar but worse results. Although for categorization problems one would usually use a classifier, we used regressors for this assignment. The regressors are not completely invaluable for this scenario as the greater the decimal value the greater the confidence the model has in that instance that a stroke occurred.

With more time, our group would first further explore the correlation between glucose levels and strokes. To do this, we could explore the ratio of total subjects above a certain glucose level threshold to the number of subjects who've had a stroke above the same threshold. We would then compare that ratio to the same ratio but below the threshold. Also, more data could be collected to make the distribution along glucose levels more balanced to potentially yield a more clear picture.

Also, we would train and test classifier models to compare to our regressors. Classifier models would also allow us to report accuracy.