Group 7 Project #2 Final Paper

Authors: Cole Anderson, William Giles, Lexi Van Metre

Summary:

Our group worked on this project to predict the likelihood of stroke occurring in individuals through utilizing different machine learning models in pertinence to specific data regarding personal, health, and demographic elements. Some of these were age, gender, marital situation, whether the individual smoked, residential status, an established heart disease, and so on. These data points, all relevant in some way, aided us in reaching our goal. Our initial method involved using a linear model with said data that was converted to numerical status, which was necessary in creating such a model. However, from this initial approach, our group's findings deemed that the linear model was rather weak in terms of prediction capabilities. This was evident through the R^2 value being 0.189, and the RMSE value of 0.194. This signified that the model explained about 19% of the variability of the stroke, which leaves about 81% unexplained. The next step involved our first version of KNN, where we implemented this methodology. After doing such, we found similar results to the initial linear approximation. RMSE for this model was 0.179, while the R^2 value increased to about 0.307. The decrease in RMSE value proposes that the initial KNN approach was more effective than the linear model, though. Thus far, it seemed that our methodology was becoming more efficient. Following the first KNN model, we proceeded to perform a second one. In this version, we cleaned the data further, dropping unknown rows, and converted other categorical variables such as 'smoking_status' to numerical using dummy variables. This model also used a classification approach, rather than a regressor method utilized in the first KNN approach. In turn, we found a very high R^2 value of 0.968, and a very low RMSE value of 0.038. These results were more efficient than the original model provided, our initial linear model, and the first KNN model. Additionally, classification trees were utilized. However, our group found that such methods weren't useful. The tree we created only predicted two points on a graph, thus providing an RMSE of 0, and an R^2 of 1 (though there were more data points). Ultimately, the KNN model approach (specifically the second one) "wins out" over the others in predictability and accuracy of a stroke, which was our initial goal.

Data:

The data sets provided numerous variables which all factored into our project and goal in predicting the likelihood of a stroke occurring in an individual. The numerous variables within this dataset reflect the numerous factors that are accounted for in stroke probability. Below is a brief summary discussing each of the variables and their significance.

Age (numeric):

- Provides the individual's age, which is significant in stroke probability in that an older person is more likely to be susceptible to stroke.

Average Glucose Level (numeric):

- Glucose level refers to blood sugar levels, which can signify other health factors or dilemmas which increase chance of stroke.

Body Mass Index (numeric):

- This variable helps signify whether one may have obesity or other underlying stroke-inducing problems.

Ever Married (dummy):

- Refers to the individual's marital status, which can relate to marital conflict increasing stress levels, thus stroke levels.

Gender (categorical):

- Refers to male, female, and other, looking at how different sexes play into stroke probability. The National Institute of Health found that men have more stroke burden, but women are associated with higher stroke occurrence in old age.

Heart Disease (dummy):

- Signals whether the individual has a present heart disease, which increases the chance of having a stroke. This is so because oxygen/blood flow is hindered if there are arteries blocked.

Hypertension (dummy):

- Does the individual have hypertension? If so, that would increase the chance of stroke due to high blood pressure causing clots, which blocks oxygen rich blood from going to the brain.

Identification (ID) (numeric):

- Study identification number for each individual, which is important to have in any dataset. However, we dropped this variable in our efficient models. We dropped this variable in the second version of the KNN model, as it was deemed irrelevant.

Residence Type (dummy):

- Refers to either urban or rural living conditions for each individual. This can be applicable, and it is found that rural living conditions are associated with higher stroke occurrence.

Smoking Status (categorical):

- Asks individuals if they are a former smoker, if they never have smoked, or if they are a current smoker. Obviously, smoking leads to many health implications, thus worsening and heightening conditions of a stroke.

Work Type (categorical):

- Goes through different types of employment options, such as government, self, private sector, homeworker, and never worked. Higher stress job environments are associated with higher chances of a stroke occurring.

Stroke (dummy):

- Specifies whether the individual suffered from a stroke during the sample period. It's essentially the target variable in our predictive models.

To reiterate, the above data and key variables are relevant in some way, aside from some missing data points. They all shed light on the possibility of having a stroke, in the sense that depending on the individual values of each, it might be more plausible for a stroke to occur. In regard to challenges in readings, cleaning, and preparing the data for analysis, there really weren't any. Obviously, we had to remove irrelevant features by .drop(), such as 'id', and had to convert various variables into numerics for data analysis through using dummy variables, such as 'smoking_status' and 'residence_type'. Utilizing dummy variables was a big part of our project in order to keep everything proper for analysis. Nonetheless, nothing deemed too challenging with the usage of notes and concepts learned from class. We were able to complete our original task.

Results

To analyze the overall results of the project, it is important to view this project holistically as a continuous process of model development and refinement. When trying to accurately predict the likelihood of stroke, different models were implemented and explored. We started utilizing tree-based models as predictors for our response variables, but quickly pivoted to experimenting with clustering based models to better suit our classification task.

First, we will analyze the results of the tree-based models. The tree based model used initially was a classification tree with a maximum depth of 5.

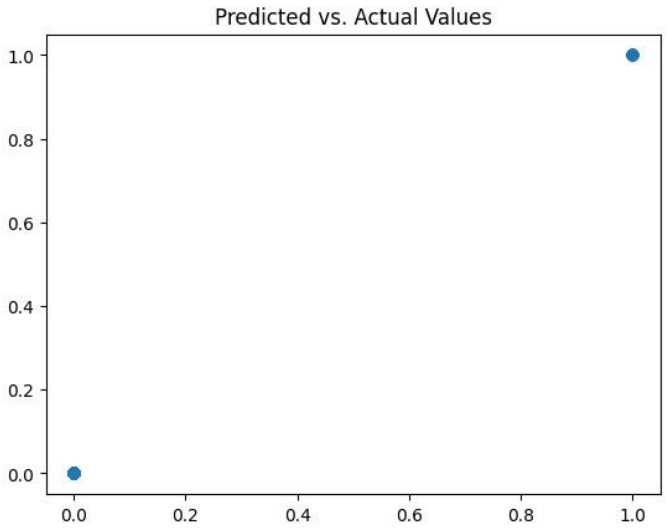


Figure 1

The presence of only two data points in figure 1 is quite unusual in this context, raising some concerns. There are a few different reasons that the graph could be taking this form. Because the decision tree is trained to output a binary outcome, it is possible that the model is too perfect and there are actually a cluster of points in the coordinates (0,0) and (1,1).

This suggests that the model could be overfitting to the training data, which means that the model could be too complex. These observations are in accordance with the recorded RMSE of 0.0 and the R^2 of 1.0. The RMSE of 0.0 indicates that there is zero error in the predictions, which would explain why the graph appears to have points in only the two coordinates mentioned. The perfect R^2 value also supports this observation. To further experiment with decision trees, we tried adjusting the depth of the tree to understand how that could change the results. Changing the depth did not cause any changes to the test metrics, indicating that this decision tree is flawed. Because of the unrealistic high accuracy of this model, we decided to pivot to clustering techniques to see if using that type of model could provide more adequate results.

Before delving into the clustering models, we quickly experimented with a basic linear model that only utilized numerical variables. The results of the linear model were as follows: an R^2 of 0.189 and an RMSE of 0.194. Again, these results are not promising and indicate a very poor model. This is because the use of a linear model is not ideal for a classification task because it tends to struggle with capturing non-linear relationships. Another basic model like logistic regression could be a more suitable option.

We next proceeded with the clustering models. First, we trained a KNN regressor model on the stroke data. We found the optimal k value by fitting models using various k values and determining the lowest SSE value. SSE is a measure of the difference between the data and the estimation model; a lower SSE indicates that there is less of an error, and the model is more closely fitted to the data. Specifically, SSE represents the sum of the squared differences between each observation and the overall mean.

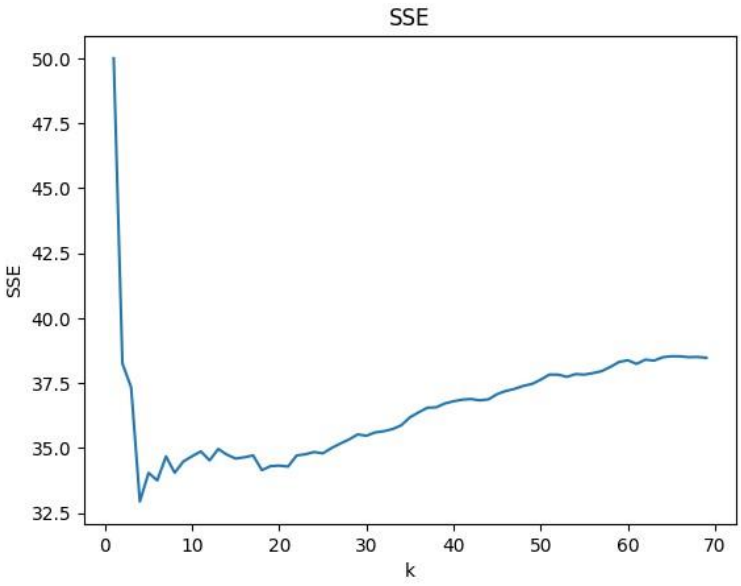


Figure 2

The SSE graph in figure 2 has a clear elbow point, which turns out to be at k=3, which indicates that this value of k should be used when training the model and predicting on new, unseen test data. Next, we looked at a plot of the residuals in figure 3a, which represent the difference between the observed values and the predicted values. The curve is centered at zero, which indicates that there is a random distribution of the errors. Additionally, the symmetry present in the graph suggests that the model is fairly accurate.

The plot of predicted vs actual values does not show promising results for the strength of the model: there is not a clear linear pattern, which would be expected. The pattern in figure 3b reveals that the model is not capturing the relationship between the variables and is severely underfitting, in contrast to the overfitted tree-based model. The RMSE of 0.179 and the R^2 of 0.307 further indicate the poor model performance.

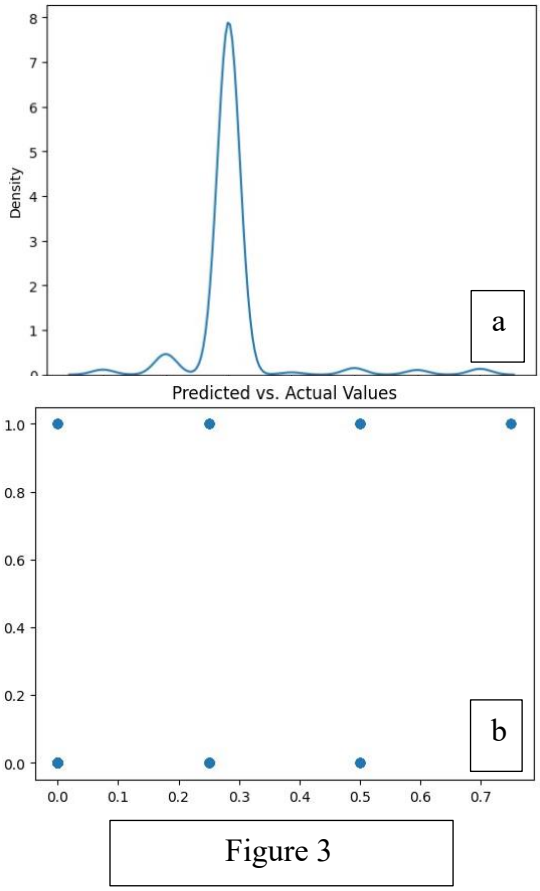


Figure 3

The second version of KNN relied on using a KNN classifier model. As done in the KNN regressor, we identified the optimal k value for the model by training various KNN classifiers, each with different k values. Ultimately, after determining the optimal k value through experimentation, the R^2 had a value of 0.968 and the RMSE had a value of 0.0380. This is therefore the most accurate model and the best at predicting the likelihood of stroke.

Conclusion:

Overall, the group is very pleased with the evaluation metrics for the final KNN classification model, and feels that it is a very good predictor of the likelihood of stroke. A very meticulous and strategic process was carried out to produce an accurate final model. This process consisted of a very detailed attention to cleaning the data and ensuring data that is well distributed and formatted for the models. Additionally, the model development process demonstrates a clear experimentation of several different types of models, showing the group's persistence and dedication to finding and training a successful and accurate model.

Although we did follow a strategic process, there are some areas that could have been improved. The most difficult component in any data science or analytics project is the data collection and cleaning phase. A model is only as good as the data it is trained on, so there is always room for improvement in this phase. Specifically, we could have sought out more data to handle issues like class imbalance, ensuring even more accurate results. Additionally, if this stroke model was to be deployed in a medical setting, it would be important to take note of where the data is collected, and potentially train several models on subsets of the data by geographic location. In terms of data cleaning, there is always more room for transforming variables and improving their distribution to better suit the model.

In terms of the model development phase, we also could have experimented with other models explored in the class. One model I wish we explored was logistic regression. This is a simple, traditional machine learning model that can be quite effective for classification related tasks, like the stroke likelihood predictor here. We also could have looked into more complex tree models, such as boosted trees and random forest models, to see if they yielded different results than the simple tree model we utilized.

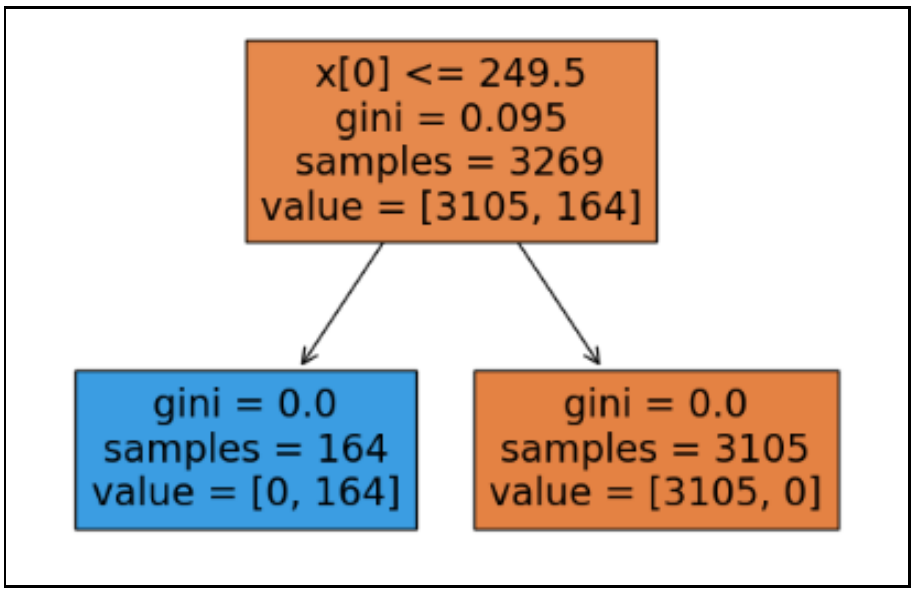
Despite the fact that we could have added more models, the ultimate model we ended up with is very pleasing and is a testament to the detailed data cleaning and model training that was conducted. Looking ahead, if we were to continue this project and further develop models, there are a few different routes that could be taken.

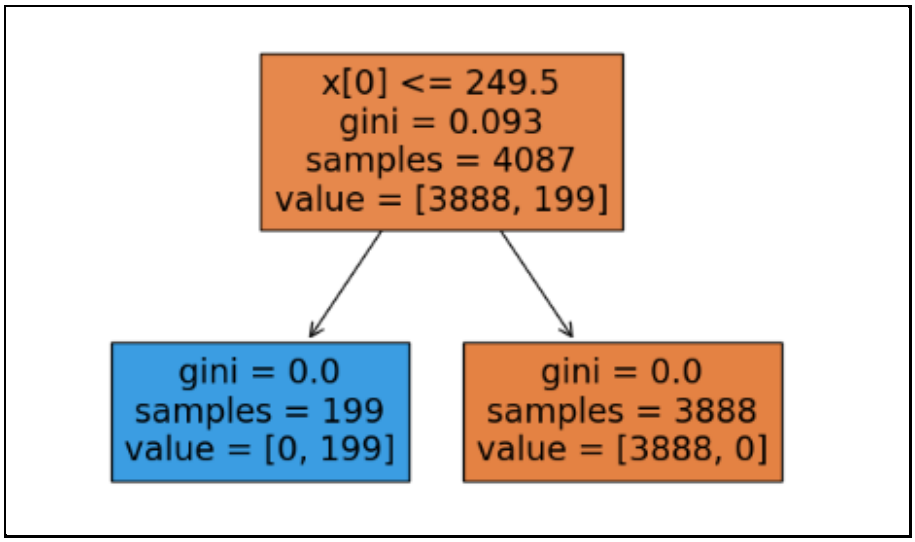
The first route would be to continue perfecting the KNN classifier model by using hyperparameter tuning techniques to further reduce the error and improve the accuracy of the model. One parameter that could be further experimented with is the k value, or the number of neighbors. This value determines how many neighbors are considered to predict the class of the data point. The distance metric used is another parameter that could be explored, which could influence the classes assigned to the new data points.

Another route that could be taken is exploring more complex models, such as neural networks that can be better at capturing complex relationships between the data. One type of neural network that we could have experimented with is a Convolutional Neural Network (CNN) with an output layer that uses a sigmoid activation function to predict a value between 1 and 0 (stroke or no stroke). This model is more difficult to implement, but has shown tremendous success in binary classification tasks, and we would be curious to see how this could influence the results of our project.

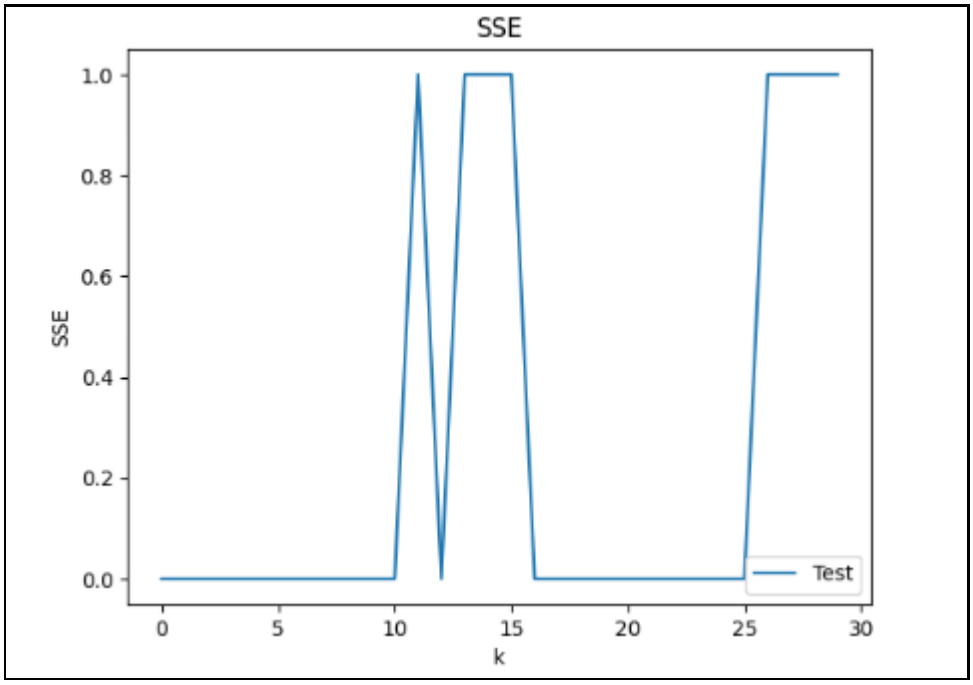
Appendix:

Classification Trees:




x[0] <= 249.5
gini = 0.093
samples = 4087
value = [3888, 199]
gini = 0.0
samples = 199
value = [0, 199]
gini = 0.0
samples = 3888
value = [3888, 0]

KNN:


SSE
1.0
0.8
0.6
0.4
0.2
0.0
SSE
0
5
10
15
20
25
30
k
Test