Group 21 – Project 2 Report

Suhaas Kanneganti, Pranav Arora, Victor Cruz Contreras, Derek Sprincis, Ayush Acharya, & Carlos Revilla

Completed in lieu of the final project.

Summary # Completed by Victor
This study examines the influence of various health indicators, notably BMI, age, average glucose levels, heart disease, and smoking status, on predicting stroke occurrences, by analyzing two datasets. The primary focus is on the 'avg_glucose_level', 'age', 'bmi', 'heart_disease', and 'smoking_status' columns. which indicate a person's average glucose levels, age, body mass index, heart disease, smoking status, and whether they got a stroke, respectively. After cleaning the dataset of missing values such as missing BMI values, histograms and clusters are generated to visually assess the relationship between BMI and stroke rates. Using histograms for visualization facilitated an immediate comparison of stroke rates among different clusters. One standout finding was the pivotal role of age, especially those older than 60 were significantly more inclined to get a stroke. Another major factor was glucose levels; in particular, those with higher glucose levels were more likely to have a higher stroke than those who had normal glucose levels, suggesting those with diabetes are more likely to get a stroke. In contrast, heart disease didn't seem to have any discernible effect on stroke rate, regardless of heart disease history as it mirrored similar patterns. Smoking status, on the other hand, seems to have an impact on stroke rates. Individuals who are smokers potentially exhibited a higher stroke rate than those who are not smokers. As current studies have significantly shown that there has been an increase in stroke frequencies with no signs of dropping, investigations of other potential determinants, like ethnicity, background, economic status, diabetic status, stroke history, cholesterol, blood disorders, and family history, would become the following fields of analysis if the research was to continue.

Data # Completed by Suhaas
The data analysis process examined two datasets, primarily focusing on clustering to predict the occurrence of strokes. The datasets included training and testing data, read from CSV files, as highlighted in the 'Code Documentation.txt' and the Jupyter notebook 'lab2.ipynb'.

The primary variable of interest was the 'stroke' variable, acting as the dependent variable (y-value) in the predictive models. This variable was removed from both the training and testing datasets for modeling purposes. Key features included 'avg_glucose_level', 'age', 'bmi', 'heart_disease', and 'smoking_status'. These were selected based on their predictive power for strokes.

In the preprocessing phase, missing values in the 'bmi' column were imputed using the mean BMI value. This approach is a standard practice in data handling to deal with missing values, though it assumes that the missing values are randomly distributed and similar to the mean of the observed values. For clustering, a combination of numeric and categorical transformations was applied. Numeric features ('avg_glucose_level', 'age', 'bmi') underwent standard scaling to normalize their distribution. Categorical features ('heart_disease', 'smoking_status') were transformed using one-hot encoding to convert them into a format suitable for modeling.

The k-means clustering algorithm was employed for the clustering approach. This involved experimenting with various combinations of the selected columns to determine the most effective grouping for stroke prediction. The final model used a pipeline that combined preprocessing steps and the k-means algorithm.

One significant challenge was dealing with missing values in the 'bmi' column. The choice to impute these using the mean value is practical but can potentially introduce bias, especially if the missingness is not random. Determining the most predictive features for stroke occurrence required trial and error. This process can be time-consuming and may not always yield a clear set of predictors, especially in datasets with numerous potential variables. The effectiveness of clustering approaches like k-means relies heavily on the chosen features and the number of clusters. Finding the optimal number of clusters (in this case, five) and ensuring that the clusters are meaningful and distinct can be challenging.

Once the clustering was performed, the data was analyzed to determine the stroke occurrence rate within each cluster. This analysis is crucial in understanding the patterns and characteristics of each cluster, but it also poses challenges in interpretation, especially if the clusters are not well-separated or have overlapping characteristics. The models and findings are based on the specific characteristics of the datasets used. This raises questions about the generalizability of the results to other populations or datasets. The Jupyter Notebook included visualizations like bar plots to illustrate the stroke rate by cluster. Effective visualization is key to interpreting the results, but it also requires careful consideration to ensure that it accurately represents the data without bias or misinterpretation.

The datasets provided a robust foundation for predictive analysis using clustering techniques. Despite challenges in data cleaning, imputation, and feature selection, the use of k-means clustering offered insightful results. The analysis highlighted the importance of careful data handling and preprocessing, as well as the critical role of feature selection in predictive modeling. The experience underscores the dynamic nature of data science, where data preparation and analysis are as crucial as the modeling itself.

Results # Completed by Carlos

Our analysis aimed to predict stroke occurrence using machine learning techniques. First, we preprocessed the data by utilizing techniques such as imputing a mean average for missing BMI values and one-hot encoding for categorical variables like smoker status and heart disease status. Using the data in a k-means clustering algorithm was done in the interest of discovering inherent groupings within the data.

Clustering Analysis

Numeric Cluster Characteristics:

The metrics used represent the mean values of features like average glucose level, age, and BMI for each cluster. These metrics provide insights into the distinguishing characteristics of each cluster. For instance, in the case of age, one cluster might represent younger individuals with lower average glucose levels, while another cluster could represent older individuals with higher BMI. As the average glucose level, age, and BMI differed significantly across clusters, we see a potential risk of certain classifications on the effect of stroke in a population. For instance, higher average glucose levels and age showed higher stroke occurrence rates, which suggests a correlation with these risk factors.

Categorical Feature Distribution:

We analyze the distribution of categorical features across the clusters, such as heart disease and smoking status (smoker or non-smoker). This helps to understand how these features vary across different clusters, indicating potential risk factors associated with stroke. Prevalence of Heart disease had shown highest in cluster 3 at 16.06%. Cluster 2, which had the highest percentage of non-smokers at 55.86%, showed a significantly lower stroke rate in comparison to Cluster 4 which had the highest proportion of smokers at 18.68%.

Our Numeric features further showed notable high average glucose levels at 207.07 and age at 60.35, which are known risk factors for stroke. In contrast, our Cluster 0 represented a younger demographic with an average age of 9.86, which had the lowest average glucose levels (93.24) and BMI (20.30) as expected.

Stroke Occurrence Rate Analysis:

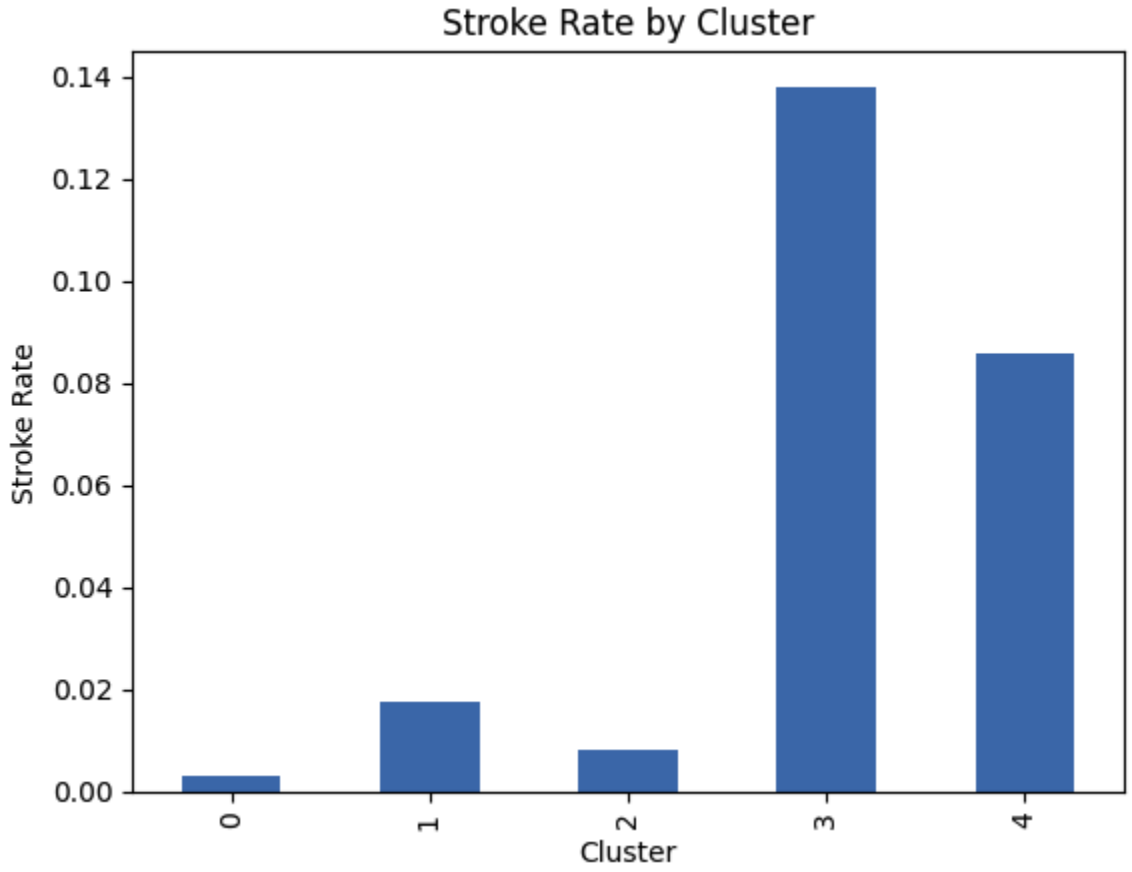

Figure 1. Stroke Rate by Cluster

Our analysis of the stroke occurrence rate within each cluster is crucial for understanding which clusters are at higher risk of strokes and could guide targeted healthcare interventions. This visualization allowed us to visually confirm the quantitative findings and effectively communicate the risk distribution. Notably, Cluster 3, characterized by older age and higher glucose levels, had a higher stroke rate, reinforcing the significance of these factors in stroke risk prediction.

Predictive Modeling

Our predictive models included a linear regression with polynomial features, a decision tree, and a stacked model combining both.

- The linear model outperformed the other models, achieving an R^2 value of 0.087 and an RMSE of 0.206.
- Interestingly, the decision tree model showed a value of negative R^2: -0.871 and an RMSE: 0.295, indicating a potential overfitting and lack of the factors ability to be generalized.

- The stacked model did not improve upon the linear model significantly, as reflected in its RMSE value: 0.226 and R^2 value: -0.094

From our quantitative factors, our models suggest that while age and glucose levels are shown to be significant predictors of stroke, BMI may not be as predictive when considered against other factors. This goes to show the complexity of stroke prediction and the necessity of selecting the appropriate features when modeling the data. Our linear model, which showed the best performance with polynomial features, suggests that the relationship between the predictors and stroke occurrence is not purely linear, which warrants further investigation into non-linear modeling approaches. Overall, the results emphasize the impact of age and glucose levels and the importance of feature selection in predictive modeling. The results of such can be very useful in informing healthcare strategies for targeted interventions and risk assessment protocols.

Conclusion: # Completed by Derek

The study provides valuable insights for predicting stroke occurrences, specifically examining BMI, age, average glucose levels, heart disease, and smoking status. Key findings highlighted the significant influence of age, with those over 60 being more prone to strokes. Higher glucose levels correlated with increased stroke risk while smoking status emerged as a potential factor. Challenges included handling missing BMI values and ensuring the generalizability of findings. Error was reduced by using k-means clustering and predictive modeling, emphasizing the importance of age and glucose levels in stroke prediction. A low error was confirmed by smaller RMSE values (<.295).