Group members:
Mariah Hudson, Zoe Van Winckel, Tsega Kidanu, Elizabeth Peterson, Hank Dickerson, Varun Pavuloori

Summary:

The purpose of this project was aimed at exploring the prediction of stroke risk using a detailed dataset. Our focus centered on how various factors, including age, BMI, average glucose levels, and lifestyle habits could influence the likelihood of an individual experiencing a stroke in their lifetime. The dataset, segmented into a training and testing group, was crucial for the training and validation of our models.

At the outset, a key focus of our project was on data cleaning and preparation. We addressed missing values, ensuring data integrity and normalized features like glucose levels and BMI to maintain consistency across the data range. Categorical variables, such as smoking status and work type were encoded to turn them into a format that was compatible with our algorithmic analysis.

In terms of data visualizations, we utilized a range of different tools to thoroughly understand the distribution and identify the potential outliers. Histograms and box plots were relied upon to get a sense of the spread and tendencies of key variables. In addition, correlation matrices offer certain insights into the relationships between different variables helping us see what variables are most predictive of stroke risk.

Overall, this project underscored the importance of machine learning in the field of health. It enhanced our understanding of how data-driven methods can predict health outcomes and emphasized the importance of meticulous data analysis in addressing complex issues. The project lays a foundation for further exploration into advanced machine learning techniques and shows a practical use in one of its numerous applications.

Data:

The dataset used in this project contained identifying factors to determine a patient’s likelihood of experiencing a stroke. The dataset contained information about various demographic, clinical, and lifestyle factors that might influence an individual's susceptibility to strokes. It included variables such as age, gender, hypertension, heart disease, average glucose level, body mass index (BMI), smoking status, and work type. The target variable, 'stroke,' signified whether an individual had experienced a stroke (1) or not (0).

Challenges included handling skewed distributions in variables like glucose levels and the imbalance in stroke outcomes, with far fewer stroke cases. We navigated these issues with appropriate statistical techniques. Furthermore, our exploratory data analysis (EDA) involved in-depth visualization to understand distributions and identify outliers. This comprehensive approach to data handling underlines our intentions in maximizing the accuracy of our model.

After importing the dataset, we cleaned it using the techniques described below:

- Missing values: We initially noticed 159 missing values in the BMI column in the training data frame and 42 in the testing data frame. We addressed this by replacing the missing values with the mean BMI value for each data frame.
- Categorical variables: We transformed categorical variables such as Gender, Ever Married, Residence Type, Smoking Status, and Work Type into numerical representations. This enabled us to incorporate the categorical data into our machine learning models and do a KNN. For example, in Smoking Status, 'Unknown' is replaced with 0, 'never smoked' is replaced with 1, 'formerly smoked' is replaced with 2, and 'smokes' is replaced with 3.

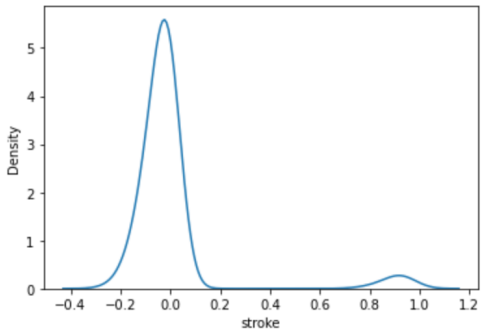
The table below lists and summarizes the relevant variables in our analysis, as well as the transformations we made.

Variable Name	**Description**	**Variable Values**
age	Patient age in numeric form	Unchanged
gender	Categorizes patients based on gender	Male == 0, Female == 1, Other == 2
hypertension	Binarily categorizes patients on whether or not they have hypertension	Unchanged (0 is does not have hypertension and 1 is has hypertension)
heart_disease	Binarily categorizes patients on whether or not they have heart disease	Unchanged (0 is does not have heart disease and 1 is has heart disease)
average_glucose_level	Lists the blood sugar level of each patient in numeric form	Unchanged

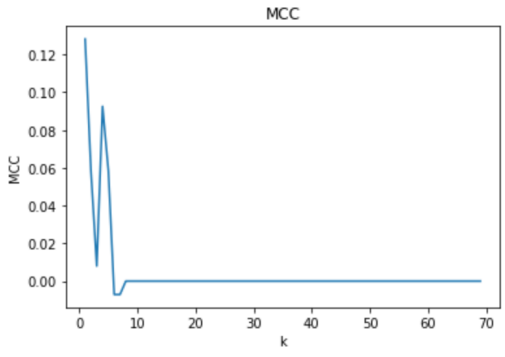
BMI	Body mass index of each respondent in numeric form	Unchanged
smoking_status	Categorizes patients based on whether or not the respondent smokes, considers both historically and currently	Unknown == 0, never smoked == 1, formerly smoked == 2, smokes == 3
work_type	Patient's employment type	Never worked == 0, private == 1, self-employed == 2, government job == 3, children == 4

Results:

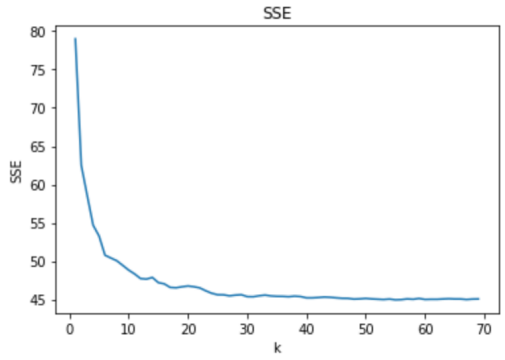
Our analysis began with visualizations, crucial for contextualizing our results. The initial visualization revealed a higher density of individuals without strokes (stroke value of 0) compared to those who have had strokes (stroke value of 1), indicating an imbalance in the dataset. This imbalance was a key consideration in our modeling approach, as it could potentially skew our predictions.



Our plot of the MCC, or the Matthews Correlation Coefficient, helps us understand the accuracy of our model based on the variation in k. The MCC plot, a measure for binary classification, indicated a peak value at a k-value of around 5. This peak suggests that the model achieves its optimal balance of sensitivity and specificity when considering the nearest five neighbors. As k increases beyond this point, the MCC approaches zero, indicating diminishing returns in model accuracy.



The SSE graph provided insights into the model's error magnitude. A significant decrease in SSE was observed around the k-value of 5, aligning with the findings from the MCC plot. This convergence indicates that our model minimizes prediction errors effectively at this point. Beyond k=5, the SSE tends to level out, reaching a plateau around 45. This trend suggests that increasing the number of neighbors beyond this point does not substantially reduce prediction errors, highlighting k=5 as the optimal parameter for our KNN model.



The R^2 value, standing at 0.0547, demonstrates that our model explains approximately 5% of the variance in the outcome variable. This level of fit is indicative of a relatively weak predictive power in the context of medical data. Nonetheless, perfect prediction is often unattainable due to the complex nature of health outcomes.

The RMSE value of 0.210 further exacerbates the bad performance of the model's . This relatively high value indicates that the model's predictions are far from the actual data, meaning the reliability of our model in predicting stroke occurrence with minimal error is low.

Our linear regression model also did produce satisfactory results. Most of the regression coefficients were very small, and some of them ended up being negative. Along with this the R squared value obtained was about .087. This tells us that the linear model could not explain much of the variance of the dependent variable. Despite this, the RMSE value we obtained from the model was .206. This low number indicates that despite not being able to explain variance in the dependent variable, it was fairly accurate when it came to predicting the outcome.

Conclusion:

This project aimed to predict whether or not someone would have a stroke in their lifetime based on various risk factors, such as age, BMI, glucose levels, smoker status, heart disease, etc.. This predictive model was made based on an extensive data set with twelve markers for prediction. In order to perform analysis on these variables, we had to convert the categorical variables into numerical values.

In this project, we explored the efficacy of two distinct predictive models: K Nearest Neighbor (kNN) and expanded linear regression model, both of which demonstrated limitations in performance. The kNN model, reflected by an R squared value of 0.0547, exhibited lower effectiveness in explaining the variance associated with stroke prediction. Conversely, the linear regression model achieved a slightly higher R squared value of 0.087, indicating marginally better performance in capturing the variance compared to the kNN model. Overall, both models faced challenges in explaining the variance within the dependent variable linked to stroke prediction. The linear regression model encountered difficulty in achieving a higher R squared value, potentially due to the binary nature of the outcome variable, restricting the model's ability to establish a linear relationship between the independent and dependent variables. We had a difficult time trying to get a relatively high R squared value when using the linear regression model. A reason for this might be that even though the independent variables in the linear regression could be significant in indicating a stroke, the binary output for the model (meaning the outcome simply can or can not happen) limits our ability to find a linear relationship between the independent and dependent variables. While there was a “better” model that had a higher R squared value, both models ultimately showcased their incapability of predicting strokes consistently. There were also too many false negatives for these to be applied in a medical setting.

Because of the nature of linear regression models and the binary output of the dependent variable it would have been better to use a different model other than linear regression in order to find the highest R squared value and lowest RMSE value possible.

A possible better fit for this project could have been to use a decision tree as that has a more practical use with a binary output from the dependent variable. It would be interesting to see if combining both K nearest neighbor and the linear regression model would have yielded better statistics in terms of the R squared and RMSE values instead of implementing them separately as we did.

Projects like this are extremely important for predicting health emergencies in order to put preventative measures in place. When we see from our predictive model what risk factors put people in the most imminent danger of a stroke, we are able to identify the most important factors to help people at risk for strokes.