AIN2002 - Stroke Predictions

Mohammed Diouri
2105264

Atena Jafari Parsa
2101183

Ava Arabi
2104906

Github URL: `https://github.com/MDiouri/AIN2002`

Table 1: Contributions

Members	Roles
Atena Jafari Parsa	Data Cleaning
Ava Arabi	Data Visualization
Mohammed Diouri	Data Modeling and Predictions

Abstract

A stroke transpires when a blood vessel in the brain becomes obstructed or ruptured, resulting in the impairment or demise of specific brain regions. The objective of this project is to apply the data analysis techniques acquired in the AIN2002 class, enabling the accurate prediction of stroke risks. This will be accomplished by utilizing both the comprehensive real-world dataset obtained from the National Center for Chronic Disease and the synthetic dataset sourced from Kaggle, employing linear regression.

1 Introduction

A stroke is characterized by the blockage or rupture of a blood vessel in the brain, resulting in the consequential harm or fatality of specific brain regions. The primary objective of this project is to apply the data analysis methodologies acquired during the AIN2002 class in order to achieve precise predictions of stroke risks. This will be accomplished by utilizing two distinct datasets: the complete dataset of real-world data acquired from the National Center for Chronic Disease, and the dataset of synthetic data sourced from Kaggle. Linear regression will be employed as the statistical technique for analysis.

2 Datasets

The datasets used in this project are publicly available from Kaggle.com

1. Real-world data: The Stroke Prediction Dataset
2. Synthetic data: The Synthetic Stroke Prediction Dataset

The clinical features in datasets are the following:

- id: unique identifier
- gender: "Male", "Female" or "Other"
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- 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"
- stroke: 1 if the patient had a stroke or 0 if not

The training set will encompass both the real-world data and the synthetic data, whereas the test set will solely comprise the synthetic data. It is important to acknowledge that the labels for the test set are undisclosed, rendering the true outcomes of the synthetic data unavailable for evaluation. To evaluate the model's performance on the test set, predictions will be submitted to the Kaggle competition from which the synthetic data was obtained. The competition will generate scores based on accuracy or other pertinent evaluation metrics, facilitating comparative analysis of the model's performance with that of other participants.

3 Data Cleaning

The initial step involves importing Pandas and NumPy libraries, followed by concatenating the two train sets using the pd.concat() method. It is important to set the axis parameter to 0 and the $ignore_index$ parameter to 1 to ensure a vertical stacking of the data frames and disregarding the original indices before concatenation.

To gain insights into the distinct categories and values within each column of the DataFrame object, we can utilize the .unique() function.

For assessing missing entries, the .isnull().sum() method is employed. The results indicate that the BMI (body-mass index) column is the only one with missing values, totaling 201 instances. To address this, a specific approach will be adopted: filtering the data based on age ranges and filling each missing value with the average BMI within that particular range. This strategy aims to enhance the accuracy of the data. Age ranges are defined from 0-90, with intervals of 10 years, and labeled accordingly. A new column denoting the age range of each individual is then created. Subsequently, the mean BMI for each group range is calculated. It is worth mentioning that no duplicate rows were detected in the dataset.

4 Data Visualization

Given the large size of the dataset, graphical representations would provide only a general overview. To obtain precise counts of occurrences for individuals who experienced a stroke and those who did not, the .groupby() function is employed. This function allows for grouping the data based on stroke occurrences, providing an accurate count for each category.

4.1 Categorical Data

To visualize the occurrence of strokes based on different categorical variables (hypertension, heart disease, age range, gender, marital status, work type, residence type, smoking status, BMI, and glucose level), count plots are created using the countplot() function from the seaborn library. Each count plot displays the count of individuals who had a stroke, categorized by the corresponding variable.

The count plots provide a clear representation of the distribution of stroke cases within each category. The x-axis of each plot represents the specific category, while the y-axis represents the count of individuals who experienced a stroke. The bars within each plot are differentiated based on whether the individuals had a stroke or not.

By analyzing these count plots, valuable insights can be gained regarding the relationship between each categorical variable and the occurrence of strokes in the dataset.

Note: The figures are available in the Jupyter Notebook.

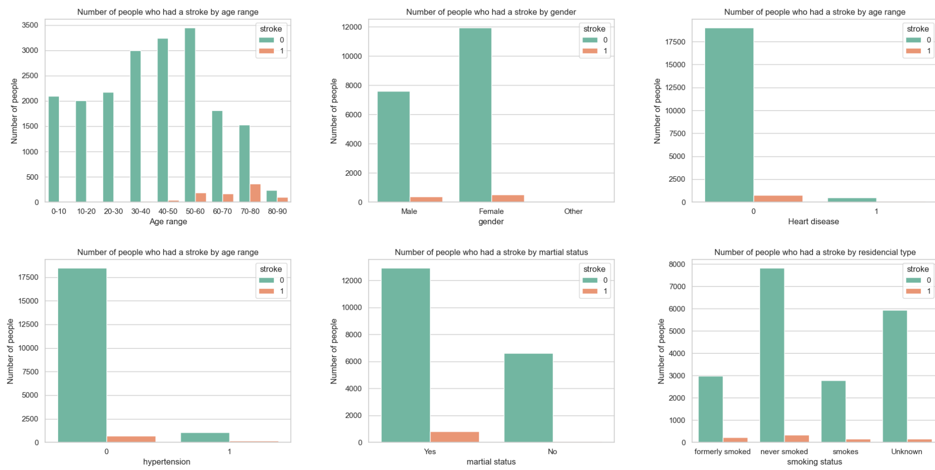

Figure 1: Categorical Plots

The analysis of the count plots reveals the following observations:

- Hypertension: The graph suggests that individuals without hypertension have a higher risk of experiencing a stroke.
- Heart Disease: The plot indicates that individuals without heart disease have a higher risk of having a stroke. It is important to note that this observation may be biased as it contradicts real-life observations where heart disease is a known risk factor for strokes.
- Age Range: The age range plot indicates that individuals between 70 and 80 years old have the highest risk of stroke. The second highest risk is observed in the age range of 50-60, followed by 60-70, 80-90, and the lowest risk belonging to individuals in the 40-50 age range. Notably, the dataset suggests that younger ages have no risk of stroke, which may not align with real-life scenarios.
- Gender: The gender plot demonstrates that females have a higher risk of experiencing a stroke, which aligns with real-life observations.
- Marital Status: The plot suggests that married individuals have a higher risk of stroke compared to other marital status categories.
- Work Type: According to the work type plot, individuals working in the private sector exhibit a higher risk of stroke compared to other work types.
- Residence Type: The residence type plot indicates that individuals living in urban and rural areas have an equal risk of stroke.
- Smoking Status: The plot reveals that individuals who have never smoked have the highest risk of stroke. However, it is important to consider that this observation may be influenced by biases in the dataset.

It is crucial to interpret these observations cautiously, considering potential biases within the dataset and comparing them with established knowledge in the field of stroke risk factors.

4.2 Numerical Data

In summary, the numerical data analysis reveals the following observations:

- The BMI plot demonstrates that the BMI values of individuals who had a stroke are normally distributed. The majority (approximately 95%) of BMI values fall within the range of 20 to 40, indicating that strokes are more prevalent within this BMI range.
- The glucose level plot suggests that individuals with an average glucose level between 50 and 100 have a higher risk of experiencing a stroke. This finding indicates a potential correlation between glucose levels and stroke risk, with values within this range being associated with increased susceptibility to strokes.

These summarized insights highlight the distribution of BMI values and the association between glucose levels and stroke risk in the dataset. However, it is important to consider these findings alongside other risk factors and consult established medical knowledge to obtain a comprehensive understanding of stroke risk.

5 Data Modeling and Predictions

Before making any predictions, it is crucial to prepare the data by converting categorical values into numerical format. This can be achieved by using dictionaries and the .replace() method. The same process is applied to both the training and test sets to ensure consistency.

Next, the linear regression module is imported from the $sklearn.linear_{m}odel$ library. The stroke column is dropped from the training set, and the stroke column itself is set as the target variable for training. After fitting the model, the .predict() method is used to generate predictions for the test set. The resulting predictions are then written into a CSV file.

Upon submitting the results to Kaggle, the predictions achieved an accuracy score of 90

Overall, this pipeline involves data preparation, model training using linear regression, prediction generation, and result submission to Kaggle. The high accuracy score obtained indicates the effectiveness of the model in predicting stroke occurrences.

6 Conclusion

The analysis revealed that several variables, including hypertension, heart disease, age range, gender, marital status, work type, residence type, smoking status, BMI, and glucose level, have implications for stroke occurrence. Future research can employ advanced statistical methods to further investigate these variables and their predictive capabilities. By gaining a deeper understanding of their relationships, researchers can enhance stroke prevention strategies and develop more accurate risk assessment models. Ultimately, this can lead to improved interventions and better management of strokes.

7 References

- Real-world data: The Stroke Prediction Dataset: `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`
- Synthetic data: The Synthetic Stroke Prediction Dataset: `https://www.kaggle.com/competitions/playground-series-s3e2/data`
- About Stroke `https://www.cdc.gov/stroke/about.htm`
- Martins et al., World stroke organization (wso): globalstroke fact sheet 2022.International Journal of Stroke
- Stroke - What Is a Stroke? NHLBI, NIH `https://www.nhlbi.nih.gov/health/stroke#:~:text=A%20stroke%2C%20also%20known%20as,begin%20to%20die%20within%20minutes.`