**Data Exploration Report for Healthcare Predictive Analytics**

**Data Exploration Report for Healthcare Predictive Analytics**

**Version 1.0**

**October 9, 2025**

1. **Introduction**

Stroke is the second leading cause of death worldwide, as reported by the World Health Organization, making early prediction and prevention a critical health priority. Predicting stroke risk can help identify vulnerable individuals and enable timely medical attention that may save lives.

In this project, we explore healthcare data to predict the likelihood of stroke occurrence using a combination of statistical methods and machine learning. Through this analysis, we aim to uncover patterns and relationships between health factors such as age, lifestyle, and medical history, and their influence on stroke risk.

The dataset used in this study is publicly available on Kaggle:  
<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

**2) Data Exploration**

**(A) Structural Exploration:**

The dataset used in this project is contained in a single CSV file named **healthcare-dataset-stroke-data.csv**. It consists of **5110 rows** and **12 columns**, including both numerical and categorical attributes that describe patients’ demographic, lifestyle, and health-related information.

Below is an overview of the dataset’s structure and the meaning of each column:

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description / Possible Values** |
| id | int64 | |  | | --- | | A unique identifier assigned to each patient. | |
| gender | object | Patient’s gender. Possible values: Male, Female, Other. |
| age | float64 | Age of the patient in years. |
| hypertension | int64 | Whether the patient has hypertension (high blood pressure). 0 = No, 1 = Yes. |
| heart\_deisease | int64 | Whether the patient has any heart disease. 0 = No, 1 = Yes. |
| ever\_married | object | Indicates if the patient has ever been married. Possible values: Yes, No. |
| work\_type | object | Type of work the patient does. Possible values: children, Govt\_job, Never\_worked, Private, Self-employed. |
| residence\_type | object | The type of residence area. Possible values: Urban, Rural. |
| avg\_glucose\_level | float64 | Average glucose level in the blood. |
| bmi | float64 | |  | | --- | |  |  |  | | --- | | Body Mass Index — a measure of body fat based on height and weight. | |
| smoking\_status | object | Smoking habits of the patient. Possible values: formerly smoked, never smoked, smokes, Unknown. |
| stroke | int64 | Target variable indicating whether the patient has experienced a stroke. 0 = No, 1 = Yes. |

After analyzing the dataset, we found that there are **no duplicate records** in the table. Among the numerical attributes, **missing values were present only in the bmi column**, while all other numerical columns were complete. For the categorical attributes, an inspection of the unique values in each column showed **no inconsistencies or unexpected entries**, indicating that the categorical data is clean and well-structured.

Here is a sample of the data:

A screenshot of a computer

AI-generated content may be incorrect.

**(B) Statistical Exploration:**

**Statistical Description:**

Below is the statistical description of the dataset:

A screenshot of a computer screen

AI-generated content may be incorrect.

**Outlier Detection**

Outlier detection was performed to identify unusual data points that could influence the statistical analysis or affect the performance of predictive models.

However, not all numerical columns in the dataset are suitable for this analysis. Columns such as **id**, **hypertension**, **heart\_disease**, and **stroke** contain identifiers or binary values (0 and 1), where outliers have no practical meaning. Therefore, we limited our examination to the three continuous variables: **age**, **avg\_glucose\_level**, and **bmi**.

The **Interquartile Range (IQR)** method was applied to these columns to detect statistical outliers. This approach identifies values that lie below the first quartile (Q1) minus 1.5 times the IQR, or above the third quartile (Q3) plus 1.5 times the IQR. The following table summarizes the number of outliers found in each

|  |  |
| --- | --- |
| **Feature** | **Number of Outliers** |
| age | 0 |
| avg\_glucose\_level | 627 |
| bmi | 110 |

(Boxplots were also generated to visually represent these outliers and their relation to stroke occurrence.)

From the visualizations, we observed that the avg\_glucose\_level and bmi features contained noticeable outliers, while age showed a more evenly distributed range of values. These detected outliers will be considered carefully in later stages, as they might represent true extreme cases rather than data entry errors.

**Key Features**

Before performing correlation analysis, the categorical attributes were label-encoded to make them numerically comparable with the target column. This straightforward encoding approach was sufficient for our purpose of gaining an initial understanding of the dataset, without delving into more complex statistical tests such as chi-square analysis.

After encoding, we computed the correlation of each feature with the target variable (stroke). The analysis revealed that **age**, **heart\_disease**, **avg\_glucose\_level**, **hypertension**, and **ever\_married** have the highest positive correlations with stroke occurrence. Among them, **age** shows the strongest correlation (≈0.25), followed by **heart\_disease** (≈0.13) and **avg\_glucose\_level** (≈0.13).

These results align well with **domain logic and medical intuition**: older individuals are naturally at higher risk, and conditions such as heart disease, hypertension, and elevated glucose levels are medically recognized as major risk factors for stroke. Even the correlation with marital status may reflect lifestyle or stress-related influences observed in healthcare studies.

A bar chart was used to visualize the **top five features most correlated with stroke**, offering a clear view of the strongest predictors that are likely to play a key role in building the model.

**Distribution:**

To better understand the composition of the dataset and the characteristics of the population, we examined the distribution of both numerical and categorical attributes. This helps reveal how the health and lifestyle factors are spread, detect imbalances among categories, and identify patterns that may influence stroke risk.

We analyzed the **numerical health indicators** — age, avg\_glucose\_level, and bmi — using histograms with kernel density estimates (KDE).

The distributions show that:

* **Age** is right-skewed, with most individuals being middle-aged or older, which aligns with the known higher stroke risk in older populations.
* **Average glucose level** also exhibits a right-skewed distribution, indicating that some patients have considerably higher glucose levels, possibly reflecting diabetic conditions.
* **BMI** values roughly follow a normal-like distribution, centered around the healthy-to-overweight range.

For the **categorical attributes** — gender, ever\_married, work\_type, Residence\_type, and smoking\_status — we visualized their frequency using count plots. The dataset appears reasonably balanced across most variables, though some categories (like “Private” work type) are more frequent. Such patterns help contextualize the population characteristics before building predictive models.

**Potential Patterns**

To explore possible relationships within the data, we examined how certain health and lifestyle factors vary between individuals who experienced a stroke and those who did not. The focus was on **age**, **average glucose level**, and **smoking status**, as these are commonly associated with stroke risk.

The visualizations suggest a **clear upward trend between age and stroke occurrence**, indicating that the likelihood of stroke tends to rise with increasing age. This observation aligns with well-known medical patterns where older adults face higher risks.

We also noticed that **higher glucose levels** appear more frequently among stroke cases, implying that blood sugar regulation could be an influential factor.

For **smoking status**, the relationship was less distinct, though the plot still helps highlight possible lifestyle effects worth exploring further.

These initial visual patterns do not imply causation but serve as valuable hints when deciding which features to prioritize for predictive modeling in later stages.

**3) Data Cleaning**

**(A) Missing Values**

The dataset contained missing values exclusively in the BMI column (approximately 201 records). Since removing these entries would result in unnecessary data loss, we applied mean imputation to fill the missing values. The mean was chosen as it provides a straightforward estimate of the central tendency and helps maintain the overall balance of BMI values in the dataset.

This approach ensures that the distribution of BMI remains consistent with the original data while preserving all patient records for later analysis. After imputation, no missing values remained in the dataset, confirming that it is now complete and ready for subsequent preprocessing steps such as outlier handling, standardization, and encoding.

**(B) Outliers**

Outlier analysis revealed extreme values in the average glucose level and BMI features, which are common in healthcare data due to natural variations among patients. Binary variables such as hypertension and heart\_disease were initially identified by the IQR method but were excluded from treatment since they contain only 0 and 1 values.

To prevent extreme numerical values from exerting excessive influence on the model while preserving clinical relevance, the continuous features were treated using IQR-based capping. Values beyond the upper or lower limits were replaced with the nearest valid threshold.

This method maintains the integrity of the data and ensures a more balanced distribution, providing a solid foundation for later modeling steps.

**(C) Standardization**

To ensure comparability among features and improve model performance, all continuous variables — age, average glucose level, and BMI — were standardized using the Z-score normalization technique. This transformation rescales values to have a mean of 0 and a standard deviation of 1, ensuring that each variable contributes equally to model training.

Standardization was chosen over simple normalization because the distributions of these health-related features are approximately bell-shaped and contain outliers, which are more effectively managed under a standardized scale. This process enhances numerical stability and prepares the dataset for reliable and efficient learning in subsequent modeling stages.

**(D) Encoding**

Categorical variables were kept unencoded since the focus at this stage is on exploratory analysis and data understanding. We chose to retain their original textual form to maintain interpretability during visualizations and statistical exploration.

This approach enables clearer insights into the relationships between health-related factors and stroke occurrence. Encoding (such as label or one-hot encoding) will be applied later — at the end of Milestone 2, once all analysis is complete and the dataset is prepared for machine learning — to ensure both computational efficiency and numerical compatibility with predictive algorithms.