**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.