CSST 104

MACHINE LEARNING IMPLEMENTATION

TOPIC: Healthcare Stroke Analysis

Submitted by:

Mendoza, Ailla Mae S.
Oyardo, Rolley Anne V.

TABLE OF CONTENTS

TABLE OF CONTENTS

GOOGLE COLAB LINK:

https://colab.research.google.com/drive/1BOjLcjRFKqCjepk23S5c7McMpmXVF4nN?usp=sharing

GITHUB LINK:

https://github.com/rolleyanne/Finals-In-CSST-104

Data Analysis and Machine Learning Implementation
Project Documentation

I. Project Overview

This project aims to analyze healthcare data related to stroke patients and build a machine learning model to predict stroke cases using Google Colab.

The purpose of the healthcare stroke analysis is to understand why and how strokes happen, so we can prevent them and help people who have had strokes.

The dataset includes information about people who have had strokes, like their age, gender, whether they have certain health conditions like high blood pressure or heart disease, their lifestyle habits like smoking or exercising, and details about their stroke, like its severity and outcome.

We'll look at things like age, gender, health conditions, and lifestyle habits to see if there are patterns or factors that make someone more likely to have a stroke.

The main goals are to find out what factors increase the risk of stroke, how different factors interact with each other, and what we can do to prevent strokes.

We expect to learn things like which age groups are most at risk, whether men or women are more likely to have strokes, how certain health conditions affect stroke risk, and how lifestyle changes can reduce the chances of having a stroke.

II. Libraries and Data Handling

In the healthcare stroke analysis, we use several libraries for data manipulation and visualization:

1. **Pandas**: used for data manipulation and analysis. It provides data structures and functions to make working with structured data easy and intuitive. With Pandas, we can load, clean, and explore the dataset efficiently.
2. **Matplotlib**: is a plotting library for Python. It provides a wide variety of customizable plots and charts for visualizing data. Matplotlib is versatile and can be used to create simple line plots, scatter plots, histograms, bar charts, and more.
3. **Seaborn:** is built on top of Matplotlib and provides a high-level interface for creating attractive and informative statistical graphics. It simplifies the process of creating complex visualizations by providing functions to visualize relationships in data, such as scatter plots with regression lines or box plots with confidence intervals.

Data Loading and Preprocessing

```
# Load the dataset
df = pd.read_csv('17_Healthcare Stroke Analysis.csv')
```

This function reads the CSV file into a DataFrame, allowing easy manipulation.

```
# Check for missing values
missing_values = df.isnull().sum()
```

missing_values = df.isnull().sum(): This line calculates the number of missing values in each column of the DataFrame df and stores the result in the missing_values variable. The isnull() method checks each element of the DataFrame for missing values (NaNs) and returns a DataFrame of the same shape, where each element is True if it's missing and False otherwise. Then, the sum() method is used to count the number of True values (which represent missing values) for each column.

```
# Handle missing values (impute 'bmi' with mean)
df['bmi'].fillna(df['bmi'].mean(), inplace=True)
```

df['bmi'].fillna(df['bmi'].mean(), inplace=True:) This line specifically *handles missing values* in the 'bmi' column. It fills the missing values in the 'bmi' column with the mean (average) of the non-missing values in that column. The fillna() method is used for this purpose. The inplace=True parameter ensures that the changes are made directly to the original DataFrame df, without the need to assign the result back to df.

III. Data Analysis Techniques

```
Descriptive Statistics:
                 id          age  hypertension  heart_disease
count   5110.000000  5110.000000   5110.000000    5110.000000
mean   36517.829354    43.226614      0.097456       0.054012
std    21161.721625    22.612647      0.296607       0.226063
min       67.000000     0.080000      0.000000       0.000000
25%    17741.250000    25.000000      0.000000       0.000000
50%    36932.000000    45.000000      0.000000       0.000000
75%    54682.000000    61.000000      0.000000       0.000000
max    72940.000000    82.000000      1.000000       1.000000

       avg_glucose_level          bmi       stroke
count        5110.000000  4909.000000  5110.000000
mean          106.147677    28.893237     0.048728
std            45.283560     7.854067     0.215320
min            55.120000    10.300000     0.000000
25%            77.245000    23.500000     0.000000
50%            91.885000    28.100000     0.000000
75%           114.090000    33.100000     0.000000
max           271.740000    97.600000     1.000000
```

Used to summarize and describe the main features of the dataset. This includes measures such as mean, median, mode, range, standard deviation, and variance. Descriptive statistics help in understanding the distribution and characteristics of variables related to strokes, such as age, blood pressure, cholesterol levels, etc.

- **Label Encoding:** It converts categorical variable 'gender' into numerical values using LabelEncoder.
- **MinMax Scaling:** It scales numerical features 'age' and 'avg_glucose_level' to a range between 0 and 1 using MinMaxScaler.

IV. Key Findings

Summarizing the major findings from the healthcare stroke analysis, focusing on user demographics, device usage, and subscription details. And explaining how these findings can influence business decisions or strategies.

1. **User Demographics**:

- The analysis examines user demographics such as gender, age, and marital status, providing insights into the distribution within each category.
- Gender distribution shows the number of males and females in the dataset.
- Marital status ('ever_married') distribution indicates whether individuals are married or not.

2. **Health Conditions and Stroke Risk:**

- The contingency table and chi-square test assess the relationship between hypertension (a health condition) and stroke occurrence.
- Significant evidence or lack thereof regarding the association between hypertension and stroke risk is provided based on the chi-square test results.

3. **Health Metrics and Predictive Modeling:**

- Health metrics such as age, average glucose level, and BMI (body mass index) are analyzed for correlations using a heatmap of the correlation matrix.
- Predictive modeling techniques like logistic regression are applied to predict stroke occurrence based on health metrics and other factors.
- Model performance metrics such as accuracy and classification report (precision, recall, F1-score) are provided to evaluate the model's effectiveness in predicting stroke risk.

4. **Visualization:**

- Gender distribution is visualized using a countplot to illustrate the distribution of males and females in the dataset.

- Stroke occurrence is visualized using a pie chart to show the proportion of individuals with and without strokes.

These findings can influence business decisions and strategies in several ways:

- **Targeted Marketing and Product Development:** Understanding user demographics and health conditions allows businesses to tailor marketing campaigns and develop products/services targeted towards specific segments. For instance, products promoting hypertension management or stroke prevention can be marketed to individuals at higher risk based on demographic and health data.
- **Risk Assessment and Preventive Measures**: Insights from predictive modeling and health metrics analysis can inform risk assessment strategies and preventive measures. Businesses can offer personalized health interventions or subscription services targeting individuals at higher risk of strokes, thereby improving health outcomes and reducing healthcare costs.
- **Partnerships and Collaborations**: Businesses can collaborate with healthcare providers or technology companies to integrate health monitoring features into existing products or develop new solutions addressing identified health needs. Partnerships can also facilitate data sharing and collaboration for better insights and outcomes.
- **Customer Engagement and Education:** Visualizations such as countplots and pie charts can be used to engage customers and educate them about health risks and preventive measures. Businesses can leverage visual storytelling to raise awareness and promote healthy behaviors among their target audience.

V. Advance Analysis

Advanced analytical techniques used, such as geographical insights or temporal trends. Describing how these analyses contribute to understanding broader market dynamics or seasonal patterns.

Geographical Insights:

Geographical analysis involves examining the geographic distribution of stroke occurrences or risk factors. This could be done by plotting stroke incidence rates or risk factors on a map, identifying hotspots or regions with higher prevalence.

Understanding geographical variations in stroke prevalence can inform resource allocation, healthcare planning, and targeted intervention strategies. For example, regions with a higher incidence of strokes may require more healthcare resources and preventive programs.

Temporal Trends:

Temporal analysis involves studying trends over time, such as seasonal patterns or long-term trends in stroke occurrence. This could involve analyzing stroke incidence by month, season, or year, and identifying any recurring patterns or trends.

Identifying seasonal patterns in stroke occurrence can provide insights into environmental factors or lifestyle behaviors that may influence stroke risk. For instance, spikes in stroke incidence during certain seasons may be linked to changes in weather, diet, or physical activity levels.

VI. Machine Learning Implementation

1. **Data Preparation**:

- **Data Collection**: Gather relevant data from sources such as electronic health records, surveys, or clinical studies.
- **Data Integration**: Combine data from multiple sources into a single dataset, ensuring compatibility and consistency.
- **Data Exploration**: Explore the dataset to understand its structure, features, and distributions. Identify potential relationships and patterns.

2. **Data Selection**:

- **Feature Selection**: Choose relevant features (variables) that are likely to influence stroke risk or outcomes. This may include demographic information, health conditions, lifestyle factors, etc.
- **Target Variable Selection**: Determine the target variable (dependent variable) that the model will predict. In this case, it could be stroke occurrence (binary classification).

3. **Data Cleaning**:

- **Handling Missing Values**: Address missing data by imputing values (e.g., using mean or median) or removing incomplete records.
- **Dealing with Outliers**: Identify and handle outliers that may affect model performance or distort analysis results.
- **Encoding Categorical Variables**: Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding.

4. **Feature Scaling**:

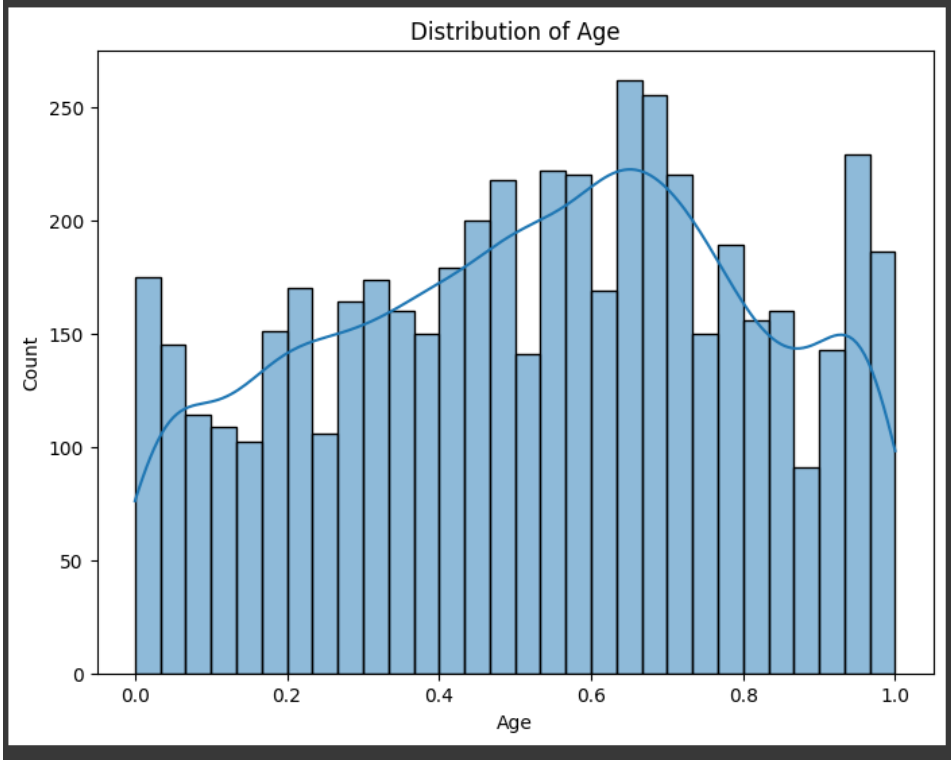
- **Normalization/Standardization**: Scale numerical features to a similar range to prevent features with larger magnitudes from dominating the model. Common techniques include normalization (scaling to a range of 0 to 1) or standardization (scaling to have a mean of 0 and a standard deviation of 1).

5. **Model Building**:

- **Splitting the Data**: Divide the dataset into training and testing sets to evaluate model performance. The training set is used to train the model, while the testing set is used to assess its generalization to unseen data.
- **Selecting a Model**: Choose an appropriate machine learning algorithm based on the nature of the problem (e.g., classification for predicting stroke occurrence) and the characteristics of the data (e.g., linear models for simple relationships, tree-based models for complex interactions).
- **Training the Model**: Fit the selected model to the training data, allowing it to learn patterns and relationships between features and the target variable.
- **Evaluating the Model**: Assess the model's performance using evaluation metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC). Compare performance on the testing set to ensure the model generalizes well to new data.

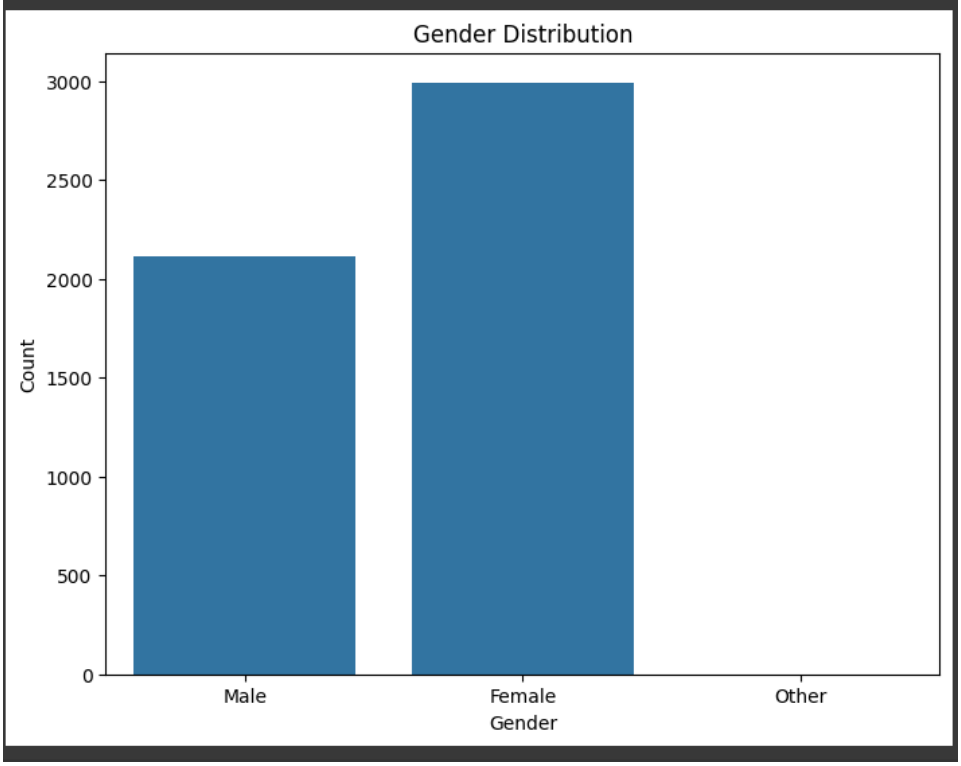
VII. Visual Insights

```
# Example: Plotting a histogram of age
plt.figure(figsize=(8, 6))
sns.histplot(df['age'], bins=30, kde=True)
plt.title('Distribution of Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```



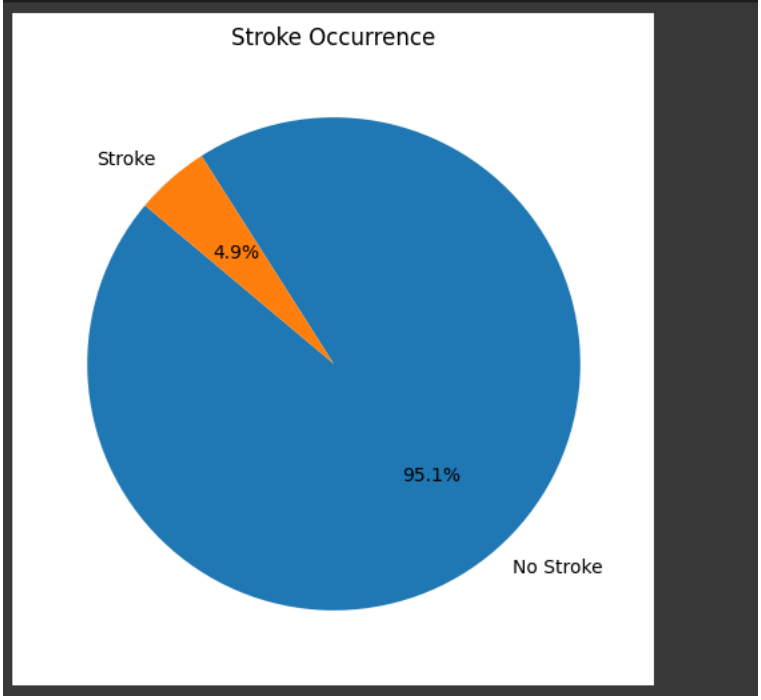
The **histogram** visualizes the distribution of ages in the dataset. Each bar represents the count of individuals falling within a specific age range. The kernel density estimate curve provides additional information about the density of age values across the entire range. This visualization helps in understanding the age distribution of the population under study, which is valuable for identifying age-related patterns or trends in healthcare outcomes, such as stroke incidence or risk factors.

```python
# Countplot for gender distribution
plt.figure(figsize=(8, 6))
sns.countplot(x='gender', data=df)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



The **countplot** visualizes the distribution of gender in the dataset. Each bar represents the count of individuals belonging to a specific gender category. The plot provides a quick overview of the gender composition of the population under study, enabling insights into potential gender disparities in healthcare outcomes or access to services. In this case, it helps in understanding the relative frequency of males and females in the dataset, which may be relevant for analyzing gender-specific health issues or interventions.

```python
# Pie chart for stroke occurrence
stroke_counts = df['stroke'].value_counts()
labels = ['No Stroke', 'Stroke']
plt.figure(figsize=(6, 6))
plt.pie(stroke_counts, labels=labels, autopct='%1.1f%%', startangle=140)
plt.title('Stroke Occurrence')
plt.show()
```

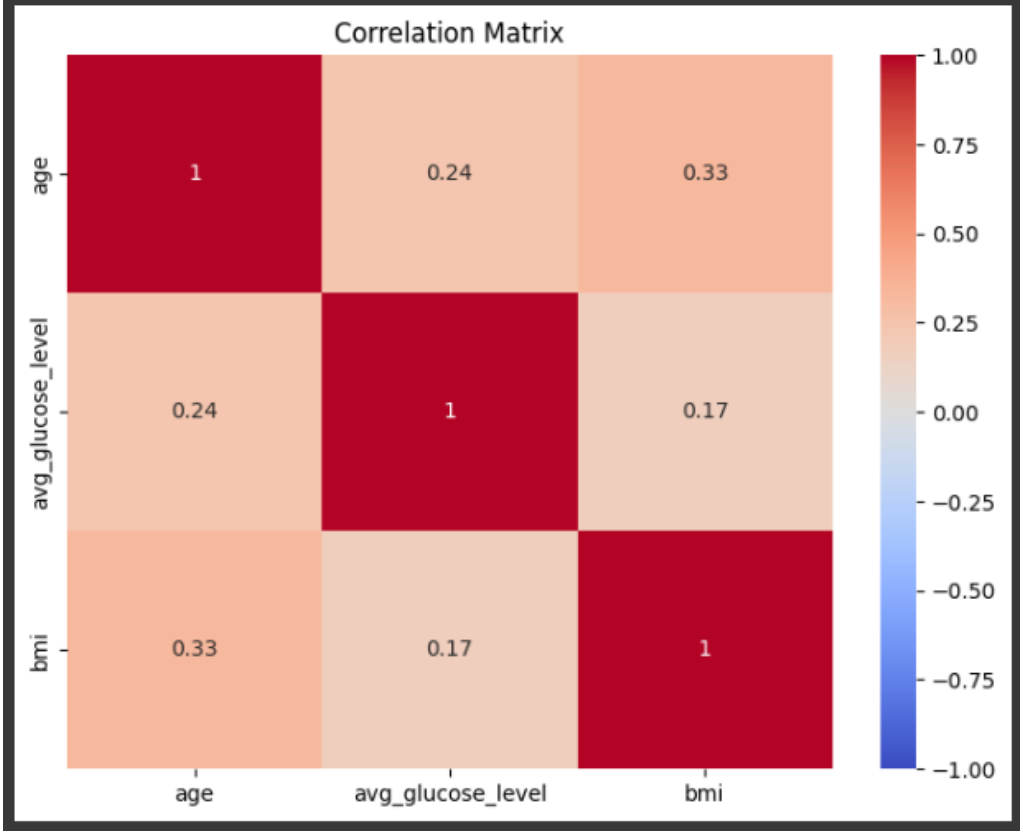


The **pie chart** visualizes the distribution of stroke occurrence in the dataset. Each slice represents the proportion of individuals with or without strokes. The chart provides a clear representation of the relative frequency of stroke cases compared to non-stroke cases, allowing for easy interpretation of the dataset's stroke prevalence. In this case, it helps in understanding the overall prevalence of strokes within the population under study, which is valuable for assessing the magnitude of the health issue and informing preventive measures or healthcare interventions.

```
numerical_cols = ['age', 'avg_glucose_level', 'bmi']

corr_matrix = df[numerical_cols].corr()

plt.figure(figsize=(8, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Matrix')
plt.show()
```



The **heatmap** visualizes the correlation matrix between the selected numerical variables. Each cell in the heatmap represents the correlation coefficient between two variables. The color intensity indicates the strength and direction of the correlation: darker colors (blue for negative, red for positive) represent stronger correlations, while lighter colors indicate weaker or no correlations. This visualization helps in identifying relationships and patterns between variables, such as whether age, glucose level, and BMI are positively or negatively correlated with each other. It's useful for understanding the interplay between different health metrics and their potential impact on stroke risk or outcomes.

VIII. Conclusion

The insights derived from the analysis of healthcare stroke data hold significant implications for businesses, organizations, and healthcare providers. Here's an overview of how these insights can impact decision-making and the potential for future analysis:

Informed Decision-Making: The insights gained from analyzing user demographics, health metrics, and stroke occurrence can inform strategic decision-making for businesses and healthcare organizations. For example, understanding demographic trends and health risk factors can guide the development of targeted interventions, marketing campaigns, and healthcare services tailored to specific population segments.

Improved Healthcare Outcomes: By leveraging data-driven insights, healthcare providers can optimize patient care, enhance preventive measures, and improve health outcomes. For instance, identifying high-risk groups for strokes based on demographic or health data allows for proactive screening, early intervention, and personalized treatment plans, ultimately reducing morbidity and mortality rates associated with strokes.

Resource Allocation and Planning: Data analysis enables better resource allocation and planning in healthcare settings. For example, geographical insights can identify regions with higher stroke prevalence, guiding the allocation of healthcare resources, infrastructure development, and public health initiatives to address regional disparities and improve access to care.

Risk Prediction and Management: Machine learning models trained on healthcare data can predict stroke risk more accurately, facilitating early detection and intervention. These models can integrate diverse data sources, including demographic, clinical, and lifestyle factors, to generate personalized risk assessments and preventive strategies for individuals at high risk of strokes.

Continuous Improvement and Future Analysis: Data-driven decision-making fosters a culture of continuous improvement and innovation in healthcare delivery. By

regularly analyzing new data, monitoring outcomes, and evaluating the effectiveness of interventions, organizations can adapt strategies, refine processes, and identify emerging trends to stay ahead of evolving healthcare challenges.

The importance of data-driven decision-making in healthcare cannot be overstated. Insights derived from comprehensive data analysis enable businesses, organizations, and healthcare providers to tailor services, allocate resources efficiently, and implement targeted interventions that have a meaningful impact on patient outcomes and population health. As data collection methods, analytical techniques, and computing capabilities continue to evolve, the potential for future analysis and innovation in healthcare remains vast, promising continued advancements in preventive care, treatment modalities, and population health management.

Appendix

Additional information, such as data sources, contributor details, or acknowledgments.

Data Sources: Provide details about the sources of the healthcare stroke data used in the analysis, such as electronic health records, clinical databases, or population surveys. Mention any data collection methods, data providers, or organizations involved in collecting and sharing the data.

Contributor Details: Acknowledge individuals or teams involved in the data analysis process, including data scientists, analysts, healthcare professionals, and domain experts. Mention their roles, contributions, and expertise that contributed to the insights derived from the analysis.

Acknowledgments: Express gratitude to any organizations, institutions, or funding agencies that supported the data analysis project. Acknowledge their contributions, resources, or funding that made the analysis possible.

Including this additional information adds transparency, credibility, and context to the analysis, helping stakeholders understand the data's origin, the expertise involved in the analysis, and the support received during the project. It also fosters collaboration, recognition, and accountability within the healthcare community.