MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

SCHOOL OF E&TC ENGINEERING

Brain Stroke Prediction

Soft Computing Project

(ACADEMIC YEAR: 2023-2024)

SEMESTER VI

Group No: 15

Name of Project Guide: Dr. Smita Kulkarni

Student Details:

PRN	Student Name	Branch	Email
202202060016	Shreyash Mahakale	ETX	shreyash.mahakale@mitaoe.ac.in
202202060008	Jyoti Sarwade	ETX	jyoti.sarwade@mitaoe.ac.in

ABSTRACT

Stroke is a leading cause of death and disability worldwide. Early detection and risk assessment are crucial for preventing strokes and improving patient outcomes. This project utilizes machine learning techniques to develop a model that predicts the likelihood of stroke occurrence based on patient data..

1. Introduction

1.1 Project Background

Stroke is a medical emergency that occurs when blood flow to a part of the brain is interrupted. This can cause brain cells to die, leading to permanent damage and a range of debilitating symptoms.

1.2 Motivation

Despite advancements in medical care, stroke remains a significant health concern. Developing a reliable stroke prediction model can empower healthcare professionals to identify high-risk patients and implement preventative measures.

1.3 Problem Statement

The challenge lies in creating a machine learning model that accurately predicts the possibility of stroke using a dataset containing patient information like age, gender, and health factors.

1.4 Objective

This project aims to develop and evaluate a machine learning model capable of predicting stroke occurrence based on a dataset of patient characteristics**.**

1.5 About the Dataset

The project will utilize a dataset containing relevant patient information for stroke prediction. This data may include:

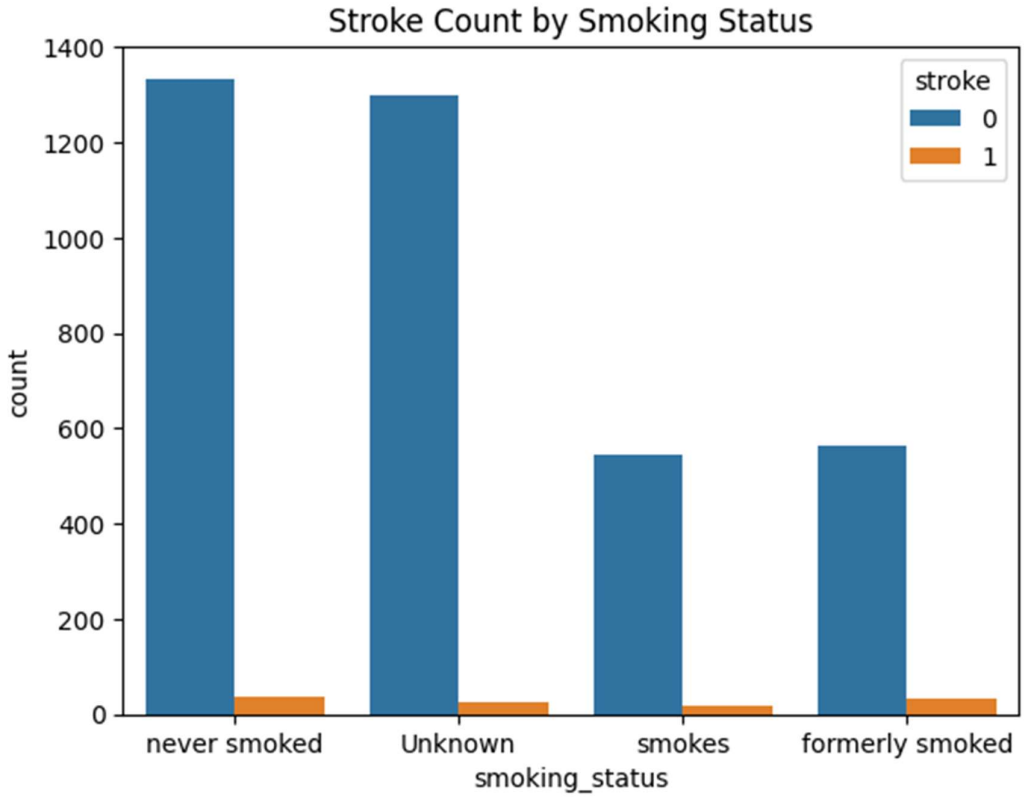
- Demographic data: Age, gender, region, state
- Health factors: Body Mass Index (BMI), hypertension (high blood pressure), history of heart disease, smoking status, blood glucose levels

2. Data Visualization

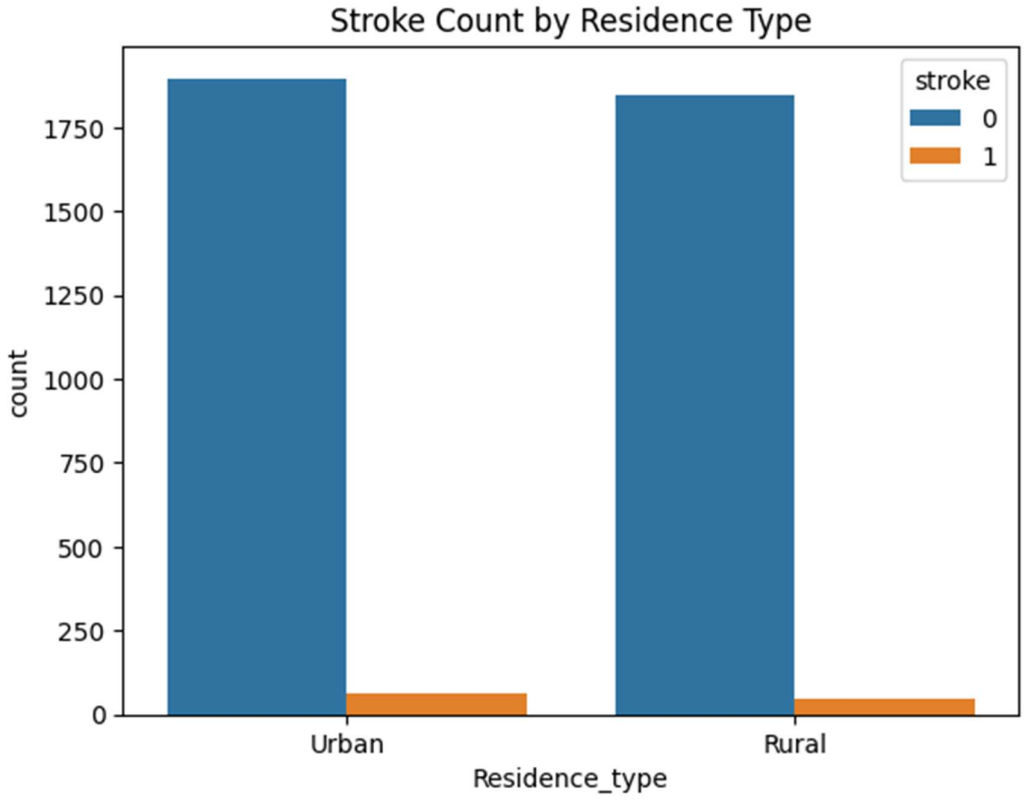
A deeper comprehension of the data and its implications for sleep and general health is made possible by the use of data visualization techniques, which visually compellingly and informatively convey the relationships and insights generated from the sleep health and lifestyle dataset.

Plots used:

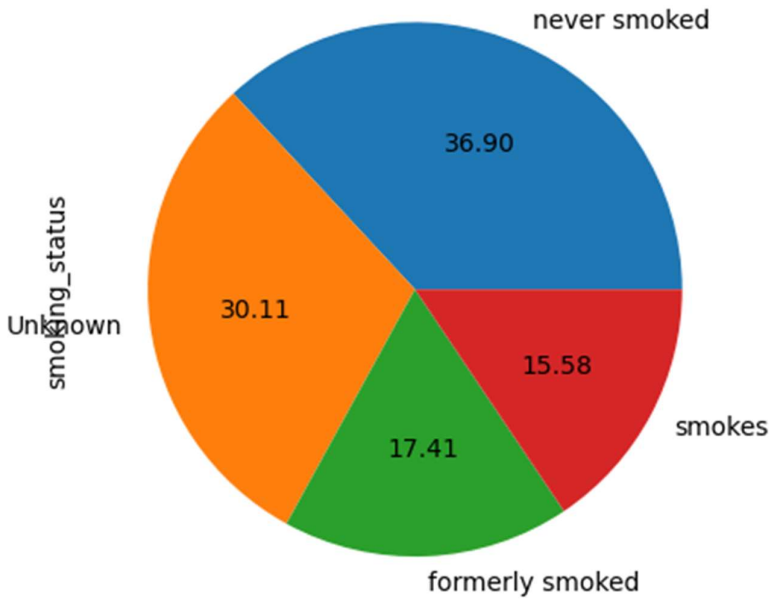
1) Bar Plot: - A bar plot is used to display the distribution of a categorical variable. Each bar represents a category, and the height of the bar represents the corresponding value of that category.



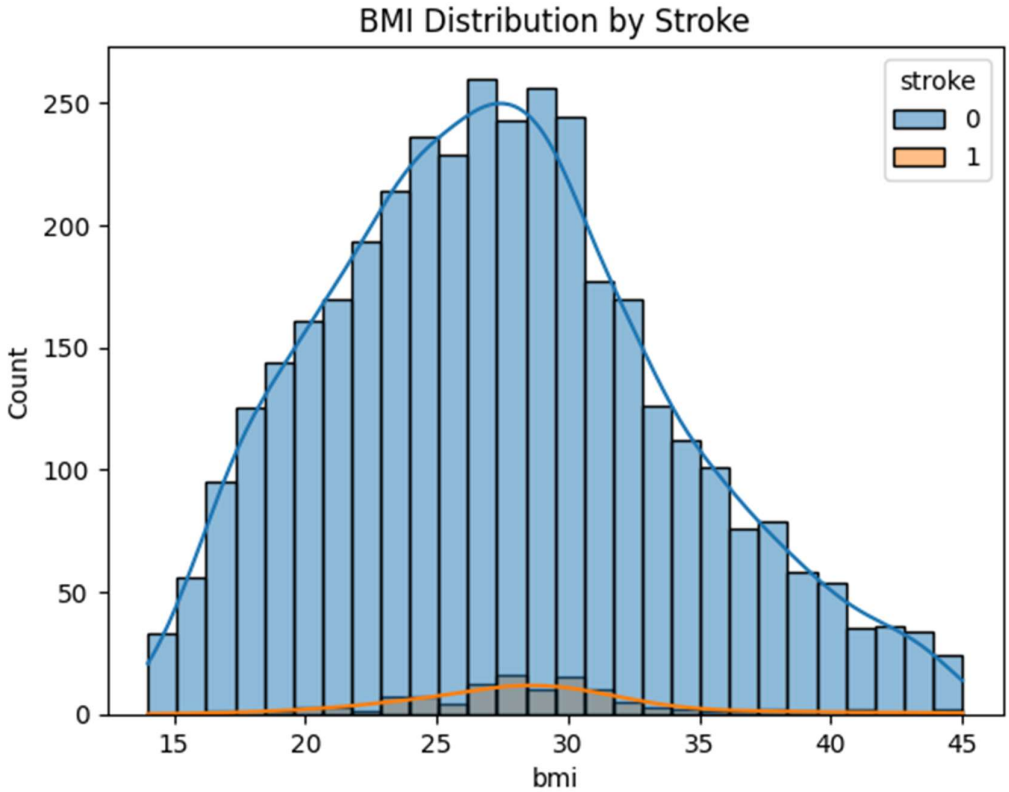
2) Count plot: - A count plot is a simple yet effective visualization tool that displays the frequency of occurrences of different categories within a categorical variable. It helps in understanding the distribution of categorical data and identifying any prominent trends or patterns present within the dataset



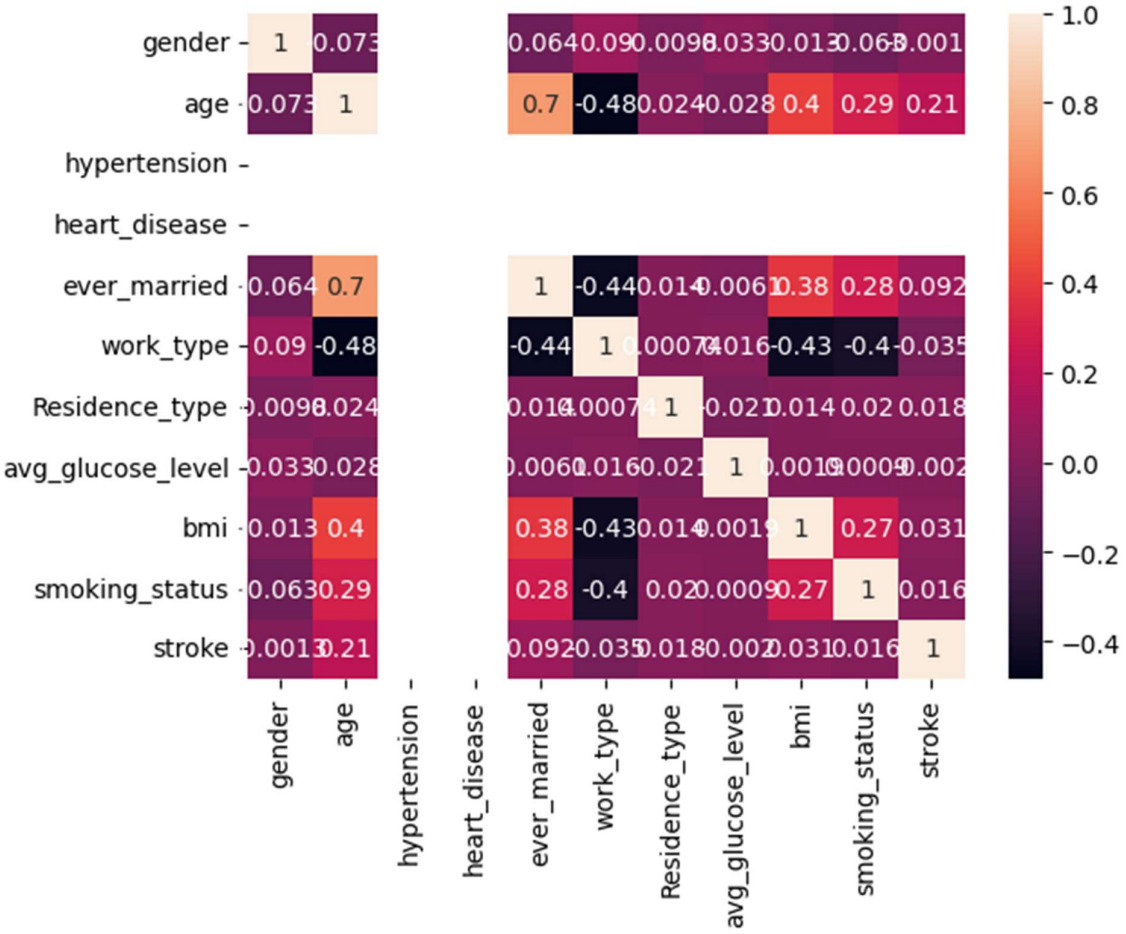
3) Pie Chart: -A pie chart is used to display the proportion of each category in a whole. Each slice of the pie represents a category, and the size of the slice is proportional to the percentage of that category's occurrence in the data



4) Histogram: -A histogram is used to represent the distribution of a continuous numerical variable. It divides the range of the data into intervals (bins) and shows the frequency of data points falling into each bin.



5) Heatmap Plot: - A heatmap plot visualizes data in a matrix format, where values are represented by colors. It's particularly useful for identifying patterns, correlations, or clusters within large datasets, providing a clear and concise overview of relationships between variables



Libraries Used:

Pandas: Used for data manipulation, cleaning, and analysis. It provides data structures for efficiently working with large datasets.

NumPy: Often used alongside Pandas for numerical operations and array manipulation.

Matplotlib: A popular library for creating static, animated, and interactive visualizations in Python.

Seaborn: Built on top of Matplotlib, Seaborn is used for creating more attractive and informative statistical graphics

Sklearn preprocessing: sklearn preprocessing is a module in scikit-learn that offers tools for preparing data for machine learning models, including scaling, normalization, encoding categorical variables, and handling missing values.

sklearn.model_selection: sklearn.model_selection in scikit-learn provides functions for data splitting, cross-validation, and parameter tuning essential for evaluating and optimizing machine learning models.

category_encoders: category_encoders is a Python library for converting categorical variables into numerical representations, useful for preprocessing data before applying machine learning algorithms.

sklearn.svm: sklearn.svm is a module in scikit-learn for support vector machine (SVM) algorithms used in classification and regression tasks.

Conclusion

This project leveraged machine learning to develop a stroke prediction model using a dataset of patient characteristics. While further validation and potential model refinement are necessary, this initial exploration demonstrates the promise of this approach for early stroke risk assessment, potentially empowering healthcare professionals to prioritize preventative measures and improve patient outcomes.