Capstone Two - Project Proposal - Hector Sanchez

Proposal

Predicting Stroke Risk Using Patient Health Data:

- The primary goal of this project is to develop a predictive model to determine the likelihood of a patient experiencing a stroke based on various health and demographic factors. This model will aid in early identification of high-risk individuals, enabling healthcare professionals to take preventive measures and improve patient outcomes.

Introduction:

Stokes, which are often referred to as brain attacks, are a medical condition caused by something blocking blood supply to part of the brain, or when a blood vessel bursts in the brain. According to the World Health Organization (WHO), stroke is the second leading cause of death globally, accounting for approximately 11% of all deaths. Strokes can have severe, long-term consequences on a patient's quality of life, making early identification of risk crucial. The dataset provided includes key features such as age, gender, presence of hypertension, heart disease, glucose levels, BMI, and smoking status, all of which have been shown to influence stroke risk.

Objectives:

- **Develop a predictive model** to accurately estimate a patient's likelihood of having a stroke based on health and demographic data.
- **Perform exploratory data analysis (EDA)** to uncover relationships between variables like age, hypertension, heart disease, and stroke occurrence.
- **Handle missing and imbalanced data** effectively, ensuring the model is robust and generalizable across different patient groups.
- **Test and compare multiple machine learning models**, such as Logistic Regression, Random Forest, and Gradient Boosting, to identify the most effective approach for stroke prediction.
- **Tune hyperparameters** to optimize model performance and reduce overfitting.
- **Evaluate the model's performance** using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, ensuring its effectiveness in predicting stroke risk.
- **Identify key risk factors** associated with stroke, providing insights for healthcare professionals on which patient characteristics contribute most to stroke likelihood.

- **Create a user-friendly tool** or dashboard that can visualize stroke risk and help healthcare practitioners make informed decisions about preventive interventions.
- **Present findings** in a comprehensive report that outlines the model's performance, data insights, and potential healthcare implications for early stroke detection.

Dataset Overview:

- **Source**: Stroke Prediction Dataset (confidential source for educational purposes)
- **Key Attributes**:
    1. **ID**: Unique patient identifier
    2. **Gender**: Male, Female, or Other
    3. **Age**: Patient's age
    4. **Hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
    5. **Heart Disease**: 0 if no heart disease, 1 if the patient has heart disease
    6. **Marital Status**: Yes or No
    7. **Work Type**: Children, Govt job, Never worked, Private, Self-employed
    8. **Residence Type**: Rural or Urban
    9. **Average Glucose Level**: Blood glucose level
    10. **BMI**: Body Mass Index
    11. **Smoking Status**: Formerly smoked, Never smoked, Smokes, or Unknown
    12. **Stroke**: Target variable (1 = stroke, 0 = no stroke)

Methods/Approach:

1. **Data Cleaning and Preprocessing**: Handle missing data (e.g., "Unknown" values in smoking status), standardize formats, and address any data imbalances using techniques such as SMOTE (Synthetic Minority Over-sampling Technique).
2. **Exploratory Data Analysis (EDA)**: Perform EDA to gain insights into the distribution of data and correlations between variables, using visualizations like histograms, boxplots, and heatmaps.
3. **Feature Engineering**: Consider additional transformations (e.g., binning continuous variables like age and glucose level) to enhance model performance.

4. **Model Development**:
    - Apply machine learning models such as Logistic Regression, Random Forest, and Gradient Boosting to predict stroke risk.
    - Tune hyperparameters to optimize performance.
    - Evaluate models using performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC curve.

Model Evaluation and Selection: Identify the best-performing model based on validation set performance and interpret its results to understand which factors contribute most to stroke prediction.

Goals/Expected Outcomes of this Project:

1. A trained predictive model capable of classifying patients based on stroke risk.
2. Key insights into factors that have the most significant influence on stroke likelihood.
3. A report summarizing the model's performance and potential implications for clinical practice.

Tools and Technologies:

- **Programming Language**: Python
- **Libraries**: pandas, numpy, scikit-learn, matplotlib, seaborn
- **Modeling Techniques**: Logistic Regression, Random Forest, Gradient Boosting
- **Evaluation Metrics**: Accuracy, Precision, Recall, F1-score, AUC-ROC

Conclusion:

This project aims to provide healthcare practitioners with an efficient tool to identify high-risk stroke patients based on various health indicators, potentially contributing to better prevention strategies and improved public health outcomes.