Heart Stroke Prediction

This project aims to predict the likelihood of a heart stroke based on various health and demographic factors using machine learning models. The dataset used in this project contains health-related information, and the workflow includes data preprocessing, exploratory data analysis (EDA), feature engineering, model training, and evaluation.

Table of Contents

Project Overview

The goal of this project is to predict whether an individual is at risk of a heart stroke based on features such as age, BMI, smoking status, and other health metrics. We applied multiple machine learning models and evaluated their performance using accuracy, classification reports, and ROC curves.

Dataset

The dataset used in this project is named `healthcare-dataset-stroke-data.csv`, which includes the following features:

- `gender`
- `age`
- `hypertension`
- `heart_disease`
- `ever_married`
- `work_type`
- `Residence_type`
- `avg_glucose_level`
- `bmi`
- `smoking_status`
- `stroke` (Target Variable: 1 for stroke, 0 for no stroke)

Data Preprocessing

1. **Missing Value Treatment:**
   - Missing values in the `bmi` column were filled with the mean value of the column.
2. **Label Encoding:**
   - Categorical variables (`gender`, `ever_married`, `work_type`, `Residence_type`, `smoking_status`, `age_group`) were converted into numerical format using `LabelEncoder`.
3. **Feature Scaling:**
   - Standardized `avg_glucose_level` and `bmi` using `StandardScaler`.
4. **Outlier Treatment:**
   - Applied winsorization to limit extreme values in `avg_glucose_level` and `bmi`.

Exploratory Data Analysis (EDA)

- **Correlation Heatmap:** A heatmap was generated to understand the relationships between features.
- **Boxplot for BMI and Glucose Levels:** Identified outliers and visualized distributions.

Feature Engineering

1. **Age Group:**
   - Categorized age into groups: `Child`, `Adult`, `Senior`.
2. **Comorbidity:**
   - Combined `hypertension` and `heart_disease` into a single feature.

Model Training and Evaluation

We implemented the following machine learning models:

1. **K-Nearest Neighbors (KNN):**
   - KNN was trained with `n_neighbors=3`.
   - Evaluation: Accuracy, classification report.
2. **Decision Tree:**
   - Trained with `random_state=42`.
   - Evaluation: Accuracy, classification report.
3. **Naive Bayes (GaussianNB):**
   - A probabilistic model based on Bayes' theorem.
   - Evaluation: Accuracy, classification report.
4. **XGBoost:**
   - Configured with `eval_metric='logloss'`.
   - Evaluation: Accuracy, classification report, and ROC curve.
5. **Logistic Regression:**
   - Used for baseline comparison.

Hyperparameter Tuning

- **XGBoost:**
  - Performed GridSearchCV to optimize hyperparameters, including:
    - `n_estimators`, `learning_rate`, `max_depth`, `min_child_weight`, `subsample`, `gamma`, and `reg_alpha`.
  - Achieved improved performance with optimal parameters.

Performance Metrics

For each model, the following metrics were evaluated:

- **Accuracy:** Proportion of correctly classified samples.
- **Classification Report:** Precision, recall, F1-score.
- **ROC Curve:** Evaluated model discrimination capabilities using AUC-ROC.
- **Confusion Matrix:** Visualized true positives, true negatives, false positives, and false negatives.

Visualization

1. **Correlation Heatmap:** Visualized feature relationships.
2. **ROC Curve:** Compared true positive rates vs. false positive rates for models.
3. **Boxplots:** Highlighted outliers in numerical features like BMI and glucose levels.
4. **Confusion Matrices:** Plotted for each model to analyze misclassifications.

How to Run

1. Clone the repository:

   ```
   git clone https://github.com/your-username/heart-stroke-prediction.git
   ```

2. Install dependencies:

   ```
   pip install -r requirements.txt
   ```

3. Place the dataset (`healthcare-dataset-stroke-data.csv`) in the root directory.
4. Run the notebook:

   ```
   jupyter notebook heart_stroke_prediction.ipynb
   ```

Conclusion

This project demonstrates the application of various machine learning models and preprocessing techniques to predict heart stroke risk. It also highlights the importance of EDA, feature engineering, and hyperparameter tuning in improving model performance.

Feel free to explore the code and experiment with different parameters to improve predictions further. If you find this project helpful, give it a ⭐ on GitHub!