STROKE PREDICTION WITH MACHINE LEARNING : CLASSIFICATION MODELS.

CONTENTS

INTRODUCTION

This project focuses on predicting the likelihood of stroke based on patient health data. Stroke is a major health concern with significant morbidity and mortality worldwide. Early identification of high-risk individuals can enable timely interventions and improve outcomes. The goal of this analysis is to build and evaluate classification models that can accurately predict stroke occurrence using demographic and medical features.

BUSINESS PROBLEM

Stroke is one of the leading causes of death and long-term disability, and timely prevention is often the key to reducing its impact. The challenge for many healthcare providers lies in identifying high-risk patients before symptoms appear.

Rather than relying solely on reactive care, there's a need for a data-driven approach that can flag individuals who might be more susceptible to strokes. By using machine learning to analyze patient data, this project aims to fill that gap — offering an efficient way to detect potential stroke cases early and enable preventative action.

Stakeholders: hospital systems that manage large numbers of patients, public health agencies

OBJECTIVES

- To develop a predictive model that can accurately classify whether a person is at risk of having a stroke based on medical and demographic data.
- To handle class imbalance in the dataset using appropriate resampling techniques and evaluation metrics that reflect the performance on the minority (stroke) class.
- To compare the performance of different classification algorithms (Logistic Regression, Random Forest, XGBoost) using metrics such as recall, precision, F1-score, accuracy, and ROC AUC.
- To select and fine-tune the best model through hyperparameter tuning in order to improve its ability to identify stroke cases while minimizing false negatives.

DATA DESCRIPTION

The dataset used in this project is Stroke Prediction Dataset which contains medical records of patients and their associated risk factors for stroke. It includes the following features:

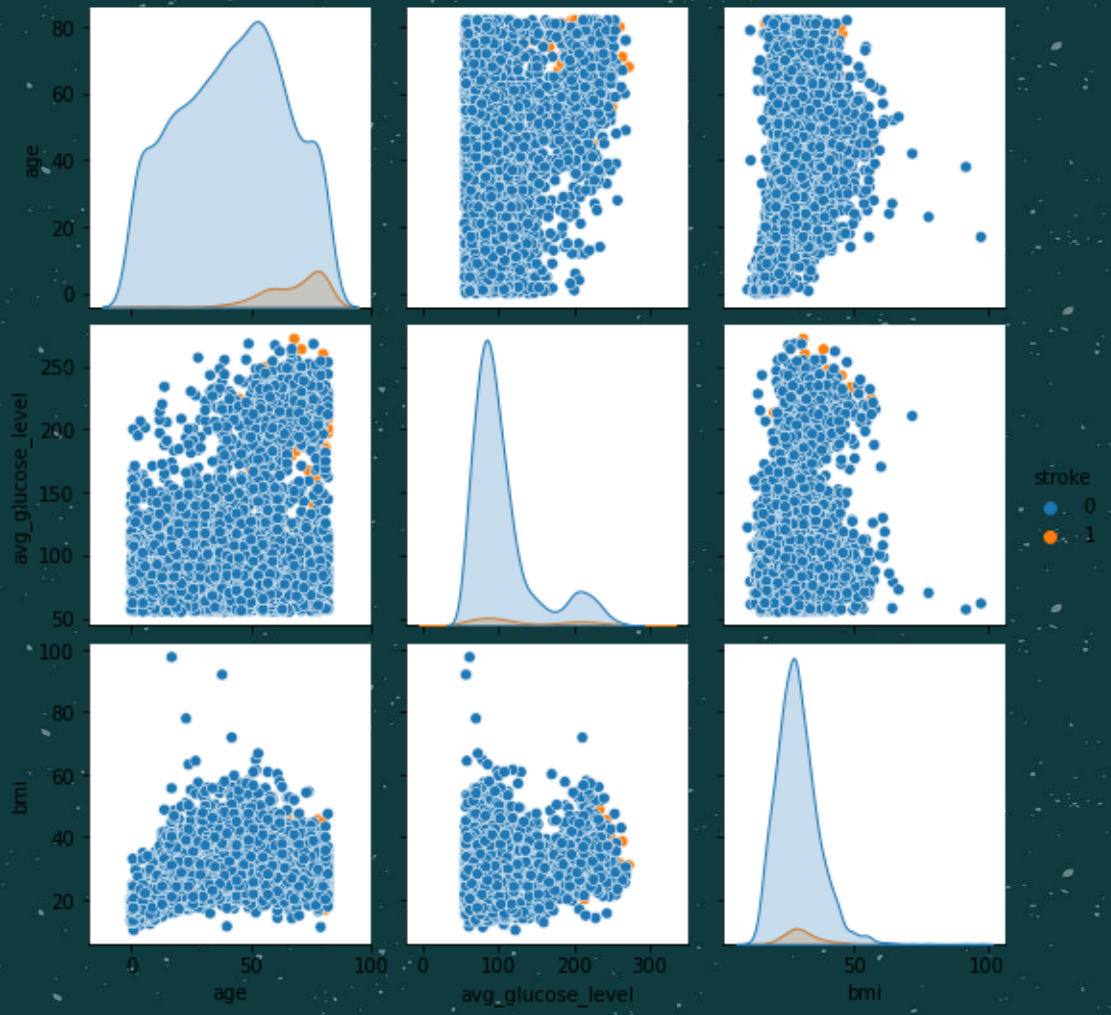
- Id: Unique identification number for each patient.
- Age: The age of the patient.
- Hypertension: Whether the patient has hypertension (1 for yes, 0 for no).
- Heart Disease: Whether the patient has a history of heart disease (1 for yes, 0 for no).
- Ever Married: Whether the patient has been married (1 for yes, 0 for no).
- Work Type: The type of work the patient does (e.g., private, self-employed, government, children).
- Residence Type: Whether the patient resides in an urban or rural area.
- Glucose Level: The patient's glucose level.
- BMI (Body Mass Index): A measure of body fat based on height and weight.
- Smoking Status: The smoking habit of the patient (e.g., never smoked, formerly smoked, currently smoking).
- Stroke: The target variable indicating whether the patient had a stroke (1 for yes, 0 for no).

TOP CORRELATED FEATURES

```
Top features correlated with stroke:

age                  0.245257
heart_disease        0.134914
avg_glucose_level    0.131945
hypertension         0.127904
bmi                  0.036110
Name: stroke, dtype: float64
```

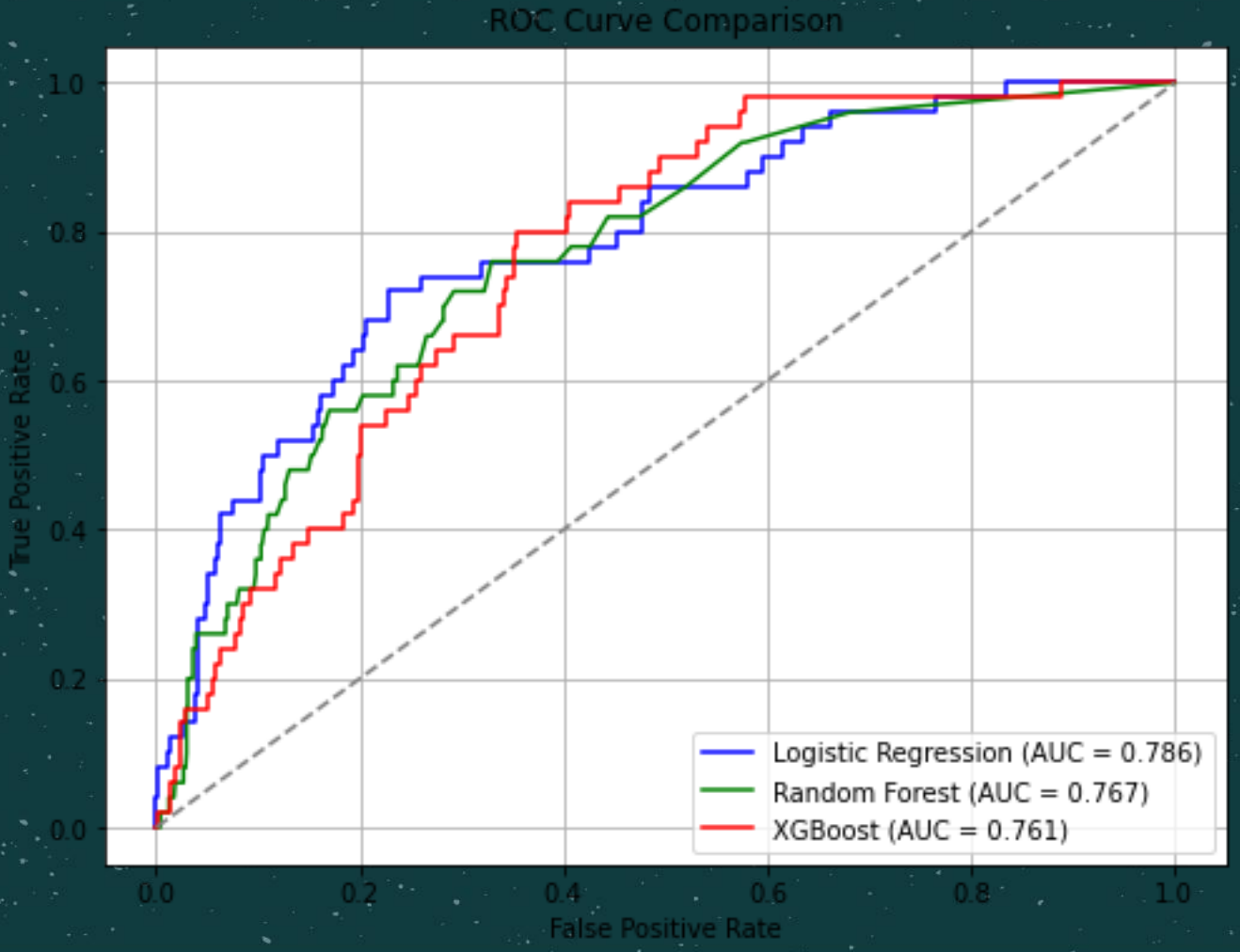
A PAIRPLOT SHOWING THE RELATIONSHIP BETWEEN THE TOP FEATURES



MODELLING

My main goal here was to create a model that can correctly identify as many stroke cases as possible. I started by trying out different classification models and used resampling techniques to deal with the imbalance. I also did hyperparameter tuning to improve model performance and used evaluation metrics like recall, precision, F1-score, and ROC AUC to make a fair comparison between models.

ROC CURVE OF THE MODELS THAT WERE CREATED AND TUNED



RESULTS OF THE BEST PERFORMING MODEL

TUNED LOGISTIC REGRESSION MODEL

	precision	recall	f1-score	support
0	0.98	0.78	0.87	972
1	0.14	0.68	0.23	50
accuracy			0.78	1022
macro avg	0.56	0.73	0.55	1022
weighted avg	0.94	0.78	0.84	1022

ROC AUC Score: 0.7860905349794238
Confusion Matrix:
[[761 211]
[16 34]]

EVALUATION

Logistic Regression is the best option for this stroke prediction task. The main reason is that it gave the highest recall for the stroke class, which is super important because we want to catch as many actual stroke cases as possible. Its precision was low, meaning there are a lot of false positives.

Random Forest and XGBoost had good overall accuracy, but they didn't do as well in detecting the minority class, stroke cases. Their recall was much lower, which means they missed more actual strokes. Since the whole point is to identify those at risk, I had to prioritize recall over accuracy or precision.

CONCLUSION

Tested multiple models including Logistic Regression, Random Forest, and XGBoost. After comparing their results, I found that Logistic Regression performed the best in terms of recall, which is very important for this kind of health-related prediction task. Although it had low precision, I preferred it over the others because it was able to catch more actual stroke cases.

Applied hyperparameter tuning to improve each model, and the ROC curve comparison confirmed that Logistic Regression had the best balance between detecting strokes and minimizing false negatives.

Overall, this project shows how important it is to look beyond accuracy, especially when dealing with imbalanced medical datasets, and focus on the metrics that really matter for the problem at hand