JemmyKuria / **Phase_3_project**

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

0 stars **0** forks **0** watching Branches Activity Tags

Public repository

1 Branch 0 Tags Go to file t Go to file + Add file Code ...

JemmyKuria Made some changes in both the index and readme aed82ea · 2 minutes ago

.ipynb_checkpoints	Made some changes in both the ind...	2 minutes ago
AgeVsStroke.png	Added an image on the Age distribu...	19 hours ago
Bestmodel.PNG	Added an image of the LOgistic regr...	1 hour ago
ClassDistribution.png	Added a Class Distribution image	19 hours ago
CorrelatedFeatures.PNG	Added an image of top correlated fe...	19 hours ago
FeatureImportance.png	Added the Feature Importance image	yesterday
OccupationVsStroke.png	Added an image of Occupation vs st...	1 hour ago
README.md	Made some changes in both the ind...	2 minutes ago
ROC.png	Made some modifications on the co...	1 hour ago
SmokingStatus.png	Added an image of Smoking status ...	1 hour ago
healthcare-dataset-stroke-dat...	Added the dataset	yesterday
index.ipynb	Made some changes in both the ind...	2 minutes ago
pairplot.png	Added an image of the pairplot	1 hour ago
presentation.pdf	Made some changes in both the ind...	2 minutes ago

README

Stroke Prediction with Machine Learning: Classification Models

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.

Stakeholder

This solution is primarily intended for hospital systems that manage large numbers of patients and aim to optimize preventive care. However, the benefits also extend to other stakeholders: public health agencies may apply these insights at a broader scale to improve community health planning. Each of these groups has a vested interest in improving early detection strategies.

Business Impact

A well-performing prediction model offers real value on multiple levels. For patients, it can mean earlier treatment, fewer complications, and a better chance at recovery. For healthcare systems, it translates to more efficient use of resources, fewer emergency interventions, and reduced long-term care costs. Over time, widespread use of predictive tools like this could lead to a noticeable drop in stroke rates, easing the burden on public health infrastructure while saving lives.

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

Data : https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

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).
```

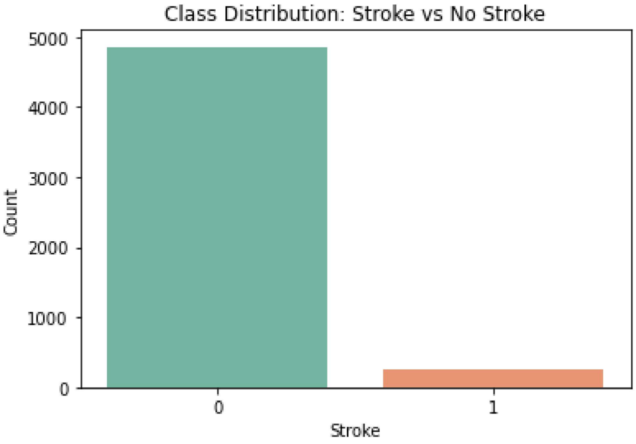
The target variable indicates whether a patient has had a stroke. The data includes both categorical and numerical variables, requiring preprocessing before modeling. The dataset is not balanced between stroke and non-stroke cases, this causes the use of SMOTE which will enable the class balance.

EXPLORATORY DATA ANALYSIS

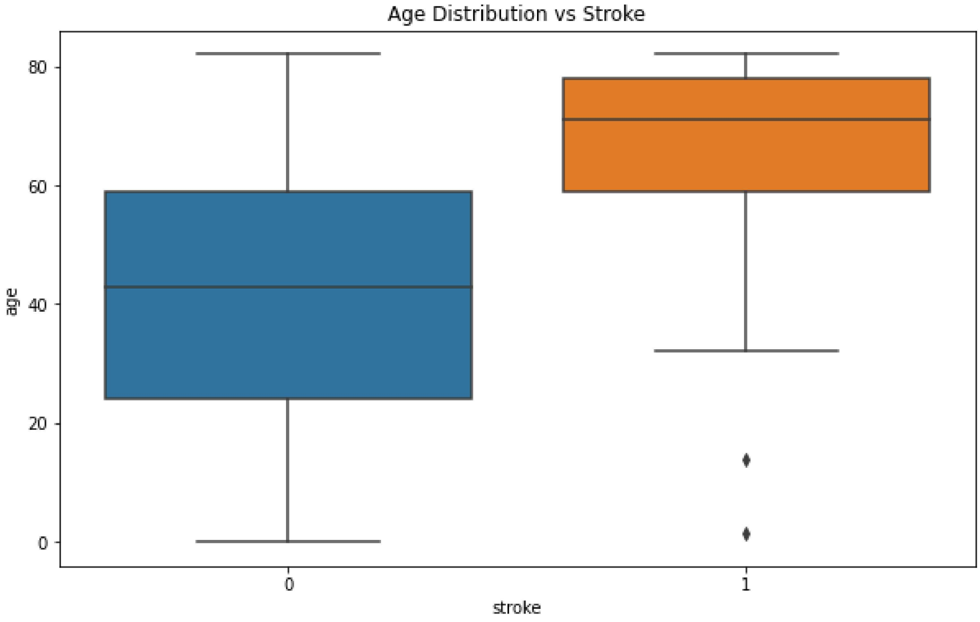
Exploratory analysis was conducted to understand feature distributions and relationships with the target variable. Visualizations such as histograms, box plots, and correlation matrices helped identify patterns and potential predictors of stroke. Initial observations revealed that factors like age, hypertension, and heart disease are strongly associated with stroke occurrence. This insight guided feature selection and engineering in the modeling phase.

Some of the insights noted were:

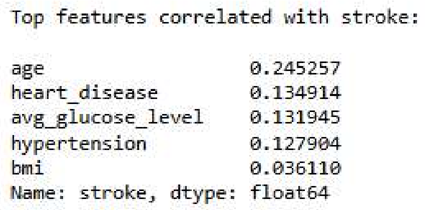
1. The classes were highly imbalanced which called for the SMOTE process.



2. Age feature has a high relationship with the target "Stroke"



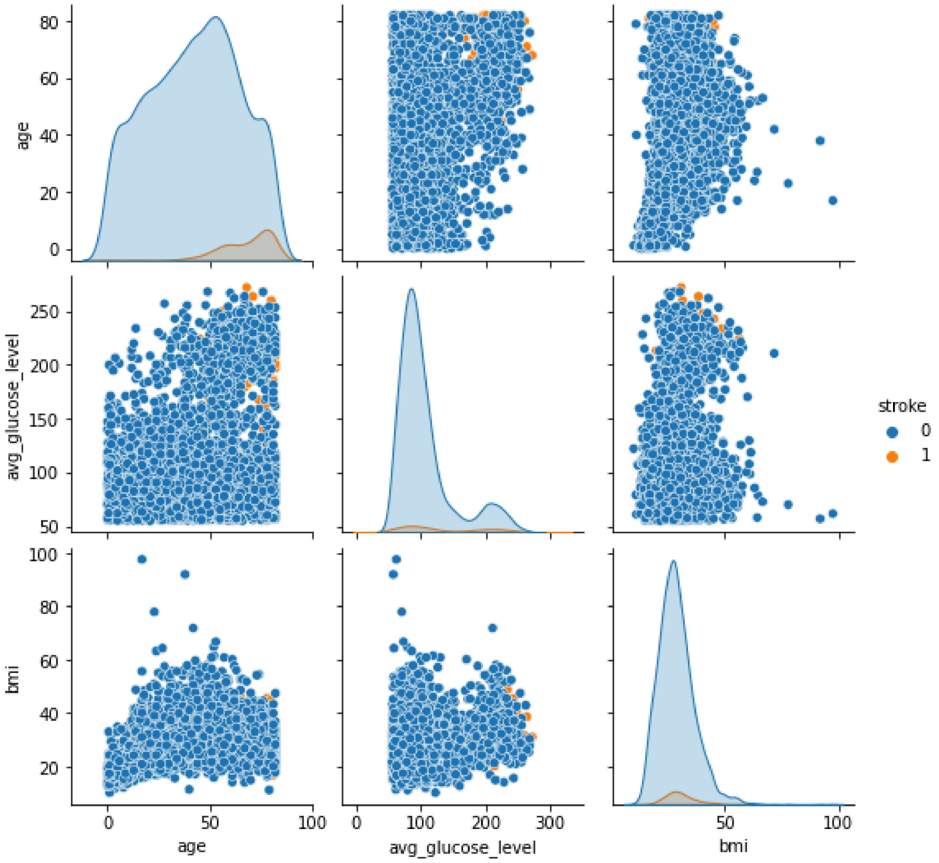
3. Generated a list of top features that are highly correlated with the target column



```
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
```

4. Made a pairplot to get a good view of the relationship



DATA PREPROCESSING

The dataset required several preprocessing steps to ensure it was ready for modeling. First, a single row with the gender labeled as “Other” was dropped, as it represented a rare category and could introduce noise into the model without adding meaningful value.

Next, one-hot encoding was applied to selected binary categorical variables: gender, Residence_type, and ever_married. Using drop_first=True helped avoid multicollinearity by dropping one category from each encoded column.

For the remaining categorical variables, ordinal or label encoding was used based on the nature of the categories:

```
smoking_status was mapped to numerical values:

never smoked = 0

formerly smoked = 1

smokes = 2

Unknown = 3
```

This preserves the general progression of exposure while allowing the model to handle the “Unknown” category explicitly.

work_type was also label encoded, with each category assigned an integer from 0 to 4:

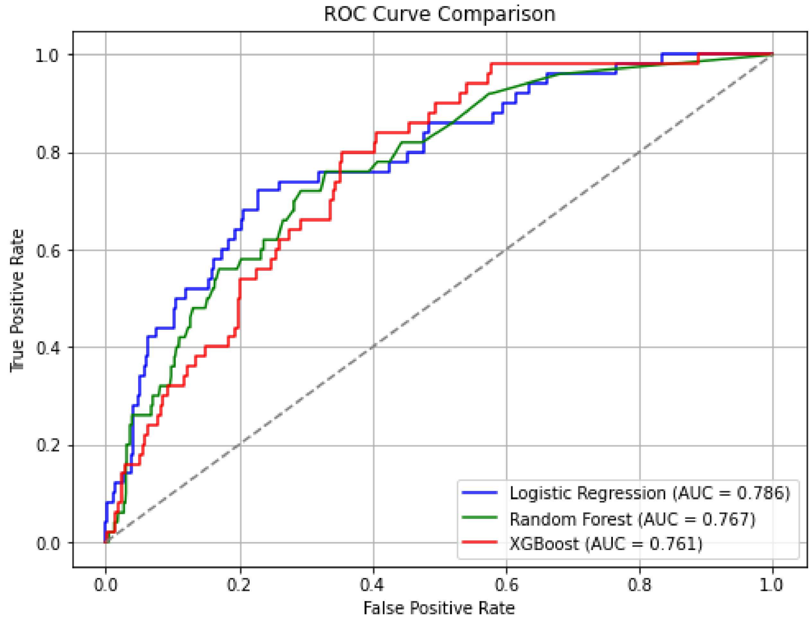
children, Never_worked, Govt_job, Private, and Self-employed were given distinct values to reflect different occupational backgrounds.

This combination of one-hot and label encoding allowed the model to process categorical data efficiently while retaining important information. These encoded features, along with the numerical variables, were then scaled and passed into modeling.

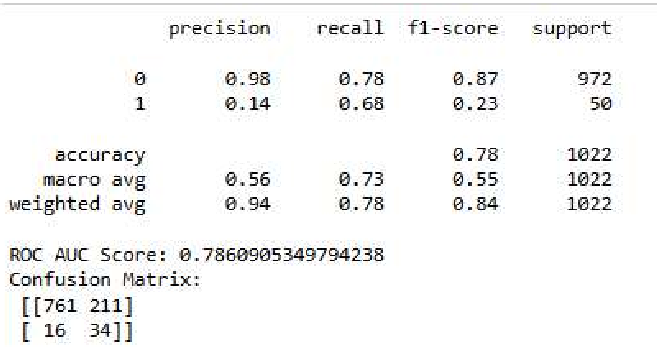
MODELLING

In this part of the project, I focused on building and evaluating models to predict whether a person is likely to have a stroke. Since the dataset is highly imbalanced (very few stroke cases compared to non-stroke), I made sure to use methods that can handle this issue properly.

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.



Evaluation

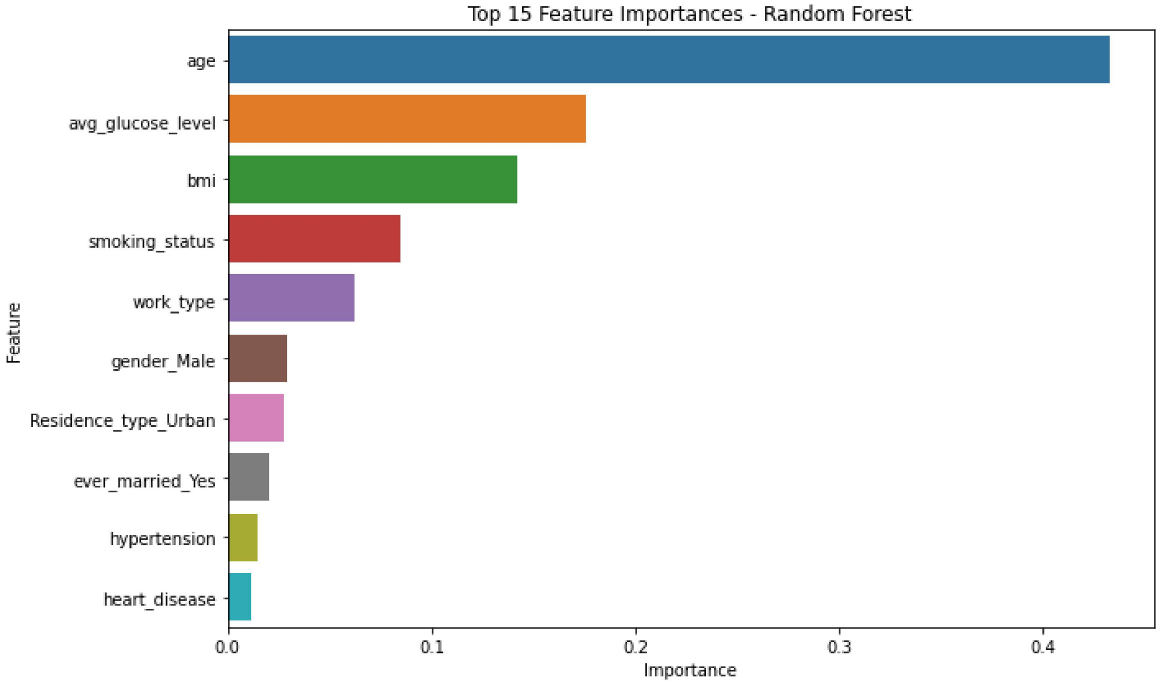

```
              precision    recall  f1-score   support

           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]]
```

A list of top features.



Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%