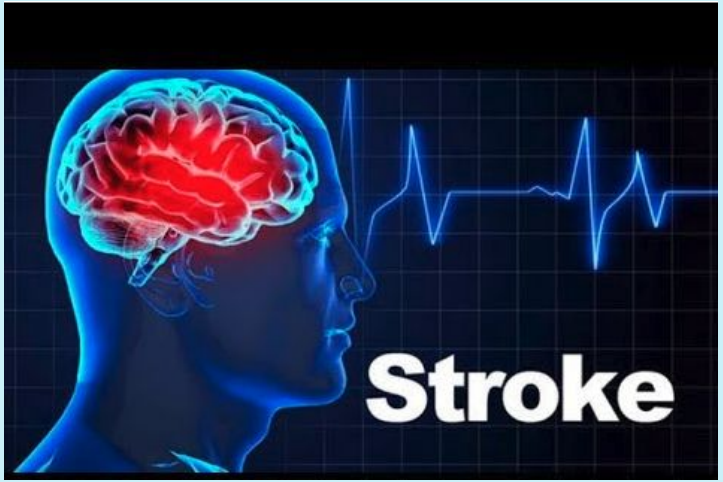
UPenn AI Bootcamp Major Project 2

Dec 2, 2024

AI disruption in the Healthcare Industry

Diagnosing Medical Issues: Using machine learning modelling to predict ***strokes***



Team Members

Larry Azar
Thomas Wells
Jason Clibanoff
Patrick Nwankwo

Overview

Objective: Build a Predictive Model to Identify Stroke

Dataset:

Healthcare dataset* with patient information (gender, age, health metrics, etc.).

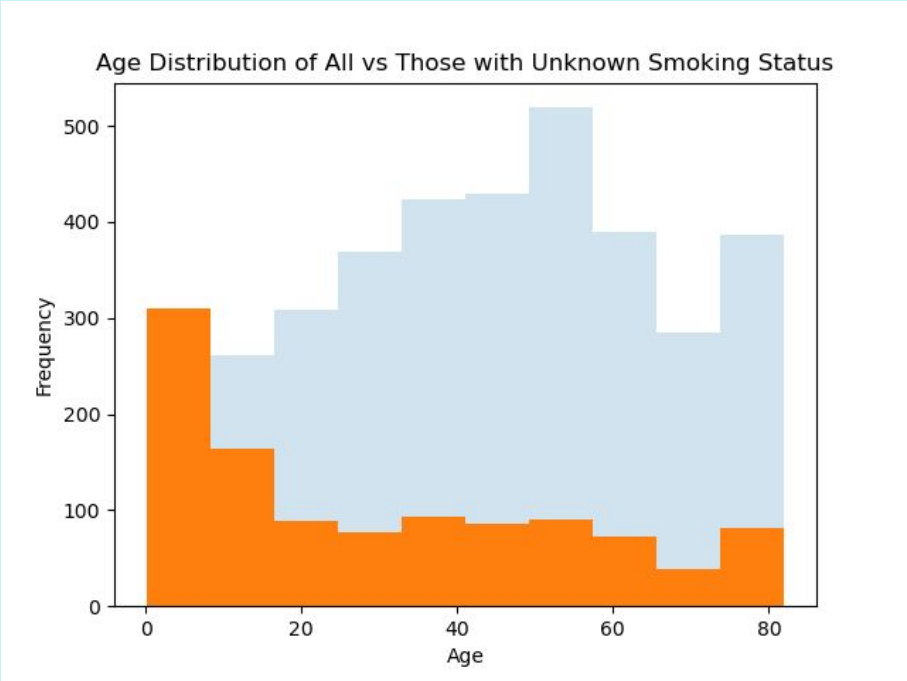
*__Stroke Prediction Dataset (2020)__. Federico Soriano.https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. Marked as confidential data. Kaggle calculated score of 100% for completeness, credibility and compatibility.

Key variables & Features:

- Age
- BMI
- Smoking status
- Glucose levels
- Hypertension
- Heart Disease
- Work Type
- Residence Type

Preparing the Data

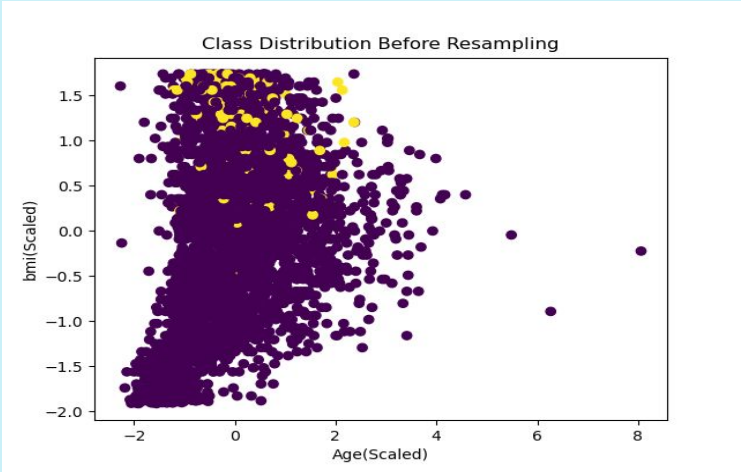
- Dropping our ID Column
- Handling nulls and “unknowns”
- Scaling the data
- Encoding the data

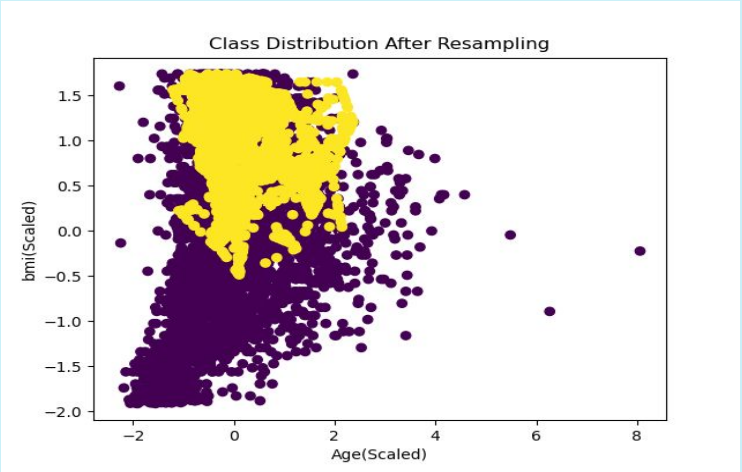


Resampling

Using Synthetic Minority Over-sampling Technique (SMOTE)

- **Imbalanced dataset**
  - Majority class “0” 4,861 (NO stroke)
  - Minority Class “1” 249 (stroke)





Finding the Right Model

- What metrics did we use to score them by?
  - Balanced Accuracy Score
- What models did we look at?
  - Random Forest ✖
  - Decision Trees ✖
  - Logistic Regression ✔
  - Adaboost ✔✔

Results

Class 1 (Stroke) Class 0 (NO Stroke)

SMOTE	Actual Class 0	Actual Class 1	Total Actual Instances	Overall Accuracy Score	Balanced Accuracy Score	Precision Macro avg	Recall Macro avg	f1-score Macro avg	Precision weighted avg	Recall weighted avg	f1-score weighted avg
Random Forest	1,169	58	1,227	0.95	0.51	0.58	0.51	0.50	0.92	0.95	0.93
Decision Tree	1,169	58	1,227	0.84	0.57	0.53	0.57	0.52	0.92	0.84	0.87
Logistic Regression	**1,169**	**58**	**1,227**	**0.74**	**0.72**	**0.55**	**0.72**	**0.52**	**0.94**	**0.74**	**0.81**
AdaBoost	**1,169**	**58**	**1,227**	**0.73**	**0.73**	**0.55**	**0.73**	**0.52**	**0.94**	**0.73**	**0.81**

Results - Tuning

Model	Overall accuracy	Balanced Accuracy Score
Logistic Regression	0.74	0.72
AdaBoost	0.73	0.73

- AdaBoost, tuned with **GridSearchCV**, yielded a best cross-validated BAC of 84.9% after applying a 5-fold cross validation, using 300 decision trees and a learning rate of 1.0 (optimal parameters).

- Logistic Regression, tuned with **GridSearchCV,** only yielded a best cross-validated BAC of 77.4%

Conclusion

Given these findings, further research and exploration are recommended to refine the AdaBoost model further.

A deeper dive into hyperparameter optimization, feature engineering, and the exploration of alternative models could potentially enhance predictive accuracy and offer insights into the underlying factors contributing to stroke risk.

Thus, additional work in this area is essential for improving the robustness and reliability of stroke prediction models.

Any Questions?