Predicting Stroke through Machine Learning Models

Group: Neuro Ninjas
Presented by:

Aya Fakhri
Kyle Goodwin
Aparajita Mondal
David Zigun

Outline

- Introduce Dataset
- Describe Approach to Data Cleaning
- Discuss Machine Learning Models
- Analysis Outcomes
- Implications
- Limitations
- Recommendations for Future Research

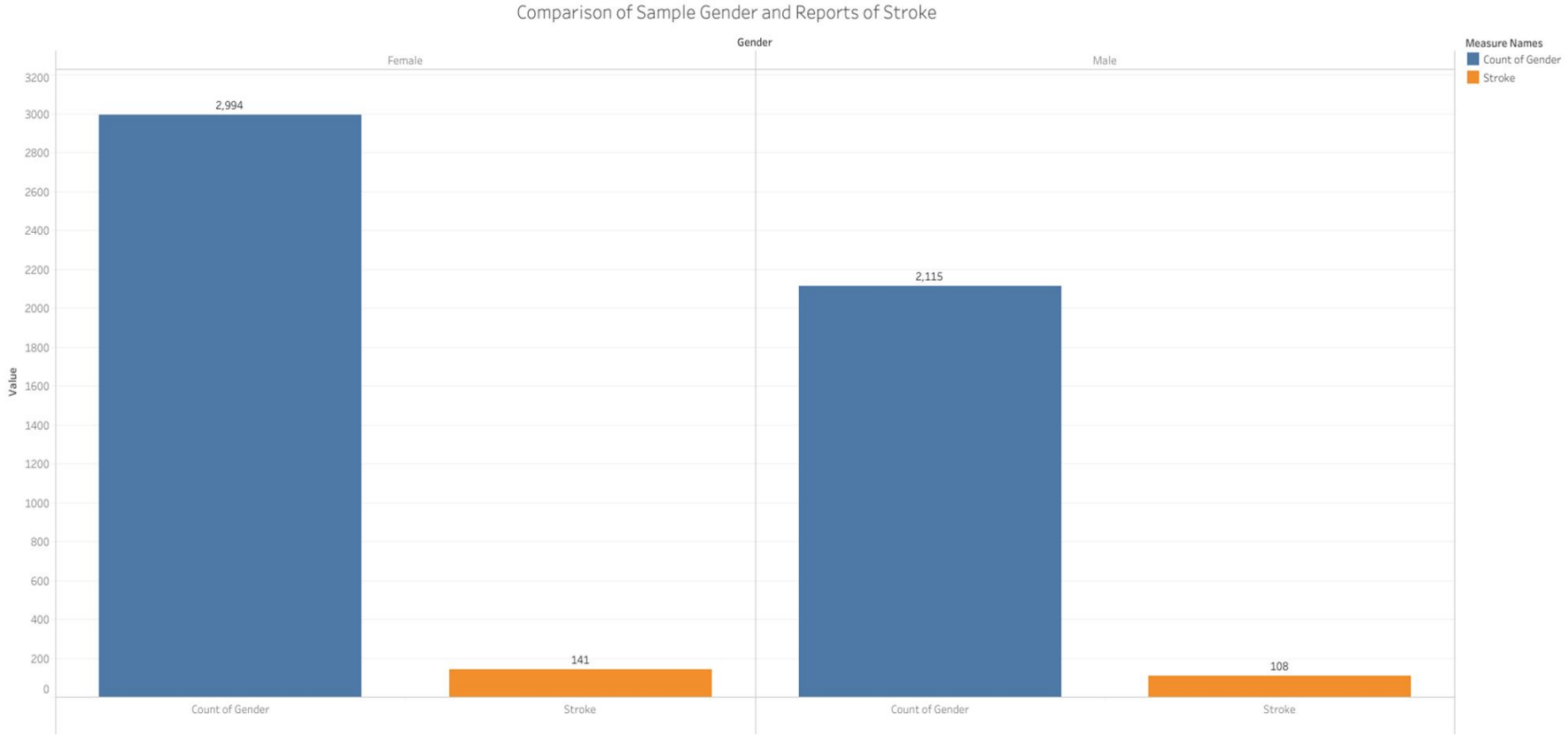
The Dataset

- Gender
- Age
- Heart Disease
- Hypertension
- Ever Married
- Work Type
- Residence Type
- Avg. Glucose
- BMI
- Smoking Status
- Stroke Status

Federico Soriano Palacios,
Associate @ Oliver Wyman

- Confidential Dataset from 2020
- https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

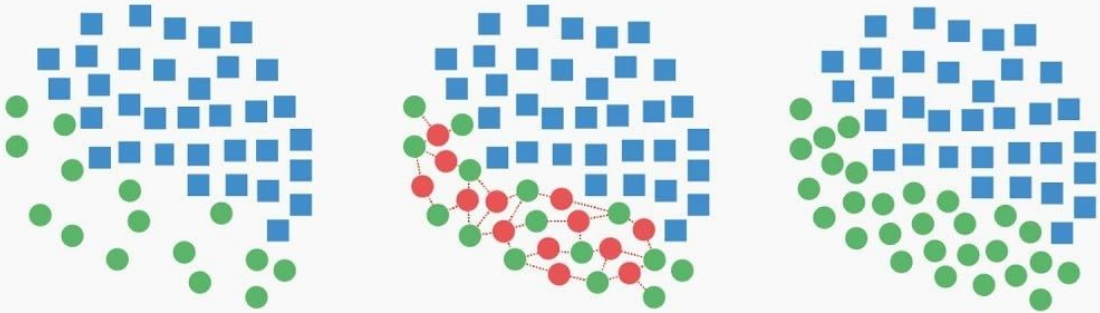
Basic Sample Pre-Processing



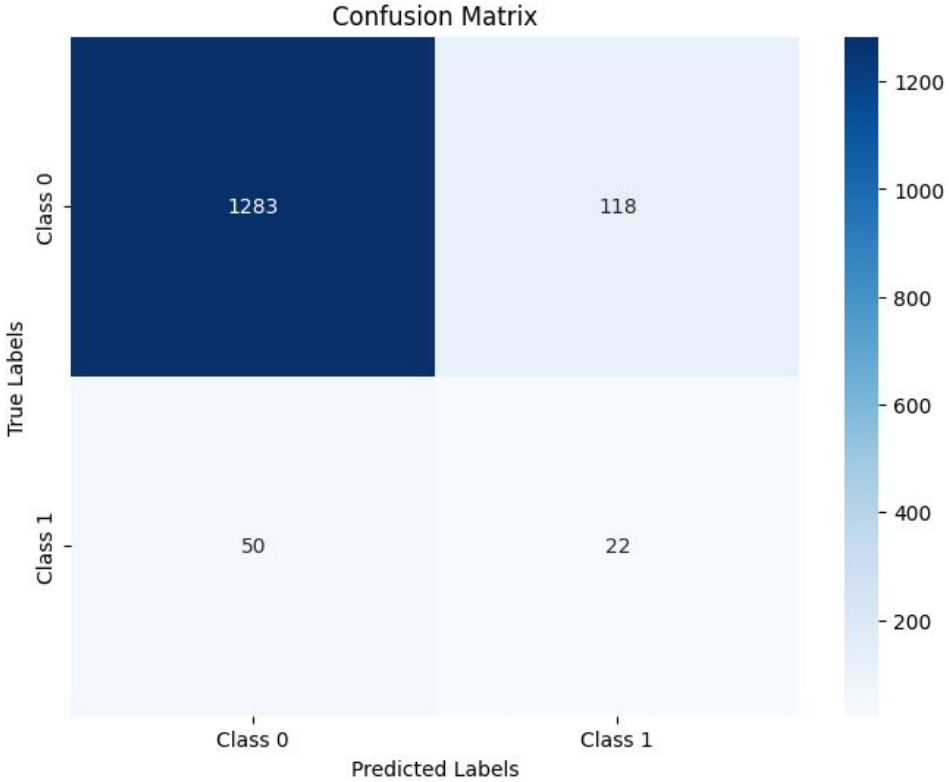
SMOTE

HANDLE IMBALANCED DATASET

Synthetic Minority Oversampling Technique



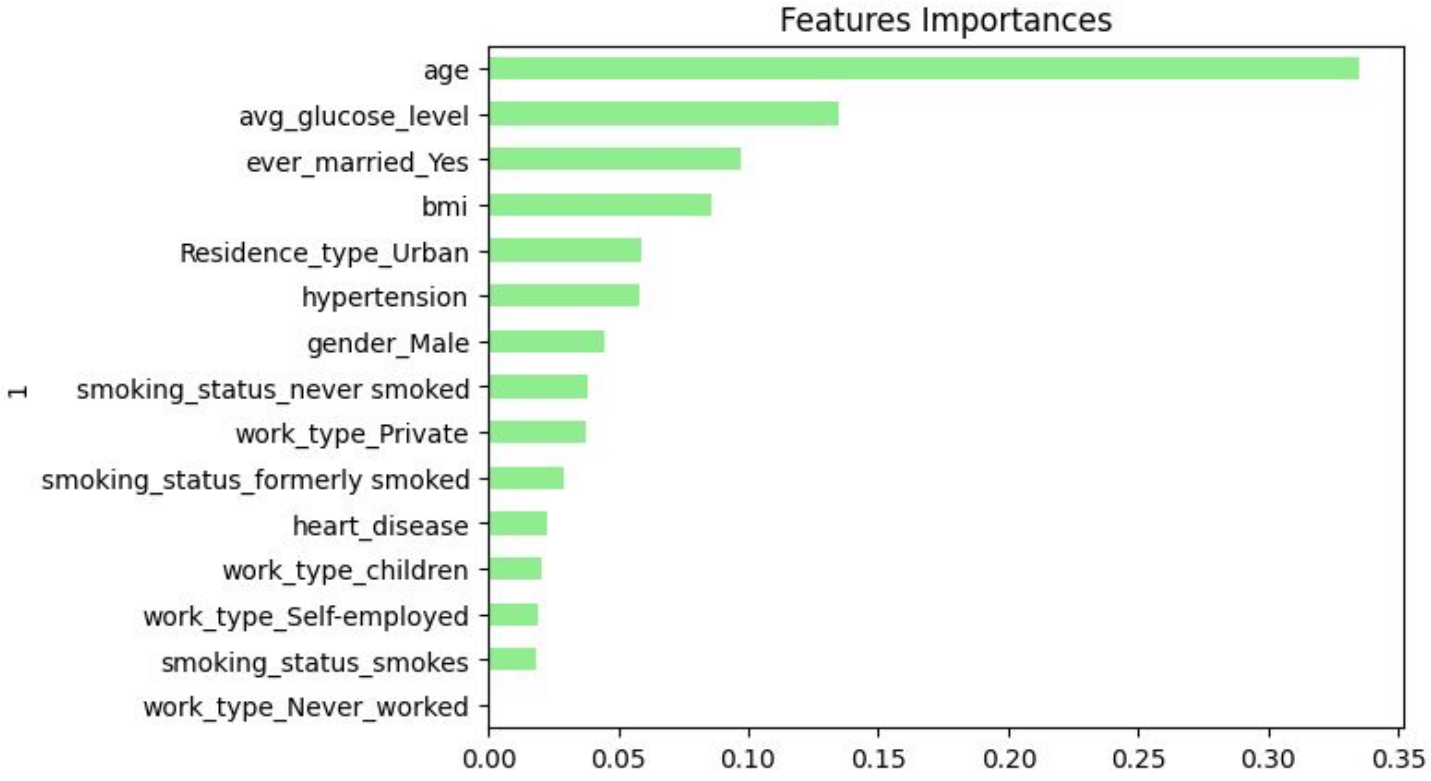
Random Forest



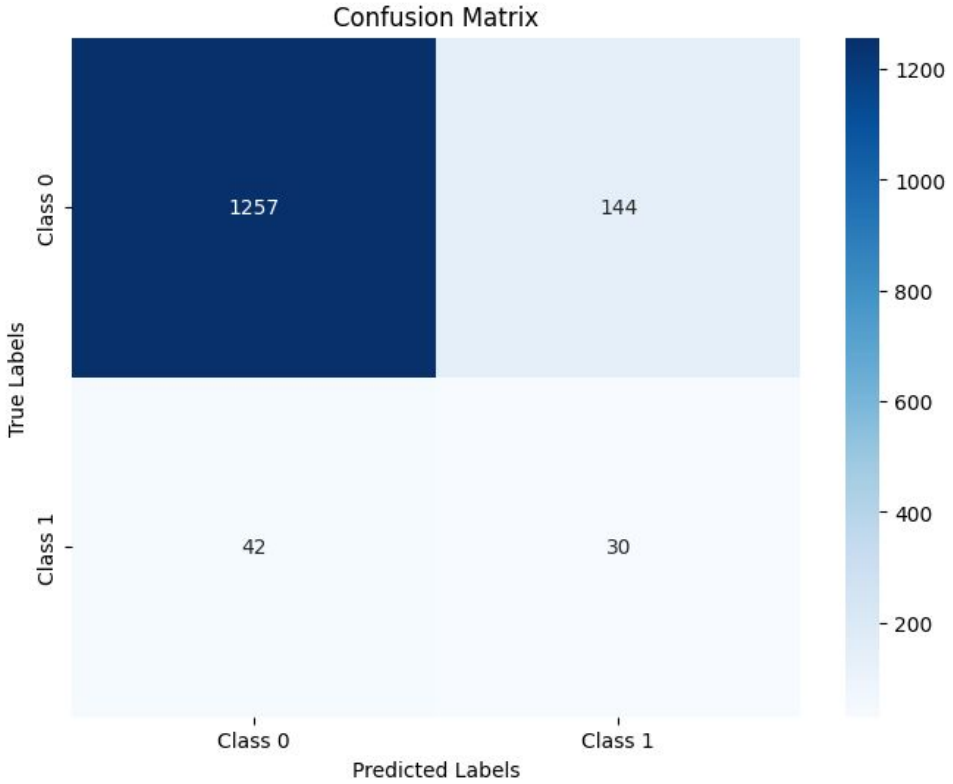
Accuracy Score : 0.8859470468431772
Classification Report

	precision	recall	f1-score	support
0	0.96	0.92	0.94	1401
1	0.16	0.31	0.21	72
accuracy			0.89	1473
macro avg	0.56	0.61	0.57	1473
weighted avg	0.92	0.89	0.90	1473

Random Forest

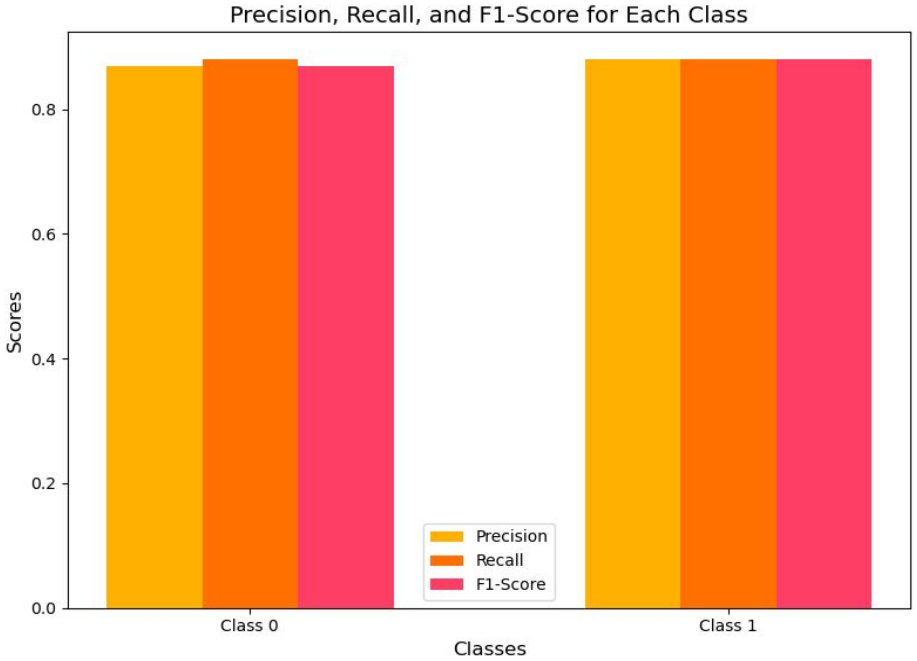


Random Forest - with work type removed



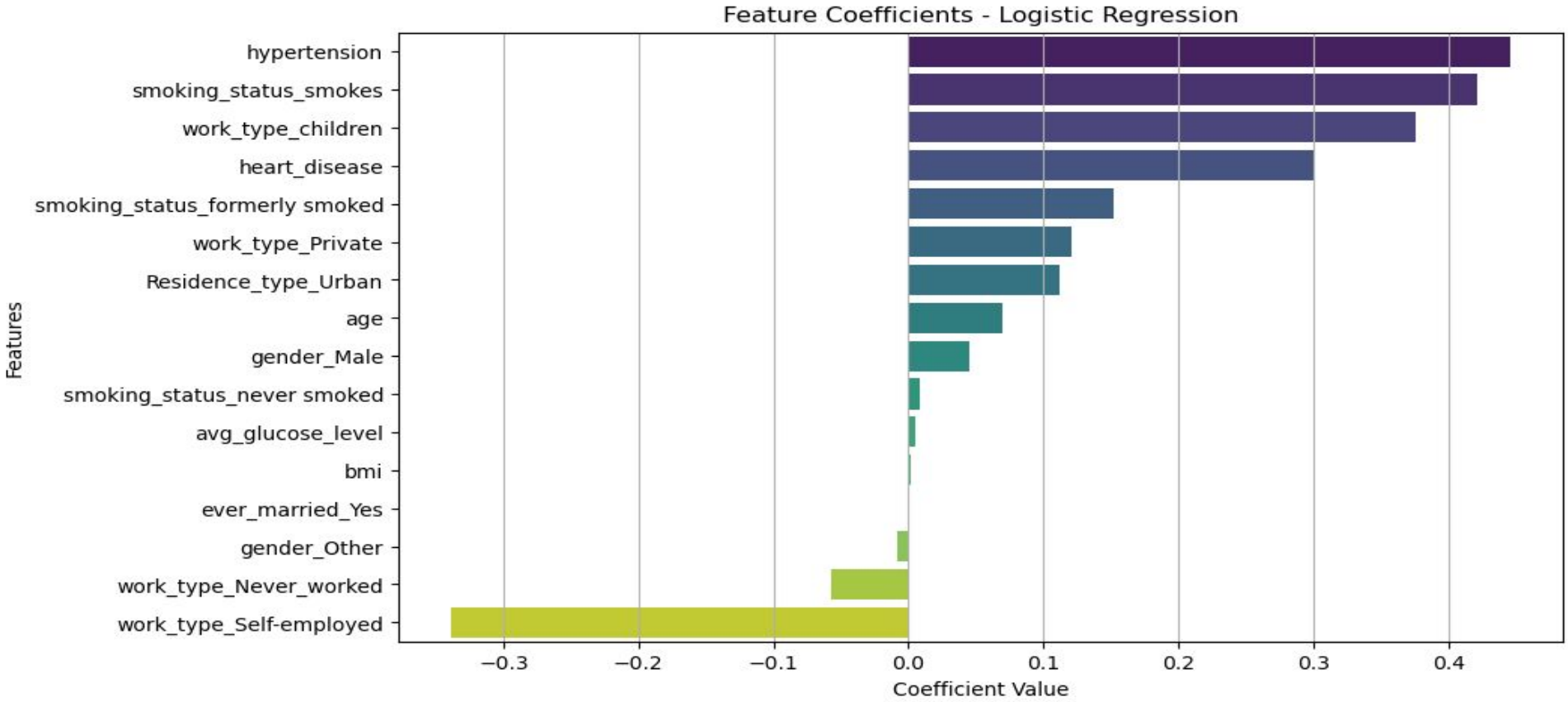
	precision	recall	f1-score	support
0	0.97	0.90	0.93	1401
1	0.17	0.42	0.24	72
accuracy			0.87	1473
macro avg.	0.57	0.66	0.59	1473
weighted avg.	0.93	0.87	0.90	1473

Classification Regression : Model Evaluation Results

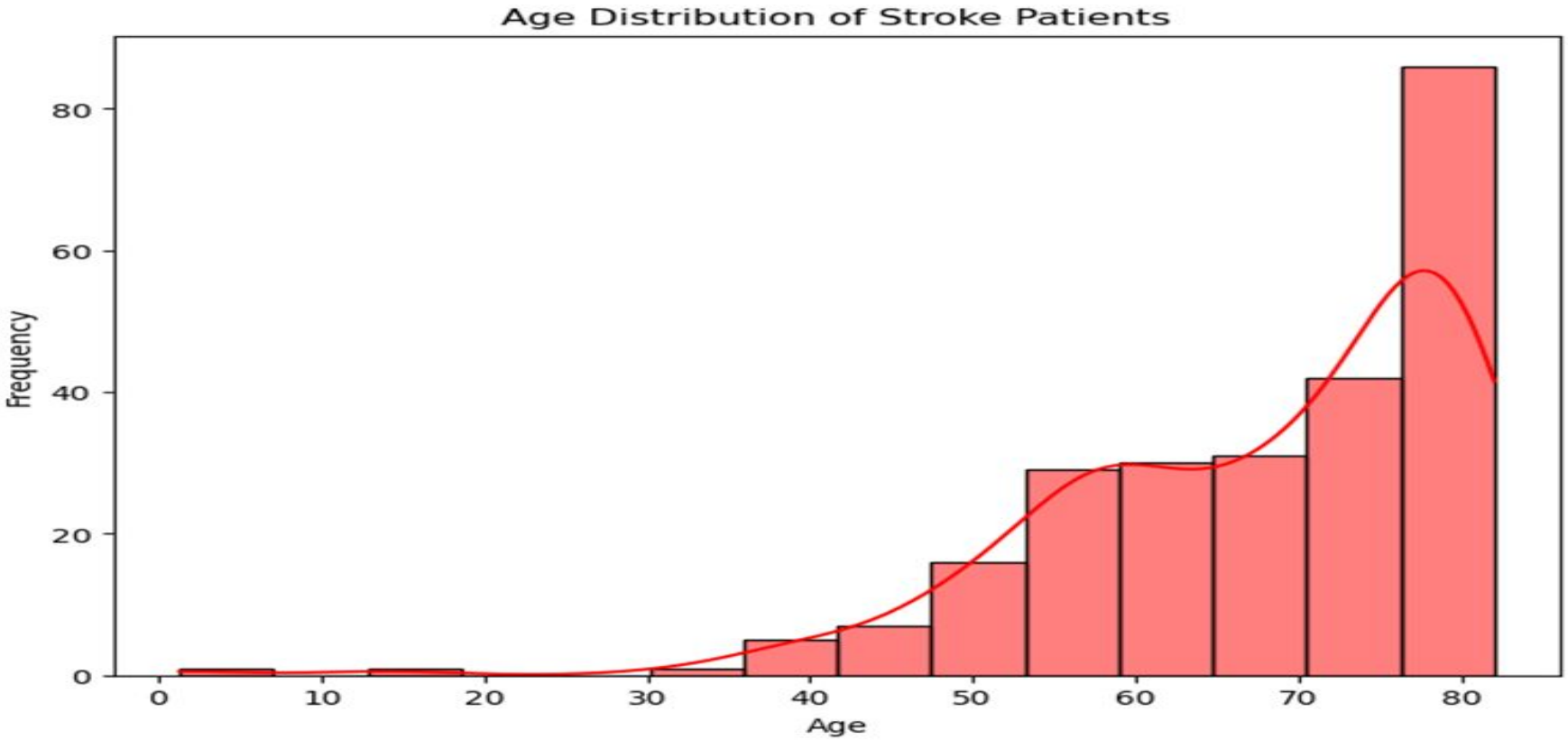



```
Best Parameters: {'C': 1, 'solver': 'lbfgs'}
Accuracy: 0.877127659574468

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.88      0.87       922
           1       0.88      0.88      0.88       958

    accuracy                           0.88      1880
   macro avg       0.88      0.88      0.88      1880
weighted avg       0.88      0.88      0.88      1880
```

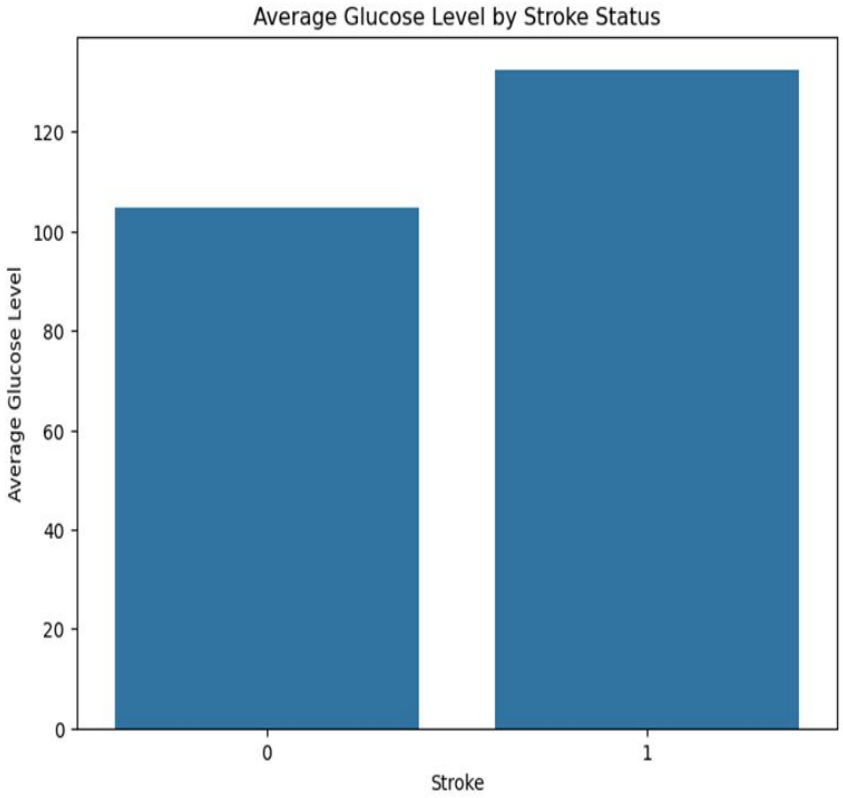
- Model Accuracy Report



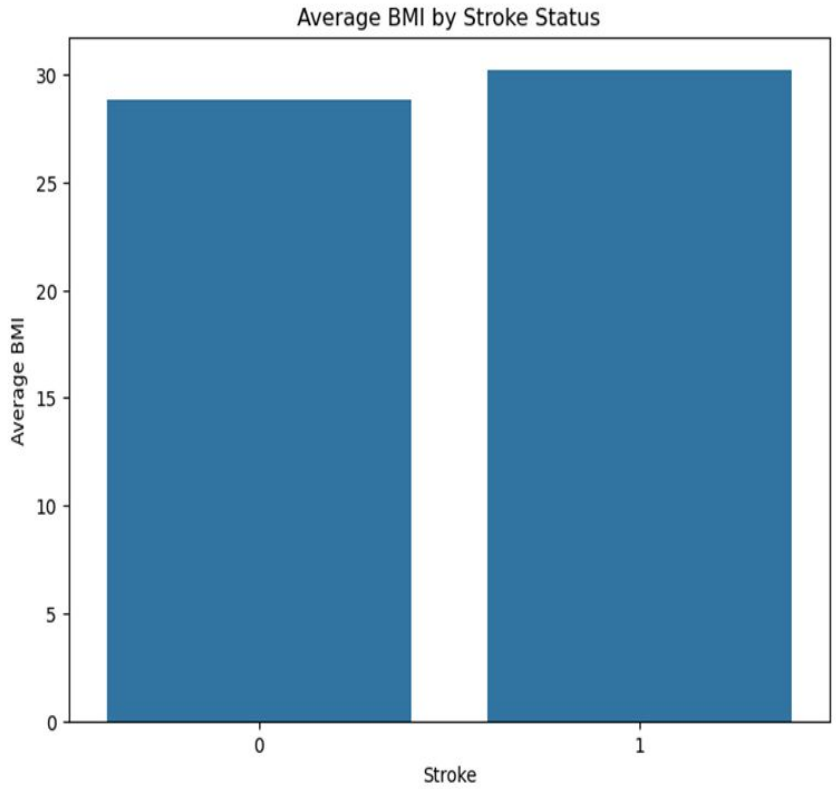
- Matplotlib Visualization : 1. Age Distribution of Stroke Patients



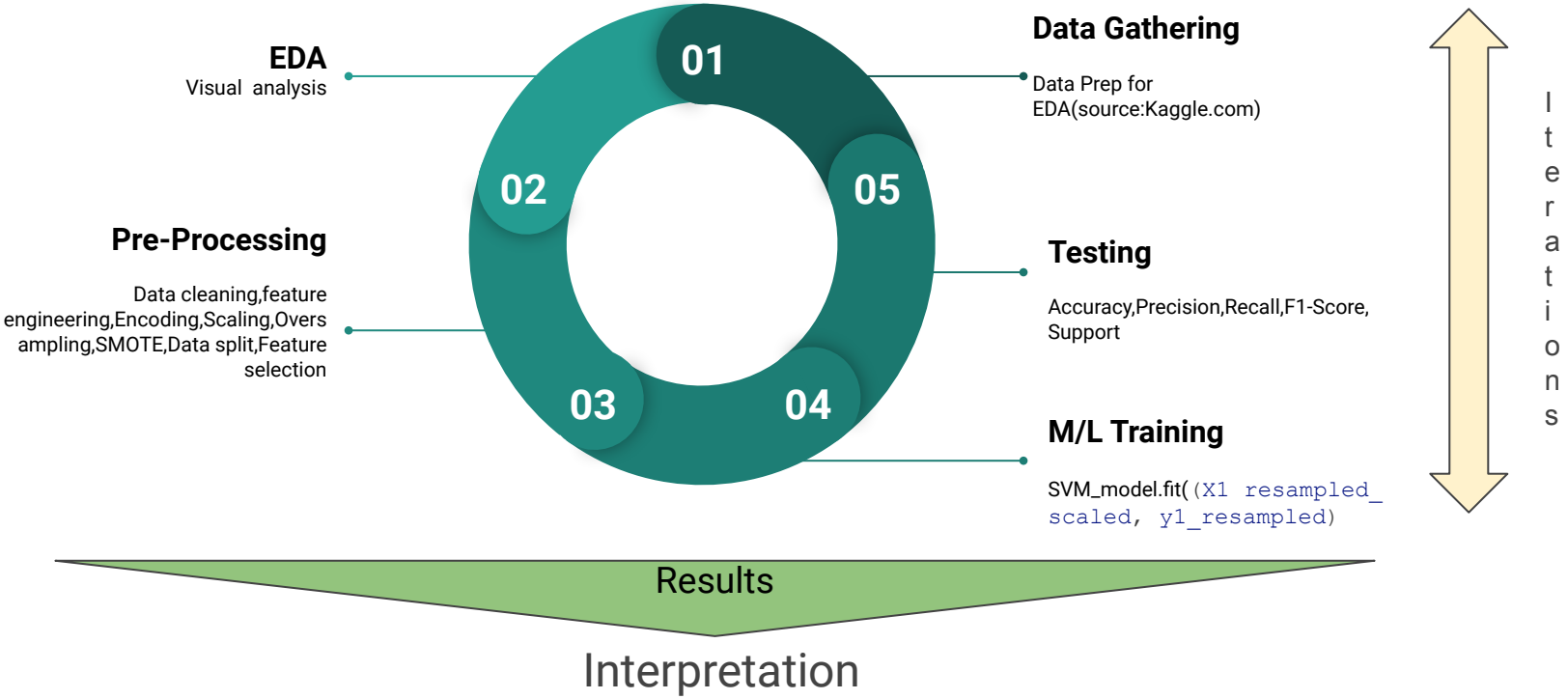
- 2. Average Glucose Levels by Stroke Outcome



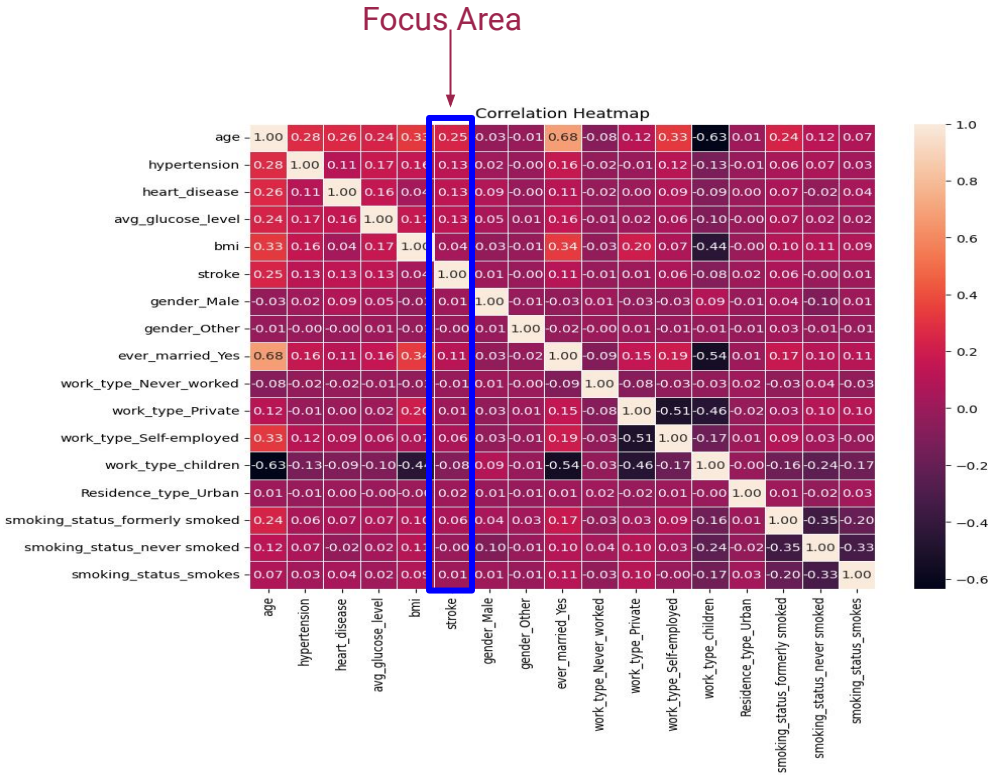
3. BMI Distribution by Stroke Outcome

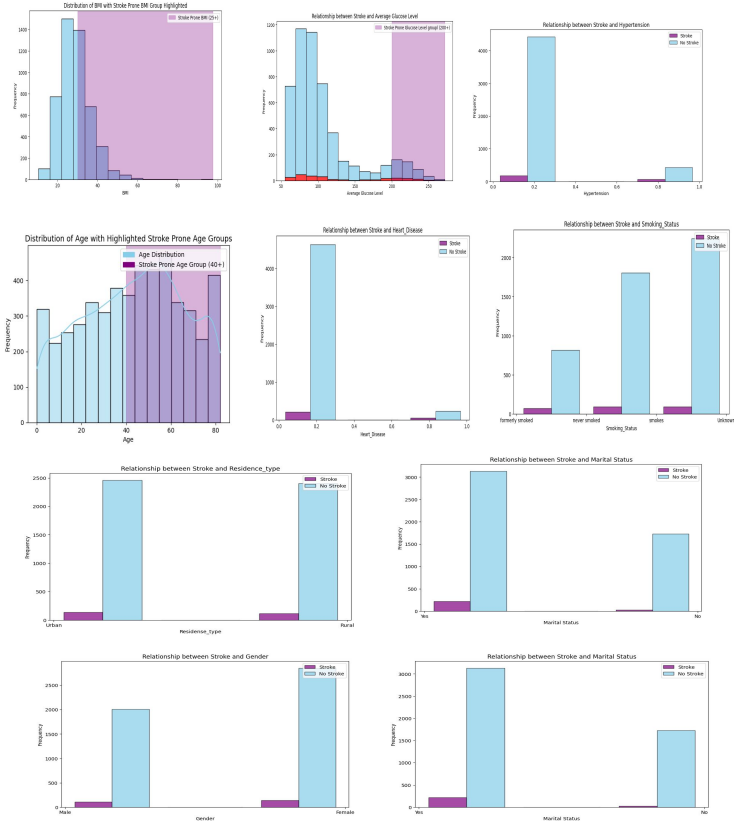


Support Vector Machine (SVM) : Process Flow



Correlation and Data Distribution between different input data attributes





Interpretation

Tests (SVM)	Description	Confusion Matrix				Classification Report(for stroke only)				
		TN	FN	FP	TP	Accuracy	Precision	Recall	F1-Score	Support
Test1(A)	With all the features-('age','Average_glucose_level','bmi' not scaled)	782	27	178	35	0.8	0.16	0.56	0.25	62
Test1(B)	With all the features-(age','Average_glucose_level','bmi' scaled)	683	14	246	39	0.74	0.14	0.74	0.23	53
Test2	only With the Health_situation features-(age','Average_glucose_level','bmi' scaled)	689	11	271	51	0.72	0.16	0.82	0.27	62

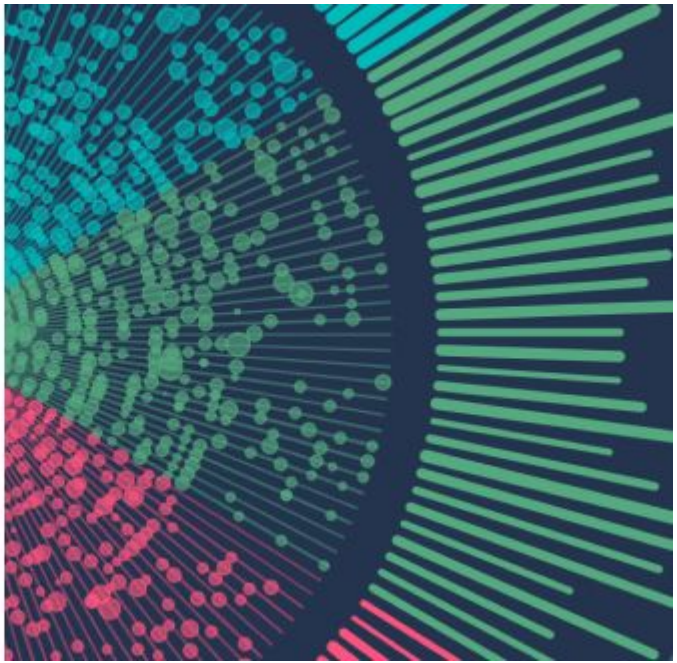
Summary:

- out of the these three testings for **Test2** has superior recall and F1-Score, ensuring better detection of stroke cases.
- This choice aligns with the goal of minimizing the risk of missing stroke diagnoses while maintaining a reasonable balance of precision and recall.

Visualization of Data Sample - Tableau

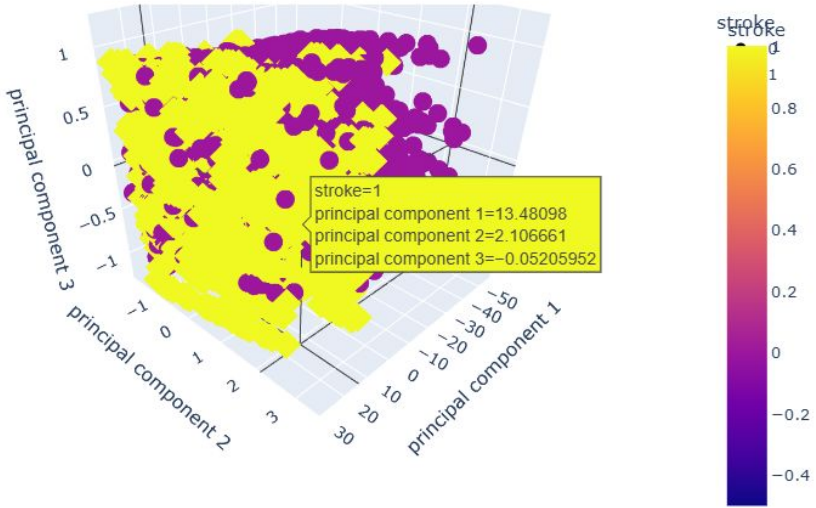
Pre-process data for better visualization

- Bin BMI into groups based on CDC info (CDC BMI grouping)

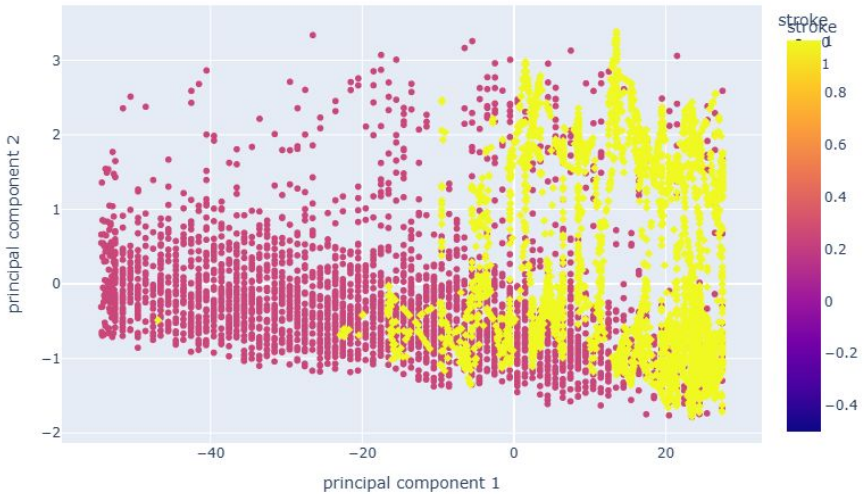


Principal Component Analysis (PCA)

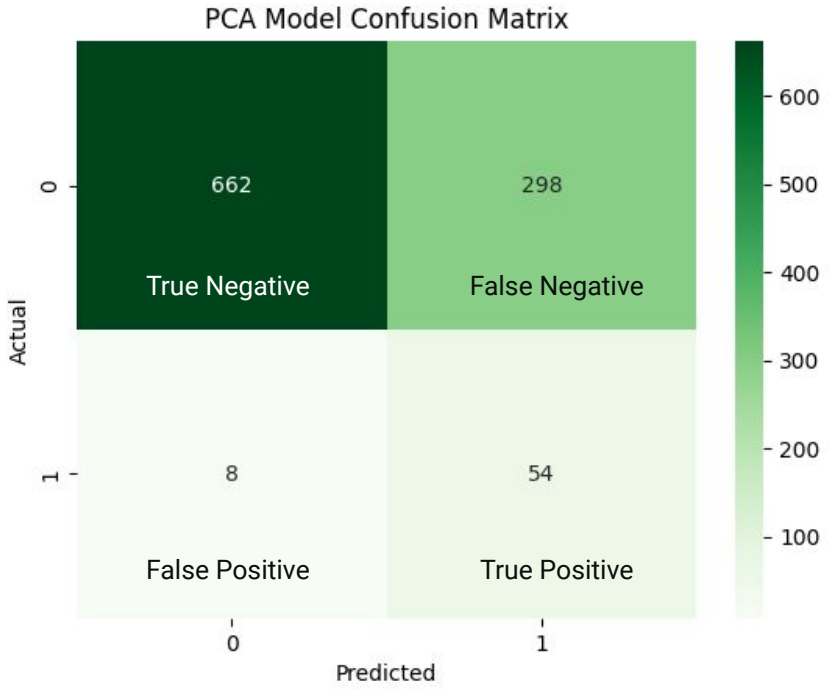
3D Plotting of Stroke on PCA 1-3



2D Plotting of Stroke on PCA 1 & 2



Confusion Matrix and Classification



PCA Classification Report

	precision	recall	f1-score	support
0	0.99	0.69	0.81	960
1	0.15	0.87	0.26	62
accuracy			0.70	1022
macro avg	0.57	0.78	0.54	1022
weighted Avg	0.94	0.70	0.78	1022

Closer to 1 = better fit

Precision
True positives / true positives + false positives

Recall
True positives / true positives + false negatives

F1 Score
The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

Support
# of actual occurrences of the class in the dataset.

Limitations

Challenge 1

Class Imbalance

- The dataset has significantly fewer instances of one class (e.g., Stroke = 1) compared to the other (Stroke = 0).
- This imbalance can lead to models being biased toward the majority class.

Challenge 2

Binary and Simplified Features

- Features like Hypertension and Heart Disease are binary (0/1), which may oversimplify complex health conditions.

Challenge 3

Potential Data Bias

- The dataset may be collected from a specific region, healthcare system, or demographic.
- This limits the model's ability to generalize to other populations.

Recommendations

- Use techniques like oversampling (e.g., SMOTE)
- Incorporate additional data, such as medical history or lab results, if available.
- Use diverse data sources and perform bias analysis to ensure fairness.

Thank you