Context

In-house Data Scientist from MOH

The semi-technical audience are stakeholders and doctors.

The stakeholders presented to us an issue about being able to detect potential of stroke as early as possible in A&E with health information gathered at the registration and triage.

We were not given patients' data due to privacy, so we had to source from elsewhere.

MINISTRY OF HEALTH

DSI-SG-42 | Michael King Sutanto | Eugene Matthew Cheong | Pius Yee

Obesity and Stroke

DSI-SG-42 | Michael King Sutanto | Eugene Matthew Cheong | Pius Yee

Being categorised as **overweight** **increases** your risk of stroke by **22%** and if you are **obese** that risk **increases** by **64%**. This is because **carrying too much weight** increases your risk of **high blood pressure**, **heart disease**, **high cholesterol** and **type 2 diabetes** which all contribute to higher stroke risk.

-World Stroke Organization

Bernard Tan

A&E Patient

A 40-year-old male stands at 175cm tall and weighs 130kg, feeling uncertain about whether his weight poses risks for obesity or stroke.

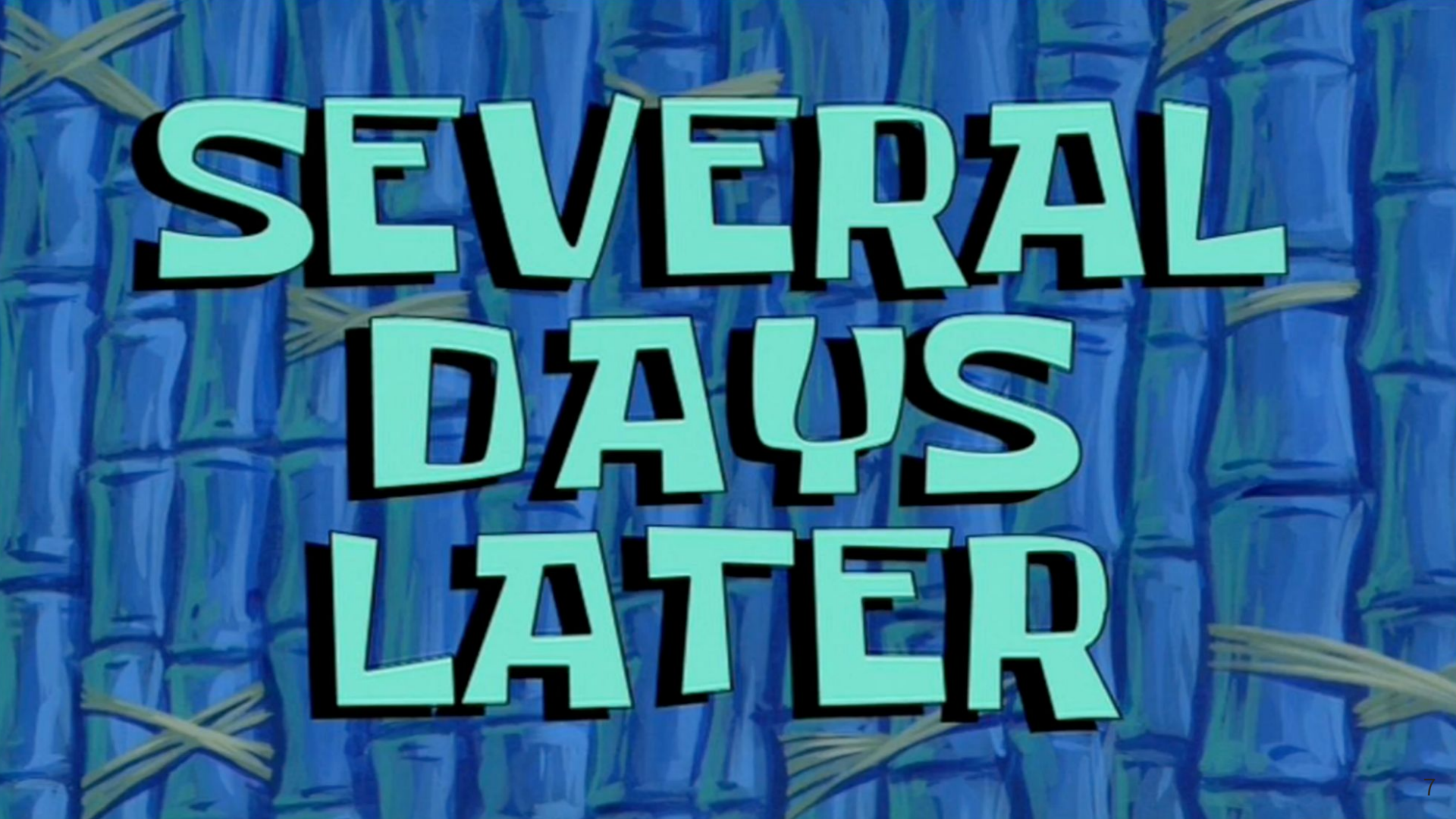
He suddenly experienced sudden numbness on the arm, and felt giddiness and headache.

George, 38

A&E Doctor

He understood that Bernard was experiencing numbness for that short period. And George knows that Bernard's weight might be a concern for risk of stroke.

Due to the lack of resources and high volume of patients that he needed to attend to, he gave some medication and told to be more active, exercise and monitor the situation.


SEVERAL
DAYS
LATER

Ischemic Stroke

An ischemic stroke occurs when a portion of the brain is deprived of oxygen and essential nutrients due to a blockage or reduction in blood supply. This deprivation leads to the rapid death of brain cells within minutes.

Problem Statement

The healthcare system faces a challenge in quickly identifying patients at risk of stroke upon arrival at the emergency department (A&E), despite standardized triage protocols. Swiftly discerning stroke risk based on initial health particulars such as **obesity** is difficult, leading to delays in referrals to specialists for preventive measures.

How might we help A&E identify patients at risk of stroke easier?

A Solution Needed

There's a need for a streamlined system using detailed patient information, such as better **classification** of obesity status, to **improve** stroke risk assessment during **triage**, allowing A&E physicians to refer at-risk patients promptly and **reduce** the likelihood of **stroke occurrence**.

Bernard

Application

George

Application

Machine Learning Models

- Obesity Classification Model
- Stroke Detection Model
- CT Scan Image Stroke Detector

Bernard

Obesity Classification Model

Stroke Detection Model

CT Scan Image Stroke Detector

George

Bernard

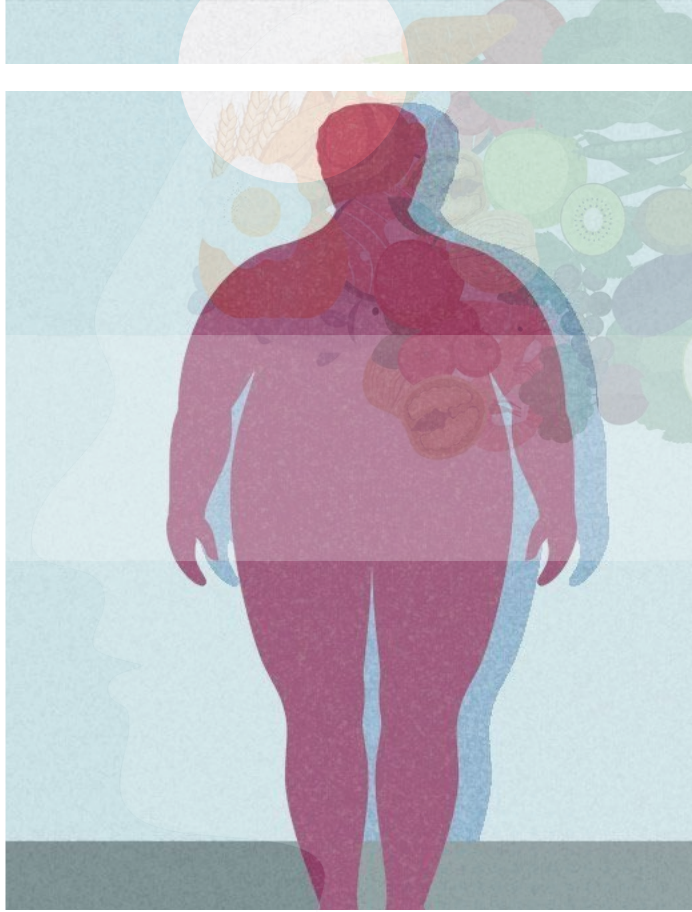
Obesity Classification Model

Stroke Detection Model

CT Scan Image Stroke Detector

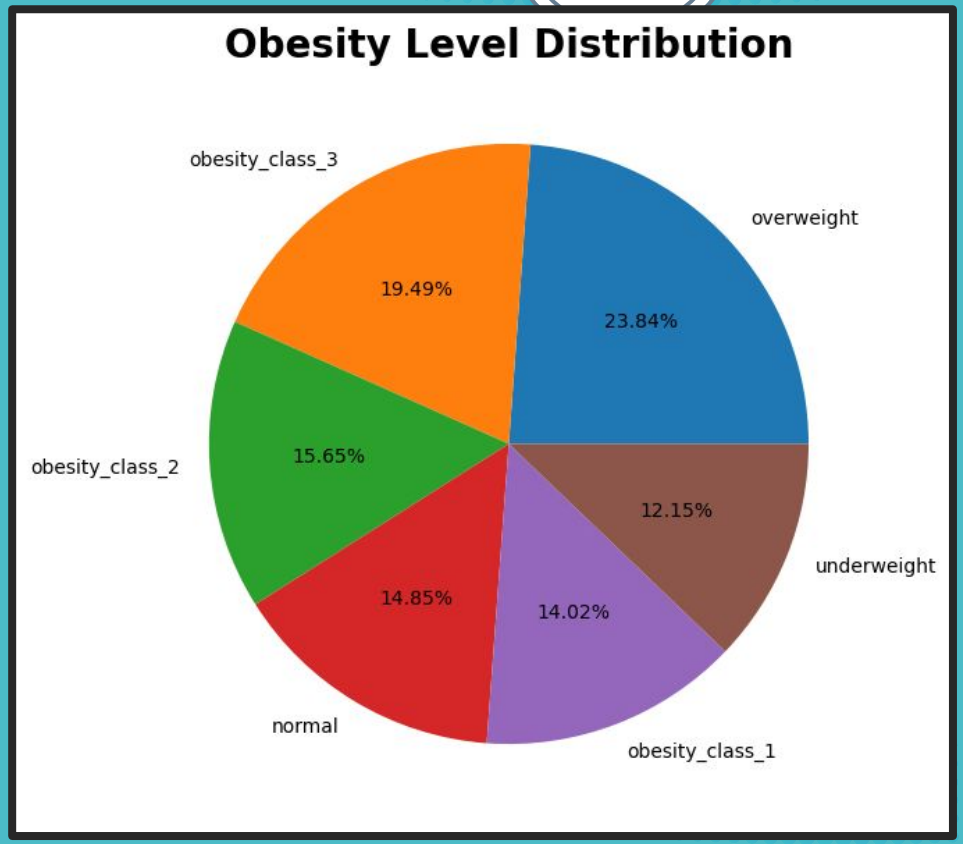
George

Obesity Dataset



Obesity Dataset

- Dataset was taken from Kaggle.
- Dataset has around 20,000 rows.
- Obesity Status distribution is quite balance.



BMI?

Body Mass Index is a person's weight in kilograms divided by the square of height in meters. A high BMI can indicate high obesity.

Obese	≥ 30
Obese class I	30.0 – 34.9
Obese class II	35.0 – 39.9
Obese class III	≥ 40

BMI?


HEALTH NEWS

BMI is a flawed way to measure obesity, experts say. What else works?

Body mass index isn't the only way to determine if a person has obesity. Here's how other assessments stack up.

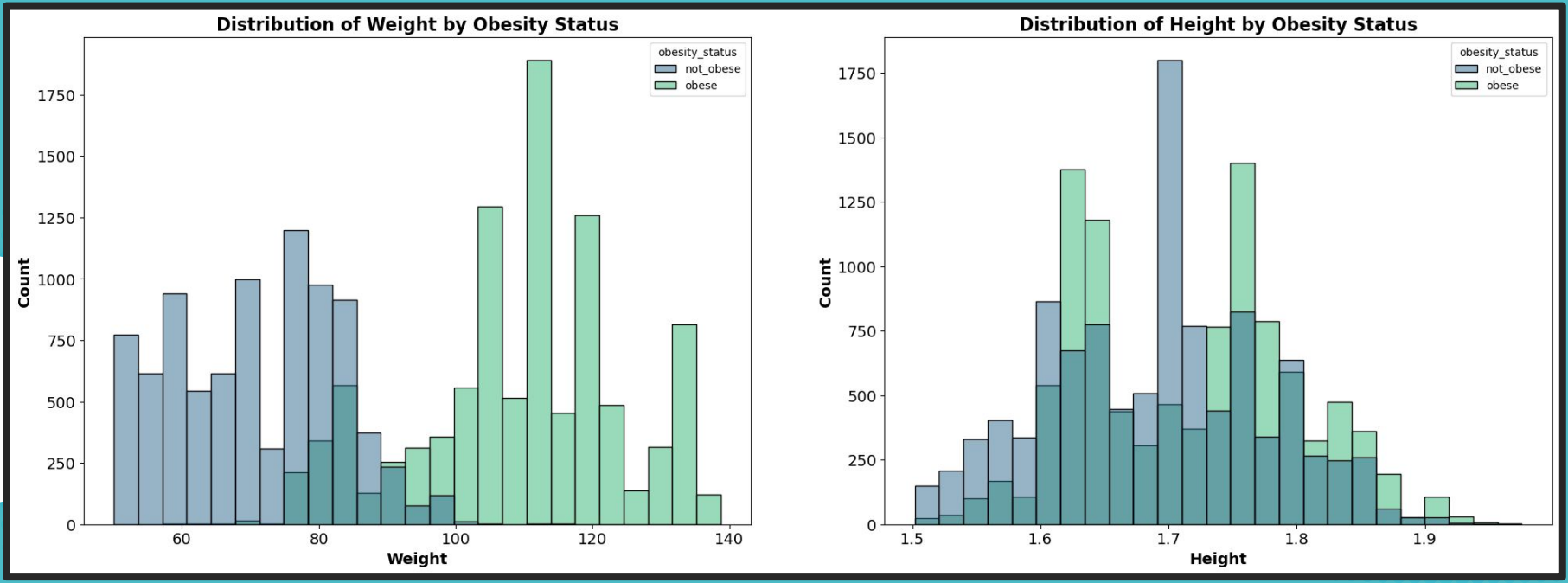
www.nbcnews.com


HEALTH

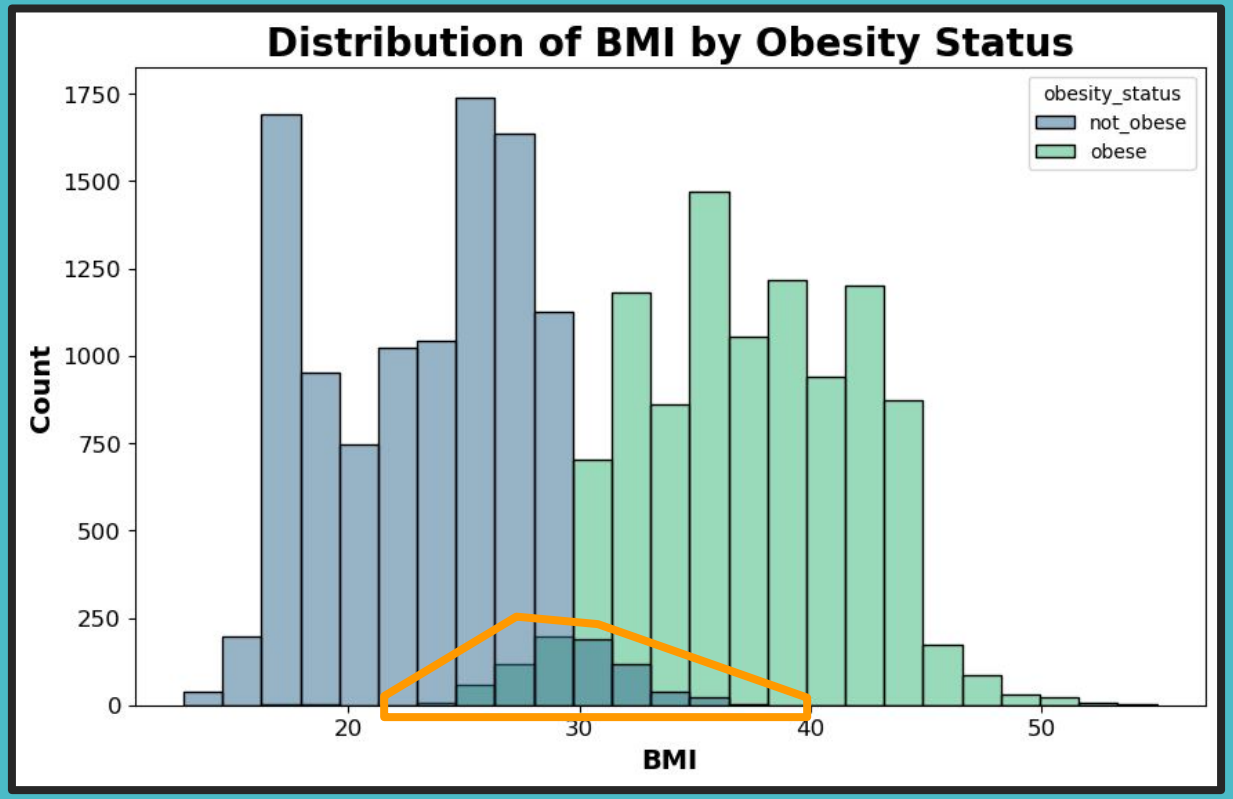
7 signs you don't have overweight or obesity, regardless of BMI

www.businessinsider.com

Weight and Height Distribution by Obesity Status



BMI Distribution by Obesity Status



BMI Above 30 and Not Obese

Obese	≥ 30
Obese class I	30.0 – 34.9
Obese class II	35.0 – 39.9
Obese class III	≥ 40

	Gender	Age	Height	Weight	BMI	obesity_status
29	Female	23.000000	1.550000	80.000000	33.298647	not_obese
124	Male	22.771612	1.769328	96.875502	30.945483	not_obese
139	Female	19.633898	1.500000	71.602622	31.823388	not_obese
186	Male	21.000000	1.650000	83.000000	30.486685	not_obese
294	Male	18.000000	1.650000	85.000000	31.221304	not_obese
...	...	...	...	...	...	...
20590	Male	17.992717	1.654067	83.016968	30.343151	not_obese
20603	Female	18.000000	1.560000	80.000000	32.873110	not_obese
20631	Female	23.000000	1.584951	80.562213	32.070055	not_obese
20725	Female	38.943282	1.565366	80.000000	32.648121	not_obese
20747	Male	33.000000	1.720000	99.000000	33.464035	not_obese

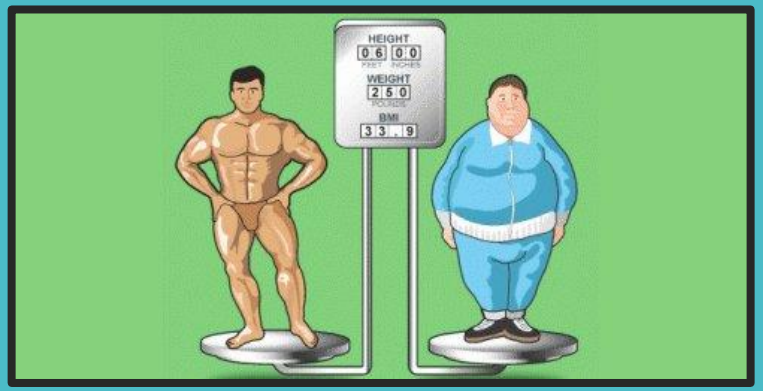
BMI Above 35 and Not Obese

Obese	≥ 30
Obese class I	30.0 – 34.9
Obese class II	35.0 – 39.9
Obese class III	≥ 40

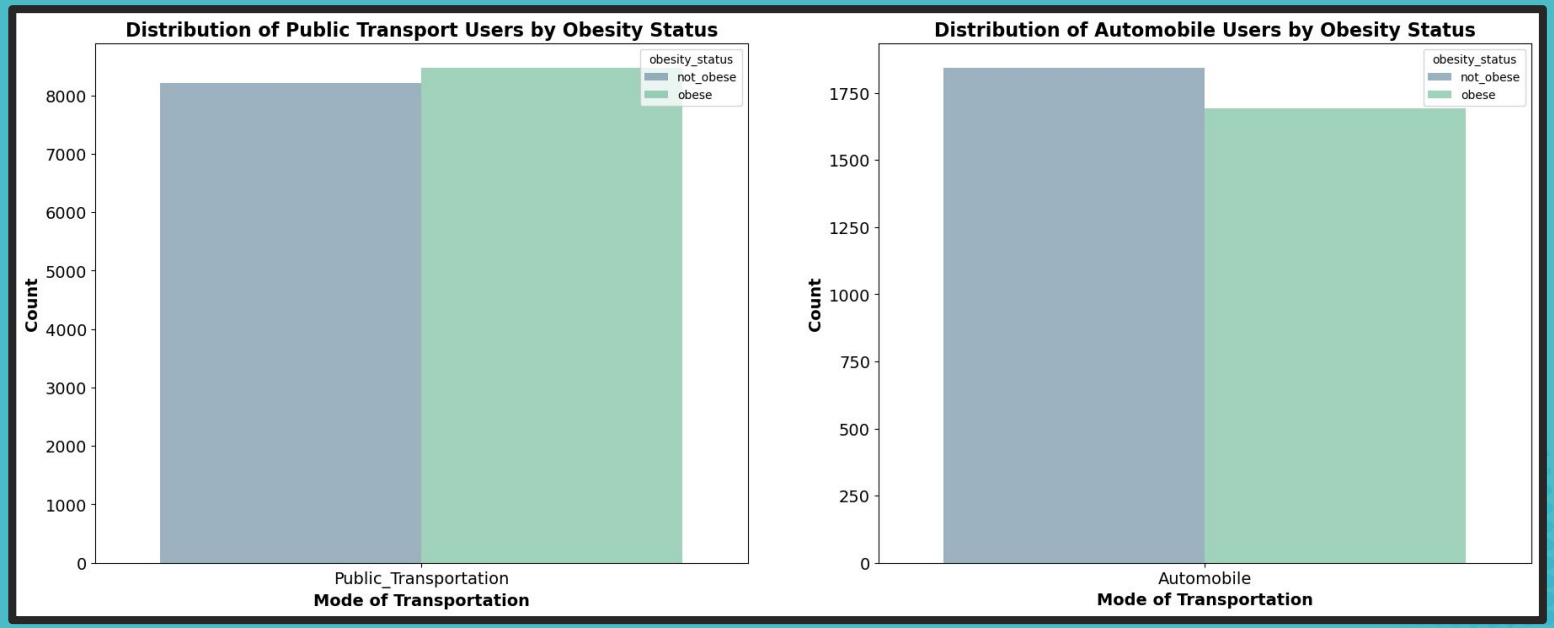
	Gender	Age	Height	Weight	BMI	obesity_status
453	Male	33.185661	1.700627	103.669116	35.845220	not_obese
1946	Female	35.125401	1.507867	79.989789	35.181024	not_obese
2841	Female	40.702771	1.500000	79.252472	35.223321	not_obese
3527	Female	37.965430	1.508908	80.000000	35.136983	not_obese
3684	Female	34.176795	1.505387	79.697278	35.167959	not_obese
4590	Male	39.825592	1.502609	79.414603	35.172918	not_obese
5965	Female	55.000000	1.500000	80.000000	35.555556	not_obese
6217	Female	25.000000	1.500000	80.000000	35.555556	not_obese
7181	Female	38.445148	1.498561	80.000000	35.623873	not_obese
7952	Female	20.242237	1.507106	80.790813	35.569223	not_obese
8087	Male	31.199261	1.699474	101.532762	35.154194	not_obese
8742	Male	28.421533	1.758382	113.714521	36.778115	not_obese
13420	Male	19.266287	1.507867	79.697278	35.052372	not_obese
15037	Male	24.751511	1.505387	83.263120	36.741456	not_obese
16404	Male	33.226808	1.669039	99.430612	35.693336	not_obese
19108	Female	36.000000	1.500000	80.000000	35.555556	not_obese
19703	Female	40.000000	1.500000	80.000000	35.555556	not_obese
20271	Male	25.999942	1.703098	103.586342	35.712744	not_obese

Other Measures / Factors

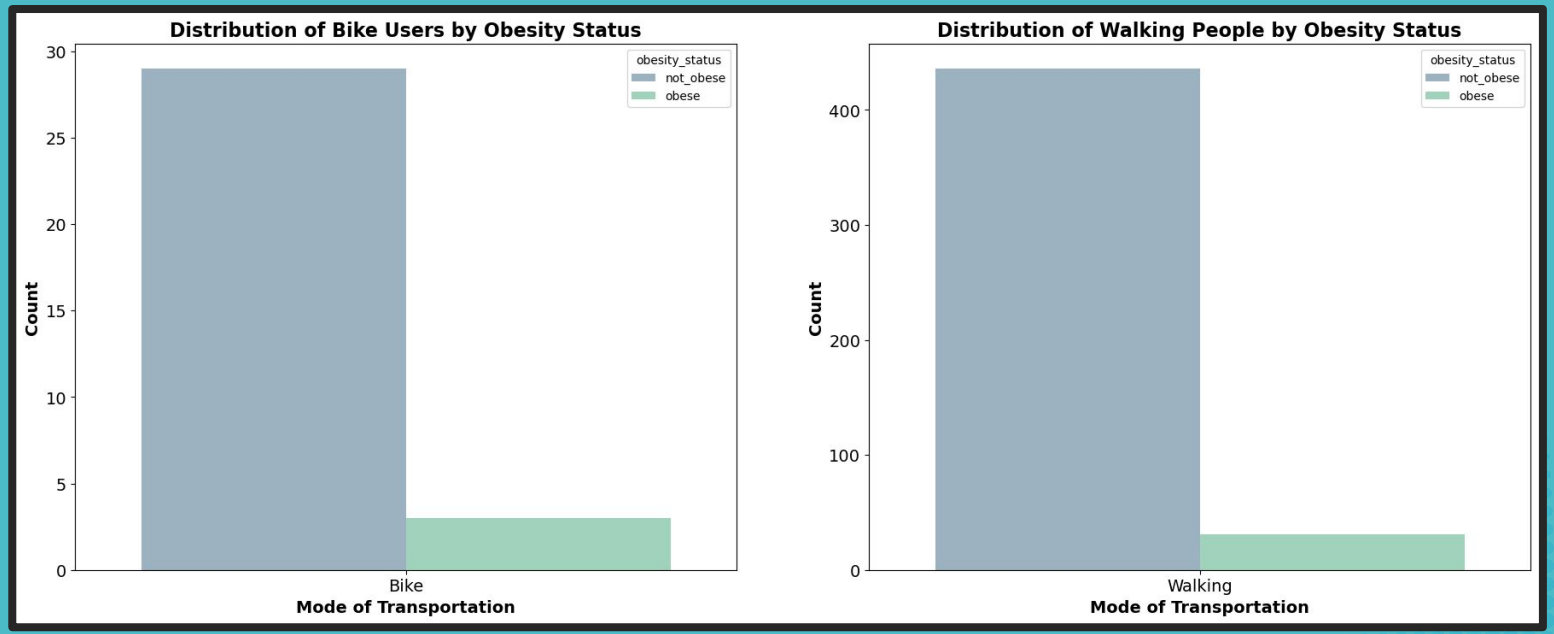
- Muscle Mass
- Fat mass
- Physical Activity
- Waist Circumference
- Age
- Ethnicity



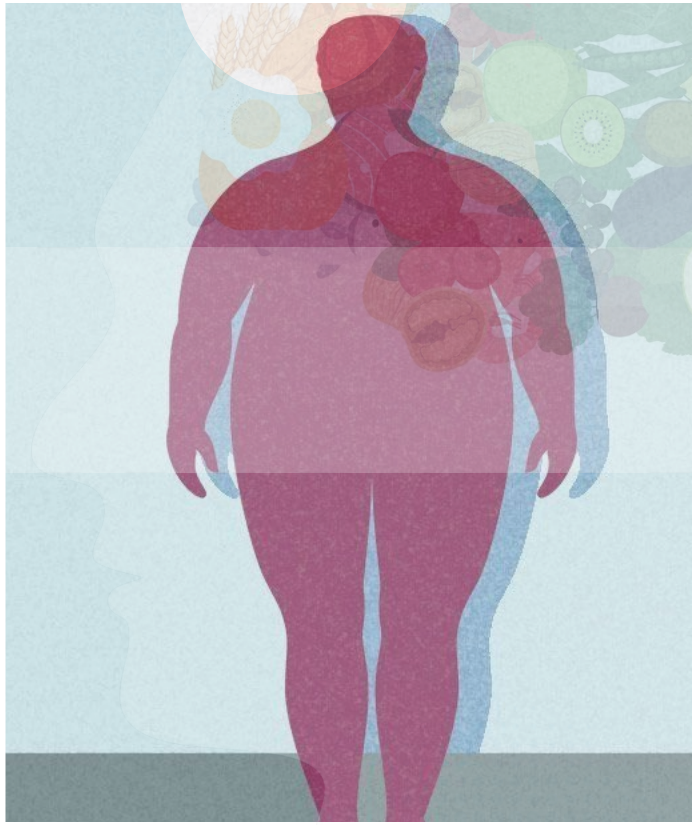
Distribution of Public Transportation Users and Automobile Users by Obesity Status



Distribution of Bike Users and Walking People by Obesity Status



Obesity Model



Features

17 Features:

- Gender
- Age
- Height
- Weight
- Physical Activity Frequency
- Mode of Transportation
- Number Of Meal In A Day
- Consumption of High Caloric Food

And other features.....

Base Models Comparison

Model	Processing Time	Train Score	Test Score	Sensitivity
SVM	2.3 s	0.9266	0.9121	0.9078
Random Forest	1.4 s	1.000	0.9227	0.9188
XGBoost	1.0 s	0.9922	0.9291	0.9265
Catboost	7.4 s	0.9726	0.9299	0.9268

Base Models Comparison

Model	Processing Time	Train Score	Test Score	Sensitivity
SVM	2.3 s	0.9266	0.9121	0.9078
Random Forest	1.4 s	1.000	0.9227	0.9188
XGBoost	1.0 s	0.9922	0.9291	0.9265
Catboost	7.4 s	0.9726	0.9299	**0.9268**

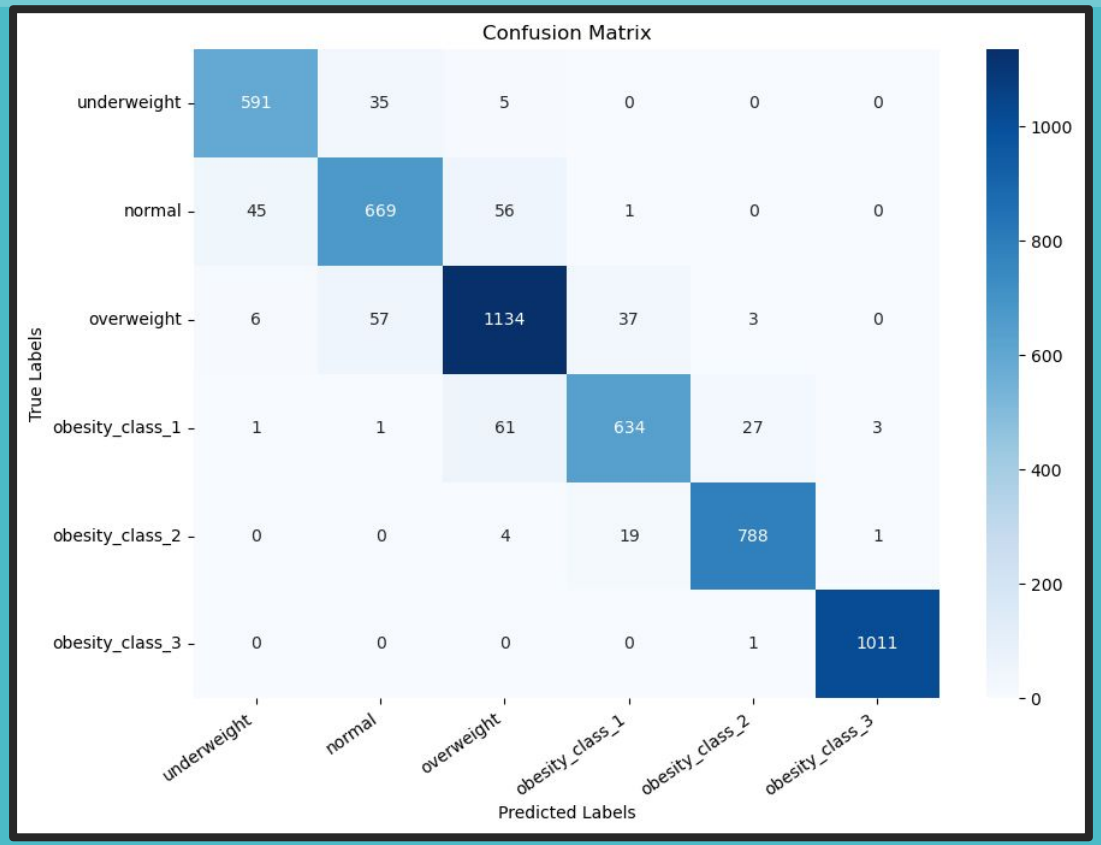
Hyperparameter Tuning

Model	Processing Time	Train Score	Test Score	Sensitivity
Catboost	7.4 s	0.9726	0.9299	0.9268
Catboost Tuned	**5.7 s**	**0.9541**	**0.9301**	**0.9271** ✅

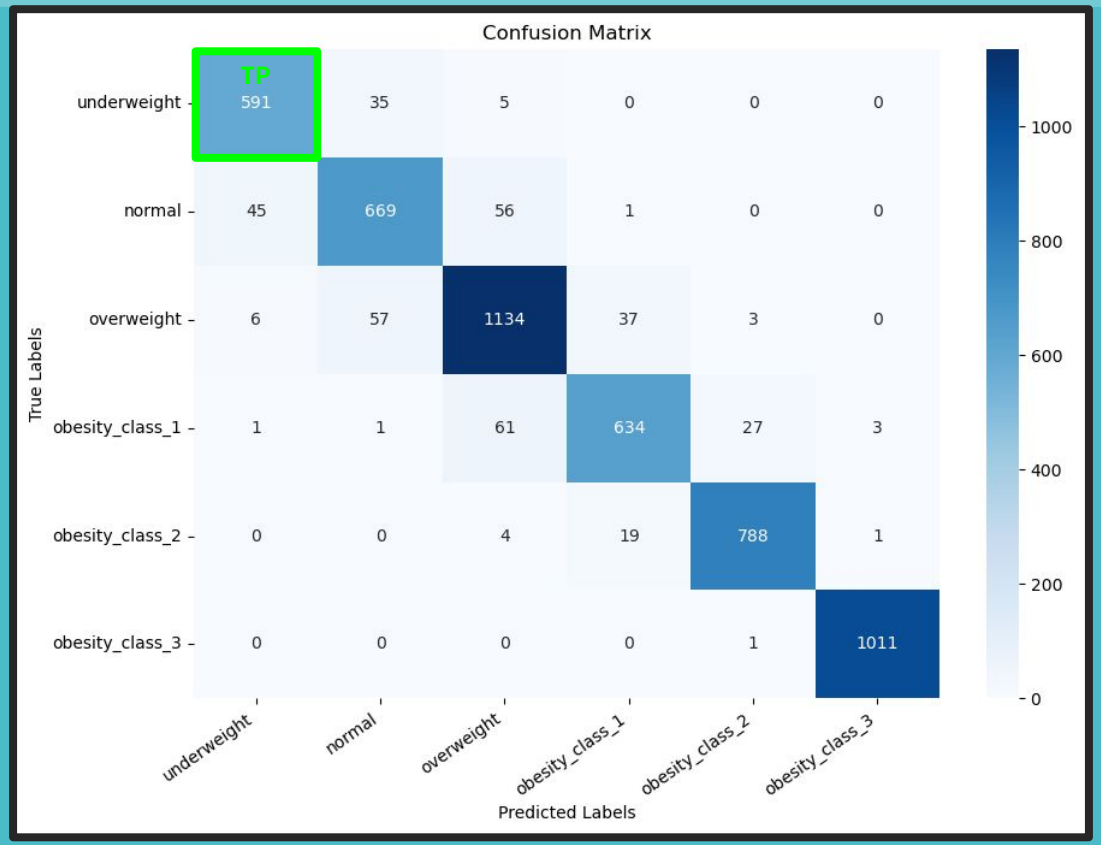
Hyperparameters:

- Learning Rate = 0.07
- Depth = 5
- Iterations = 1500

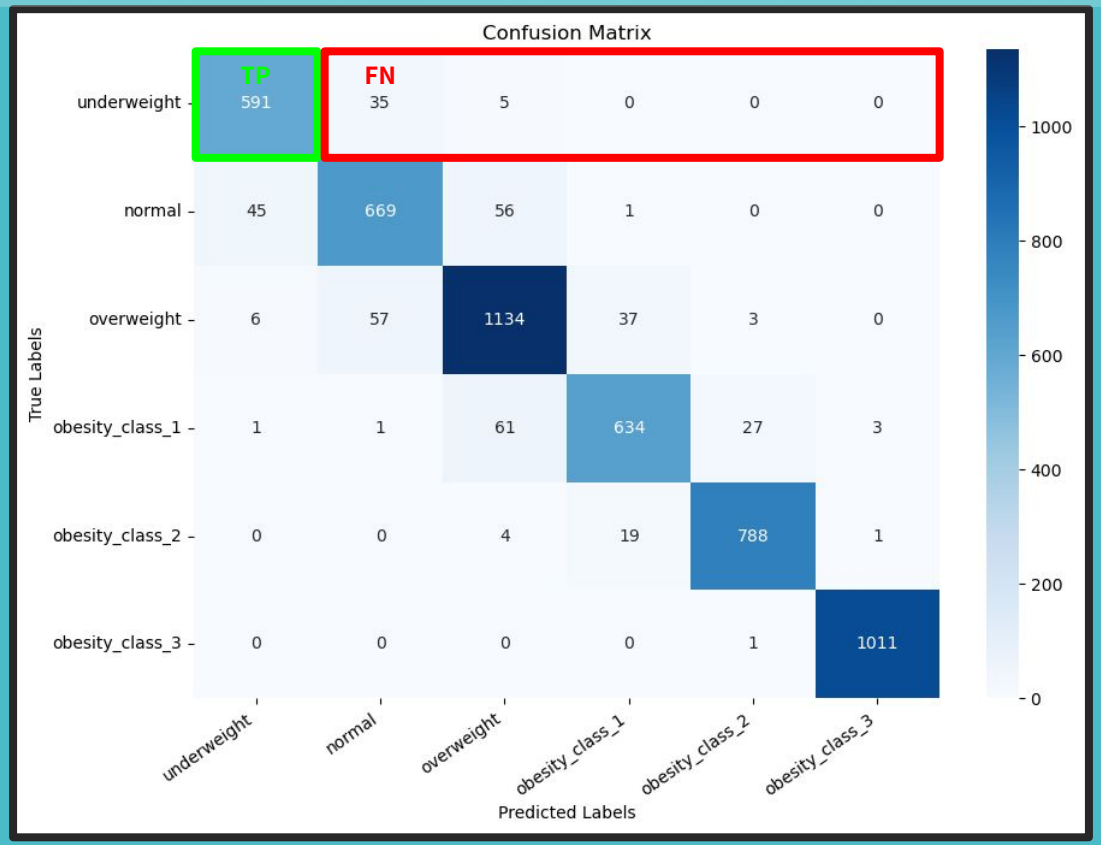
Tuned Confusion Matrix



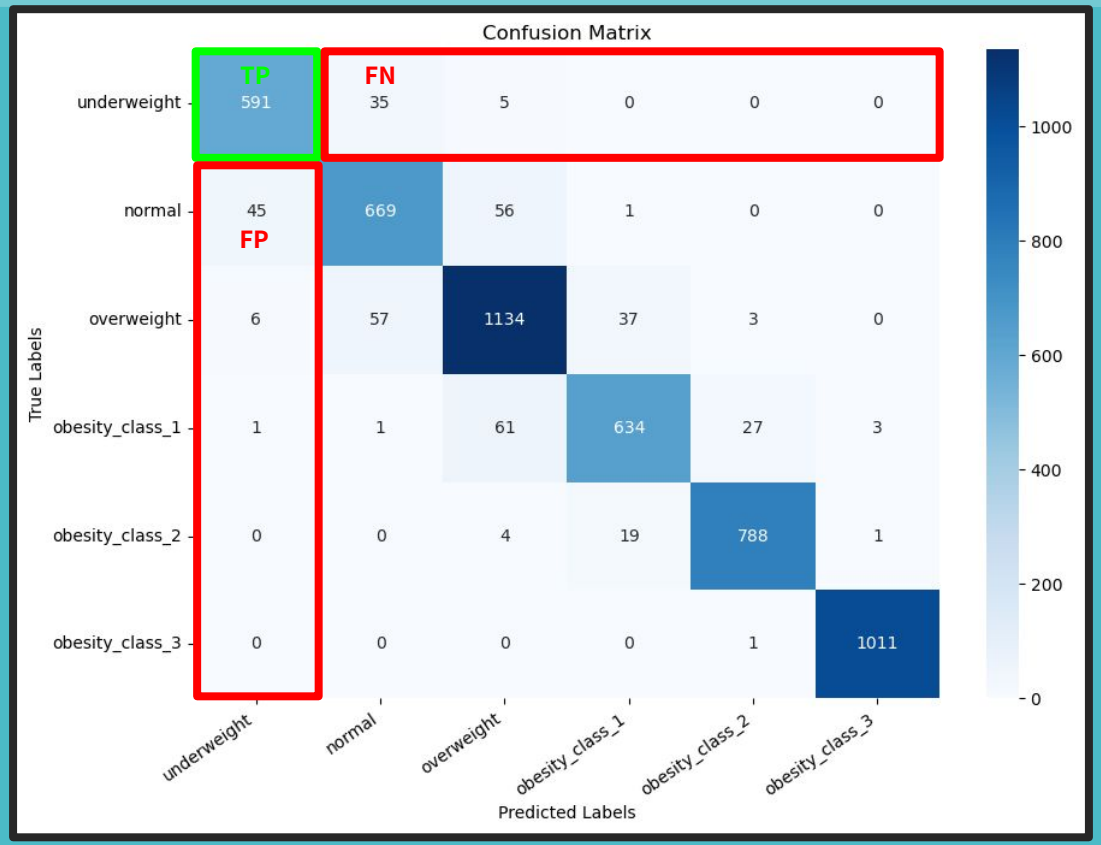
Tuned Confusion Matrix



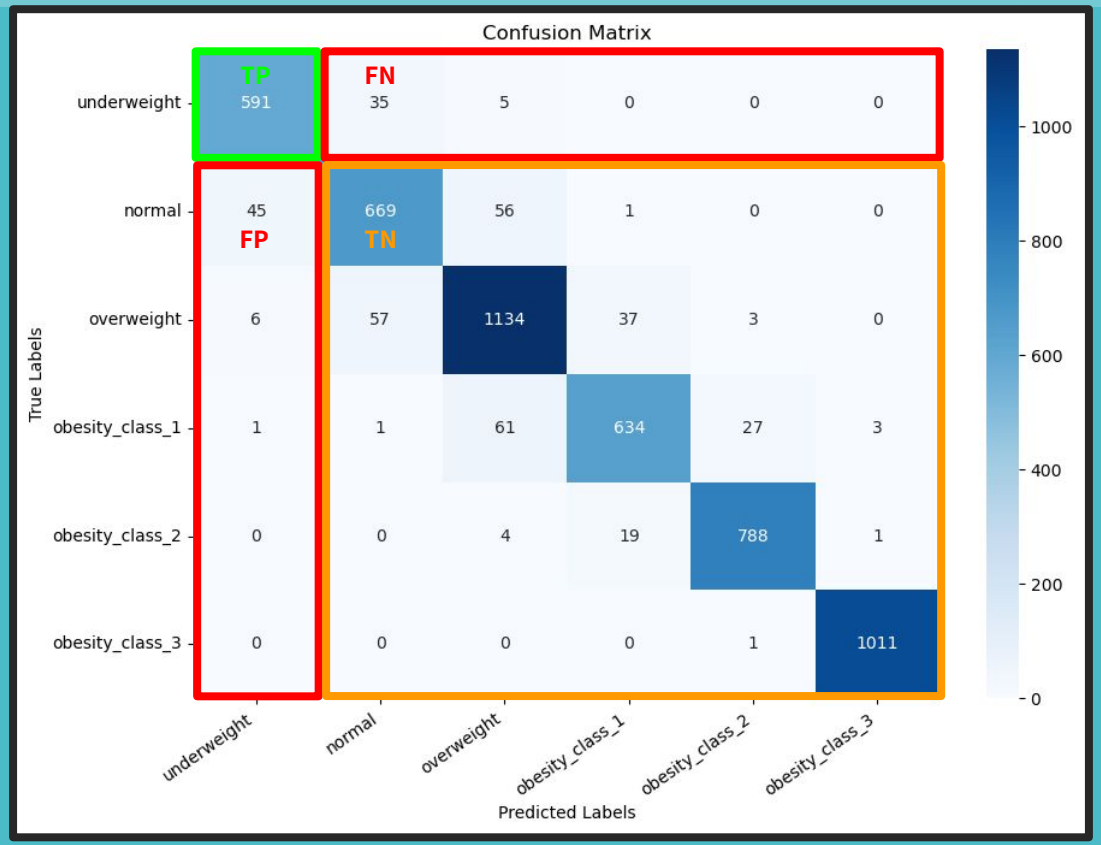
Tuned Confusion Matrix



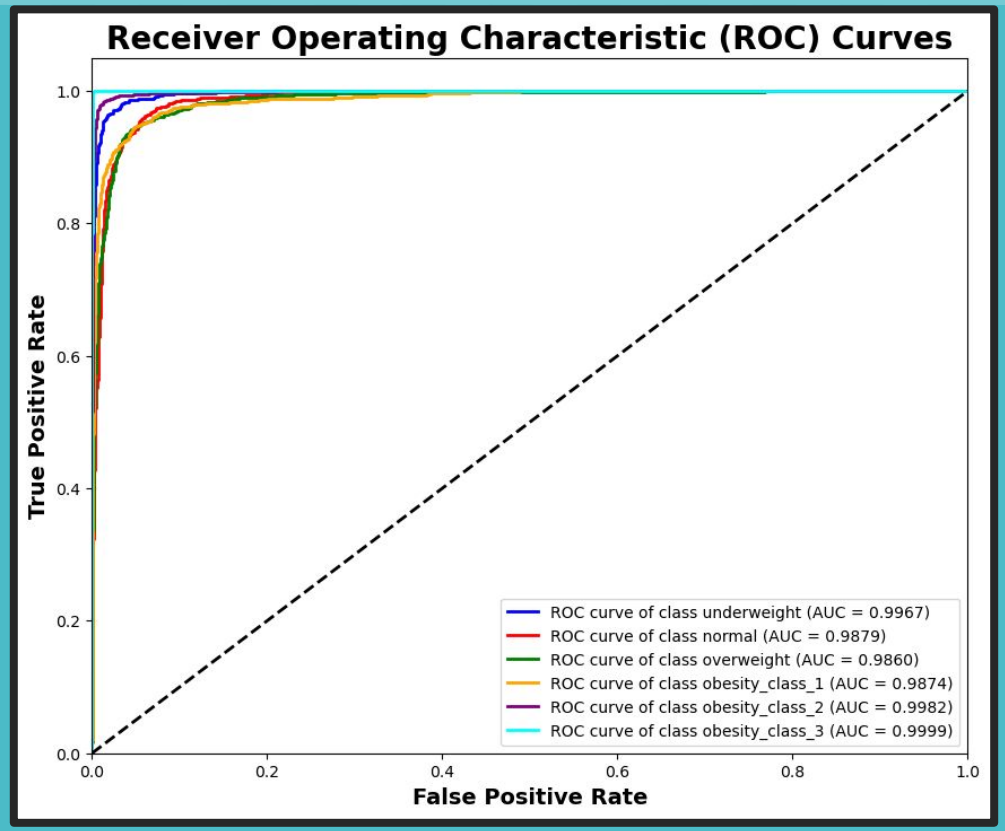
Tuned Confusion Matrix



Tuned Confusion Matrix



ROC AUC


Receiver Operating Characteristic (ROC) Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class underweight (AUC = 0.9967)
ROC curve of class normal (AUC = 0.9879)
ROC curve of class overweight (AUC = 0.9860)
ROC curve of class obesity_class_1 (AUC = 0.9874)
ROC curve of class obesity_class_2 (AUC = 0.9982)
ROC curve of class obesity_class_3 (AUC = 0.9999)

Bernard

Obesity Classification Model

Obesity Class 2

Stroke Detection Model

CT Scan Image Stroke Detector

George

Bernard

Obesity Classification Model

Obesity Class 2

Stroke Detection Model

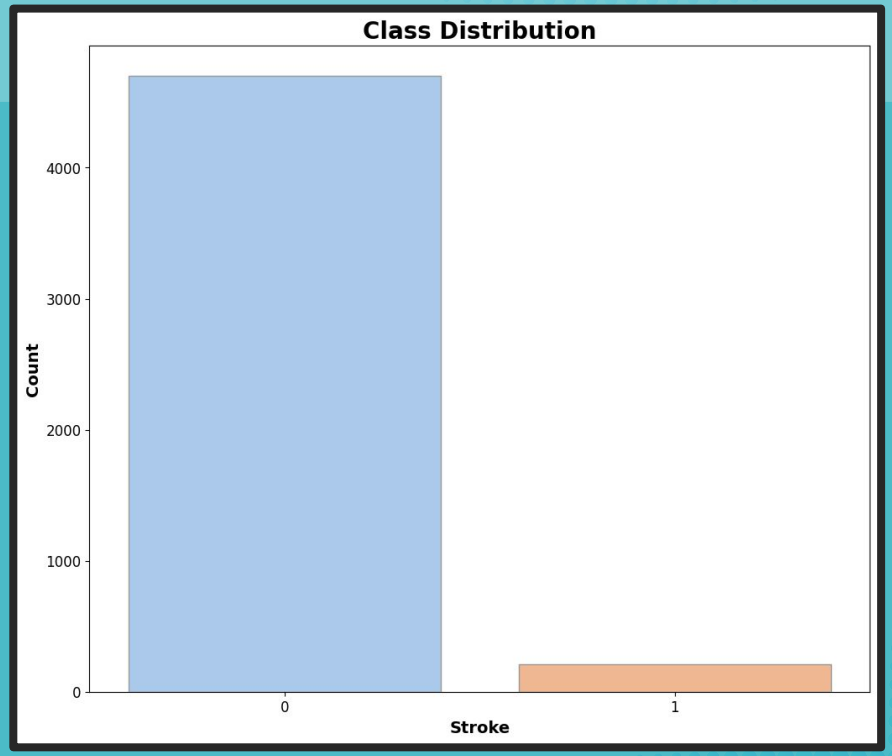
CT Scan Image Stroke Detector

George

Stroke Dataset

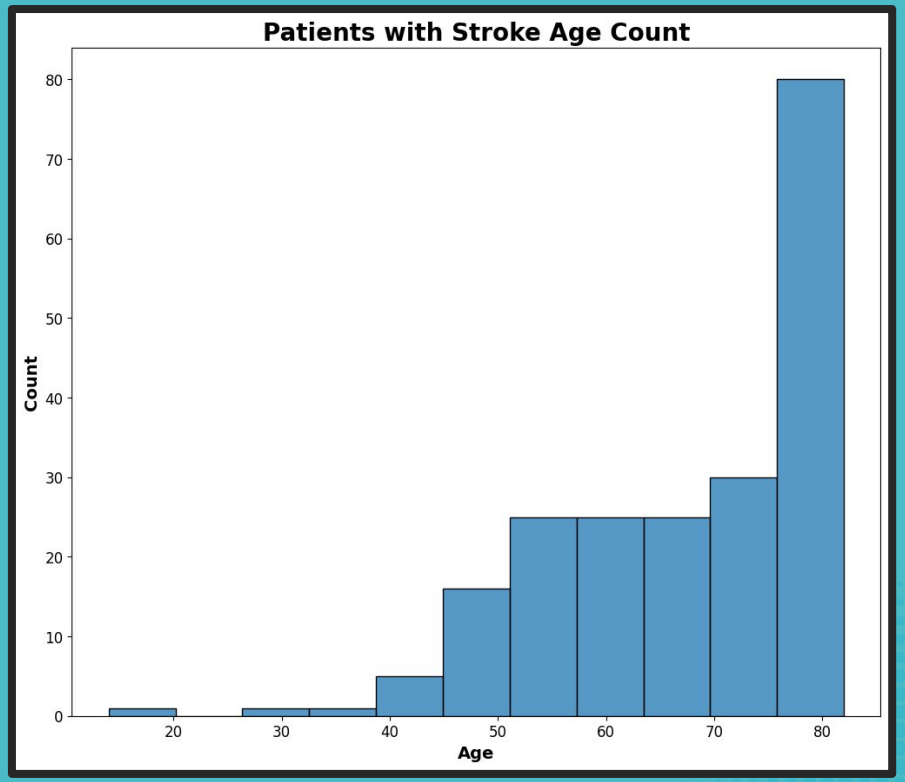
Stroke Dataset

- Dataset was taken from Kaggle.
- Dataset has around 5100 rows.
- 10 features
- Dataset is imbalance.



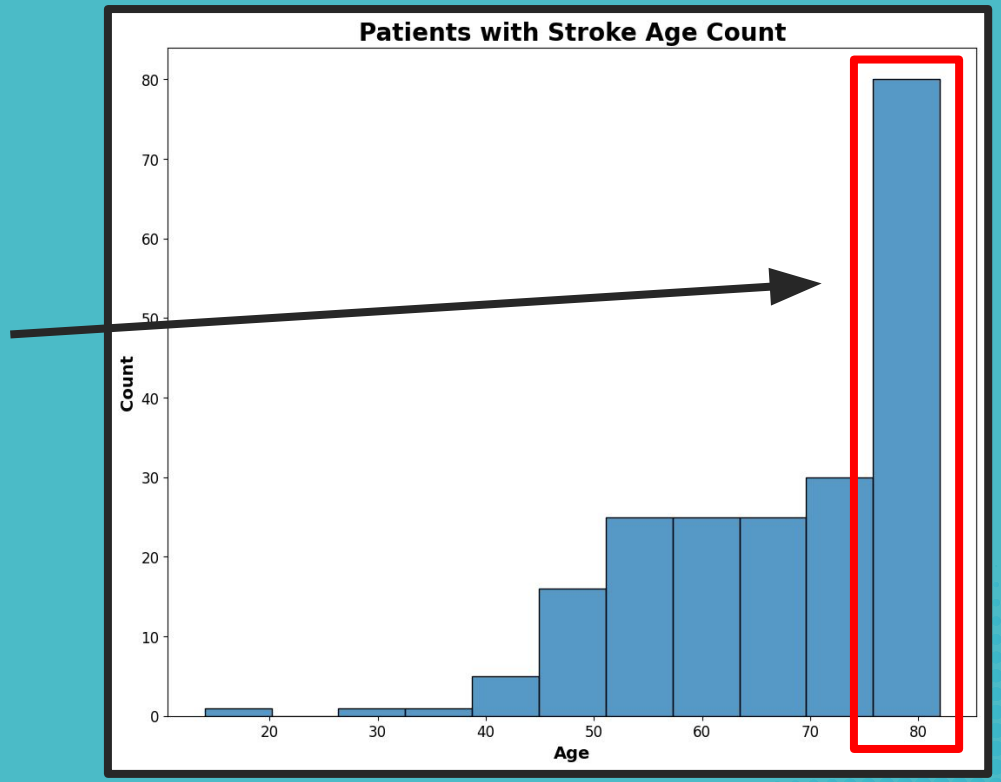
Analyzing the Data

- Highest number of stroke patients were around age 75 and above.

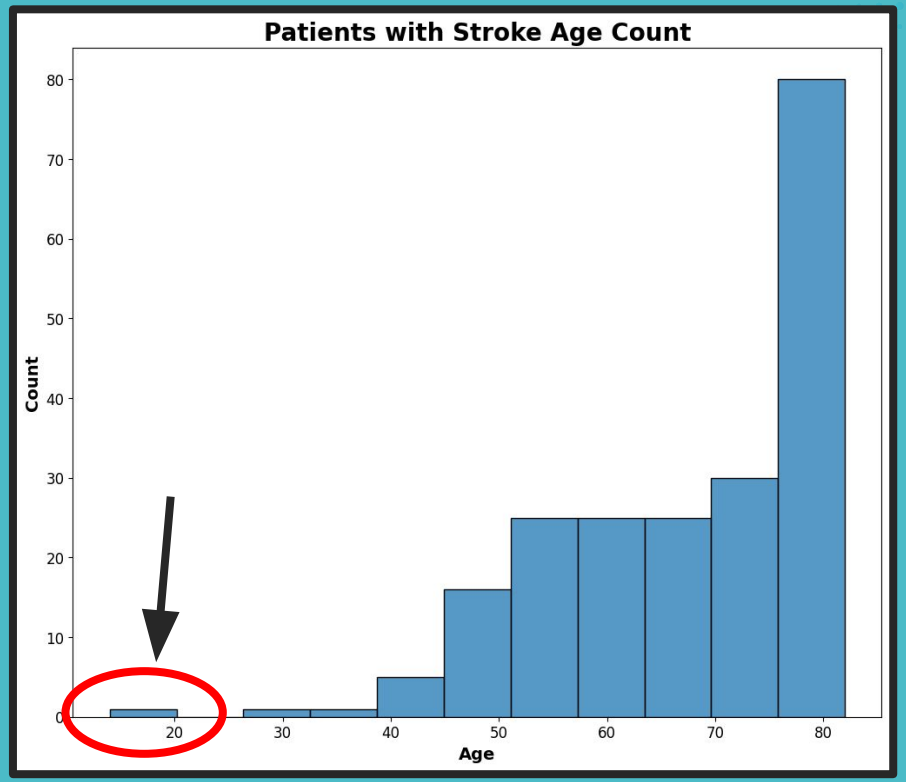


Analyzing the Data

- Highest number of stroke patients were around age 75 and above.

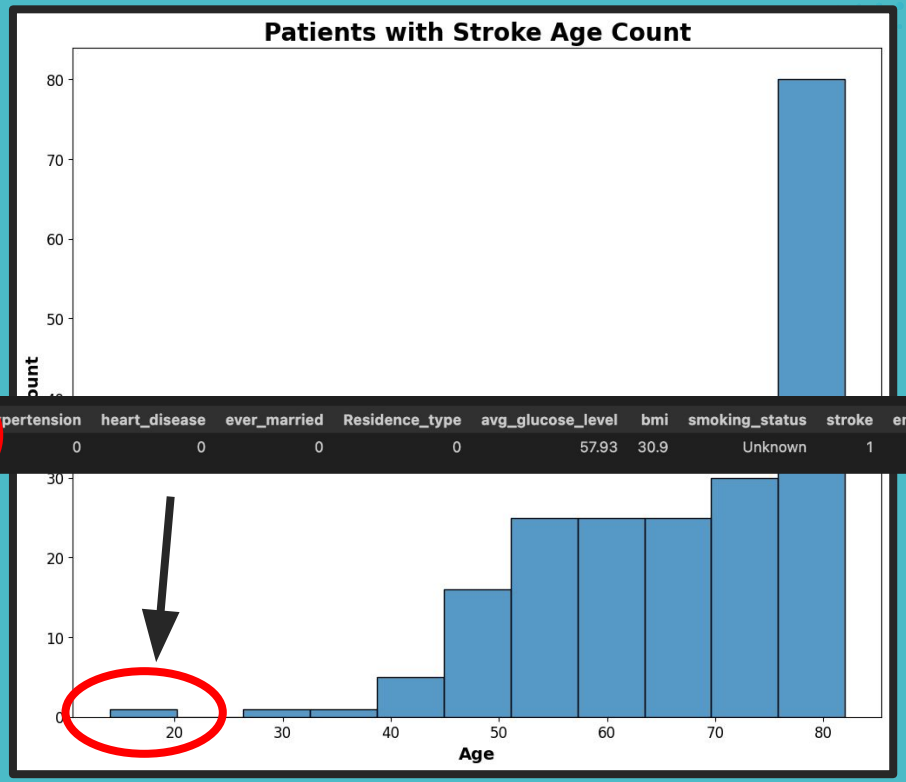


Analyzing the Data



Analyzing the Data

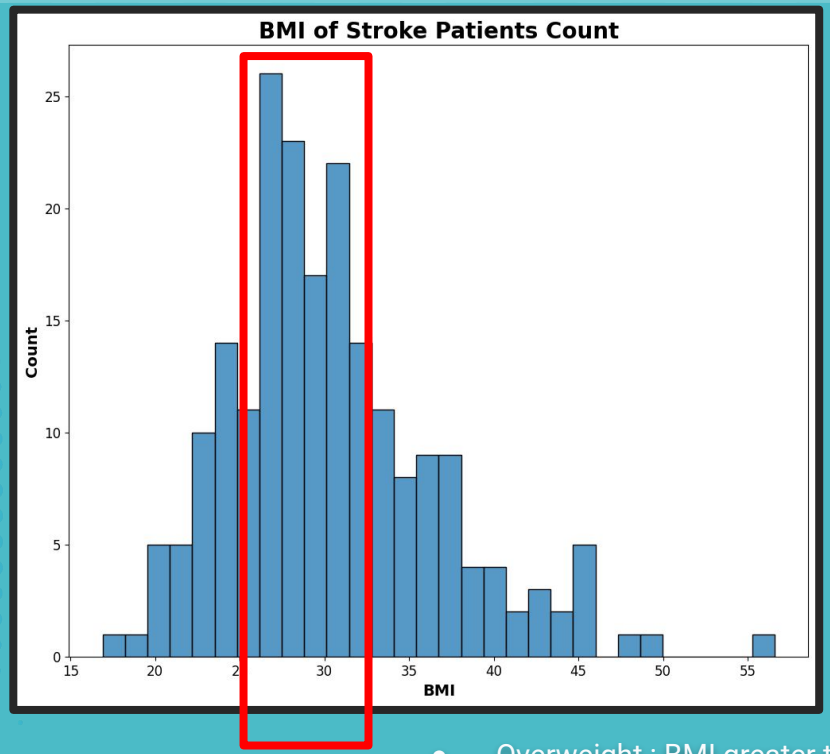
- Decided not to remove her from the dataset.

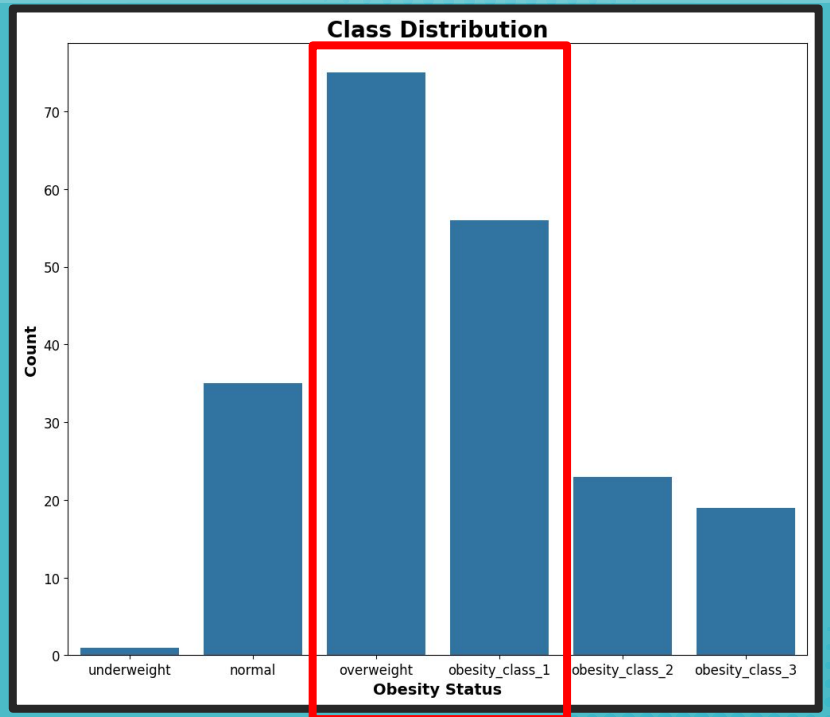


	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	employment_type	diabetic_status	obesity_status
245	1	14.0	0	0	0	0	57.93	30.9	Unknown	1	0	0	3

Obesity Class 1
BMI 30 to 34.9 kg/m^2

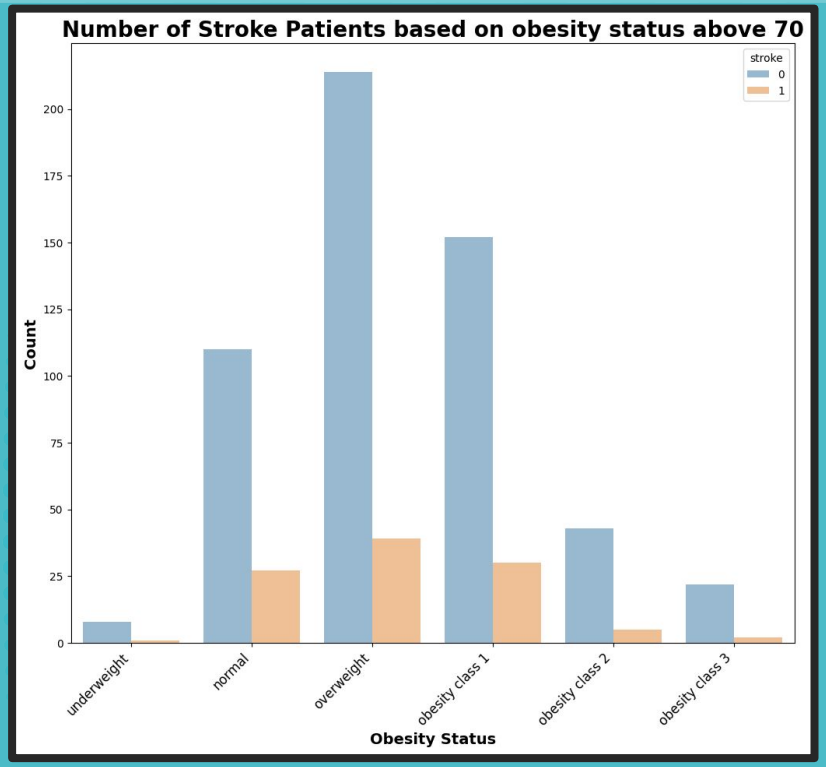
Analyzing the Data

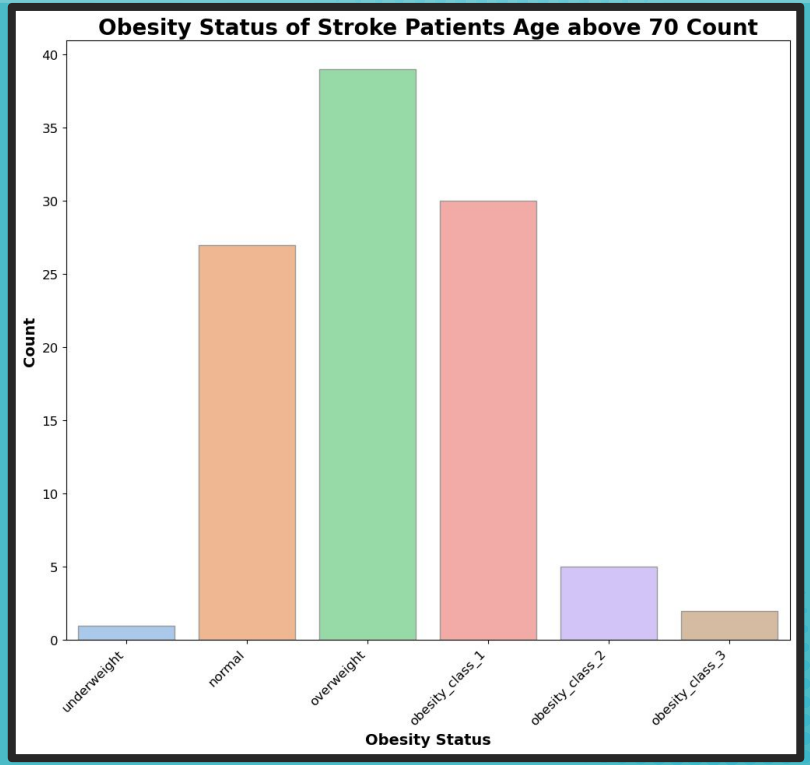




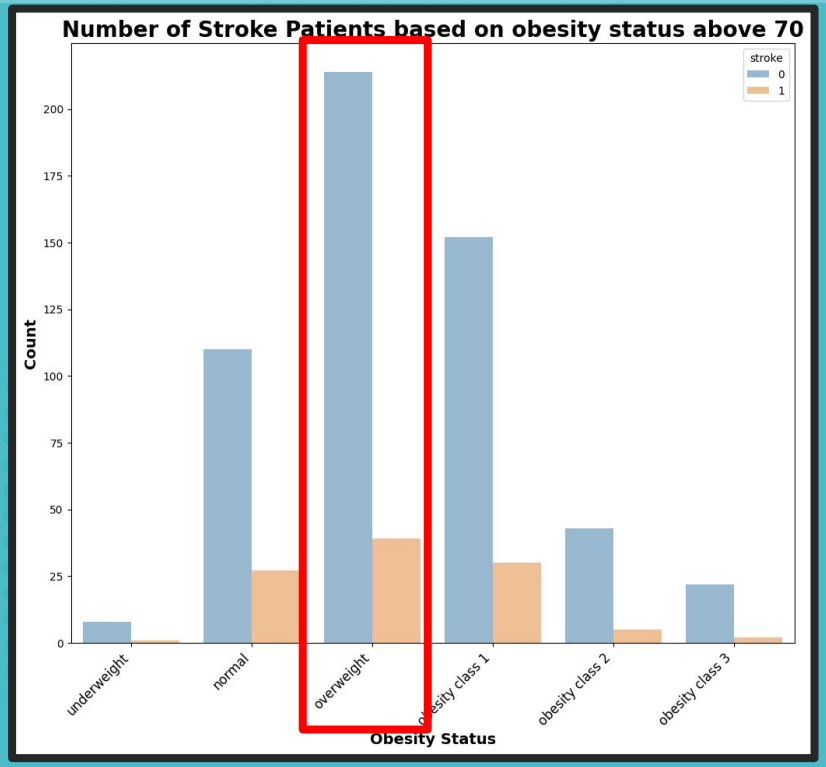
- Overweight : BMI greater than or equal to 25 to 29.9 kg/m^2
- Obesity class 1 : | BMI 30 to 34.9 kg/m^2

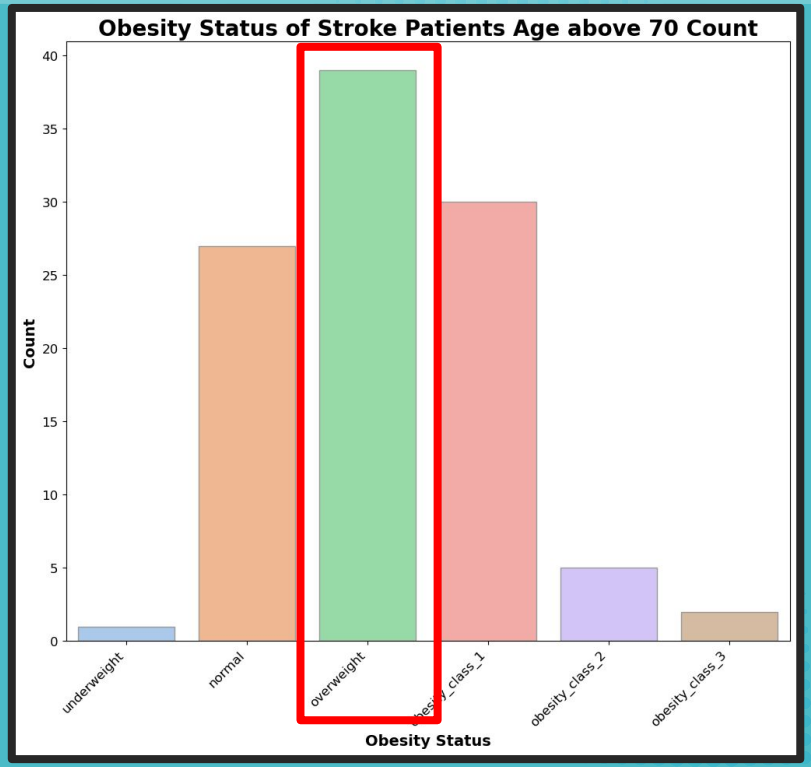
Analyzing the Data





Analyzing the Data





Analyzing the Data

- **Other Features**
  - Gender
  - Hypertension
  - Heart Disease
  - Average Glucose Level
  - Smoking Status
  - Employment Status
  - Diabetic Status
  - Obesity Status (Classified by Obesity Model)

Analyzing the Data

- **Other Features**
  - Gender
  - Hypertension
  - Heart Disease
  - Average Glucose Level
  - Smoking Status
  - Employment Status
  - Diabetic Status
  - Obesity Status (Classified by Obesity Model)

Obesity Classification Model

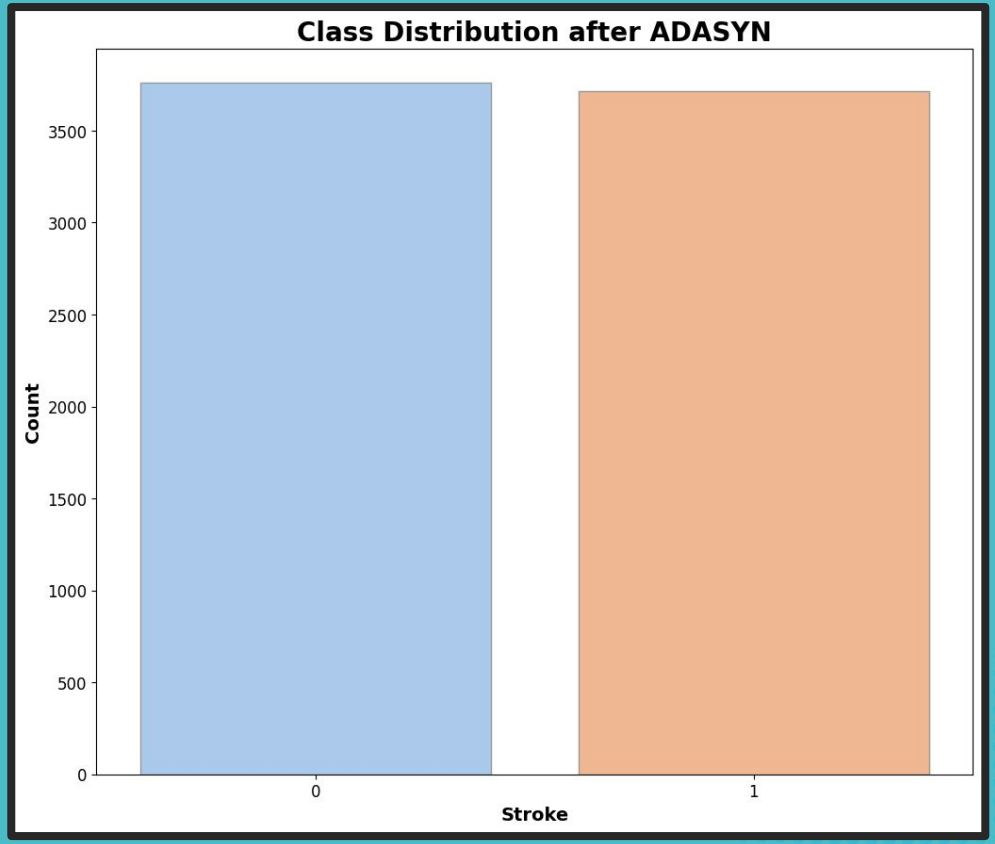
Rebalancing the Data (ADASYN)

Adaptive Synthetic Sampling (ADASYN)

Generate more synthetic examples in regions of the feature space where the density of minority examples is low, and fewer or none where the density is high.

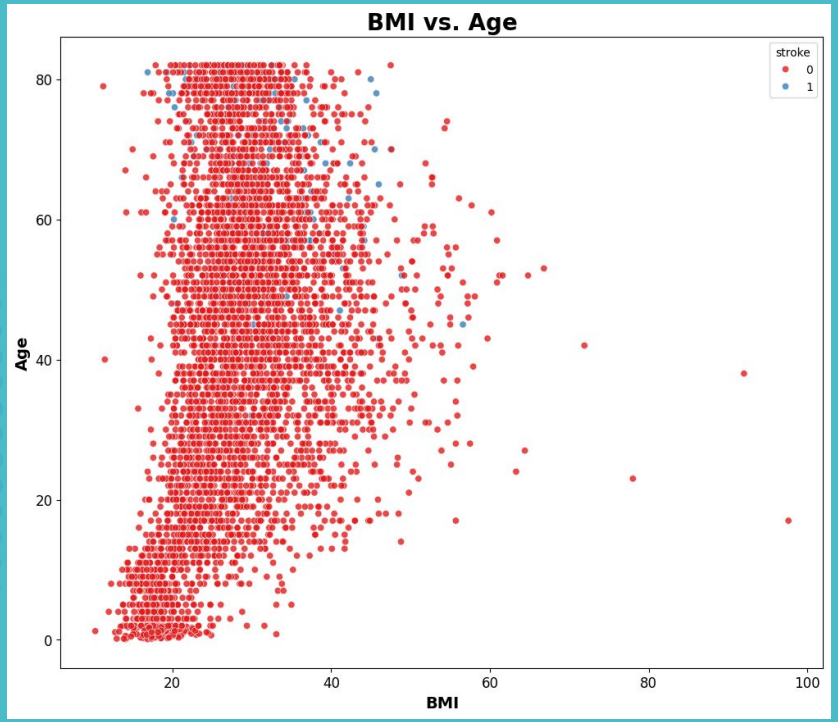
Why oversample the minority?

Machine learning algorithms tend to be biased towards the majority class because they aim to optimize overall accuracy. As a result, the minority class may be poorly represented in the model's predictions.

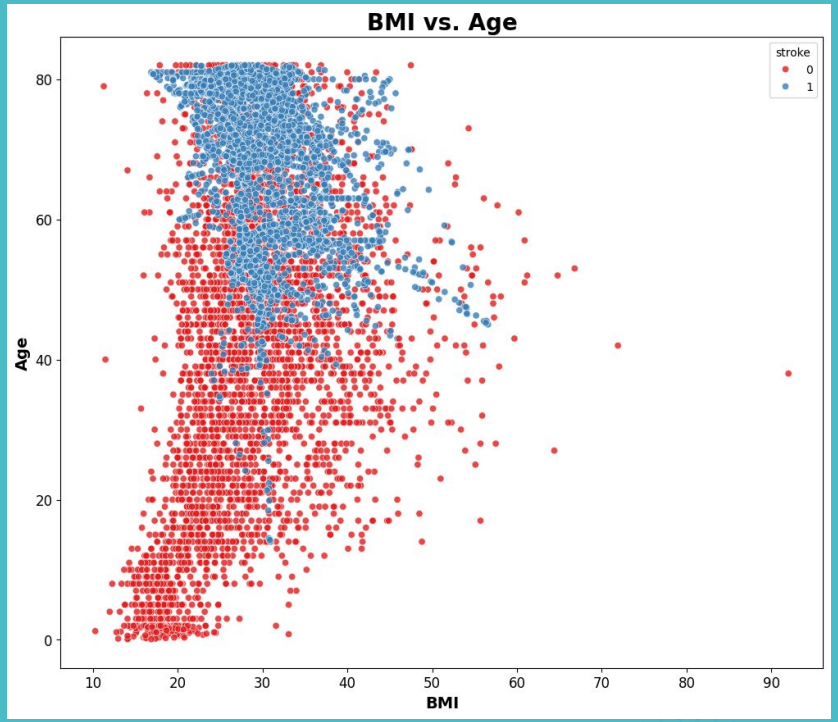


Rebalancing the Data (ADASYN)

Without Data Augmentation



With ADASYN oversampling



Stroke Detection Model

Base Model Comparison

Model	Processing Time	Train Score	Test Score	Sensitivity
Log Reg	21 ms	0.8120	0.7760	0.5714
KNN	26 ms	0.9317	0.8004	0.3095
Decision Tree	44 ms	1.0000	0.8625	0.1429
Bagging	199 ms	0.9968	0.8819	0.1667
Random Forest	723 ms	1.0000	0.8890	0.2143
AdaBoost	237 ms	0.8426	0.7709	0.4762
SVM	1.01 s	0.8616	0.7984	0.3571
XGBoost	135 ms	0.9941	0.9002	0.1667

Base Model Comparison

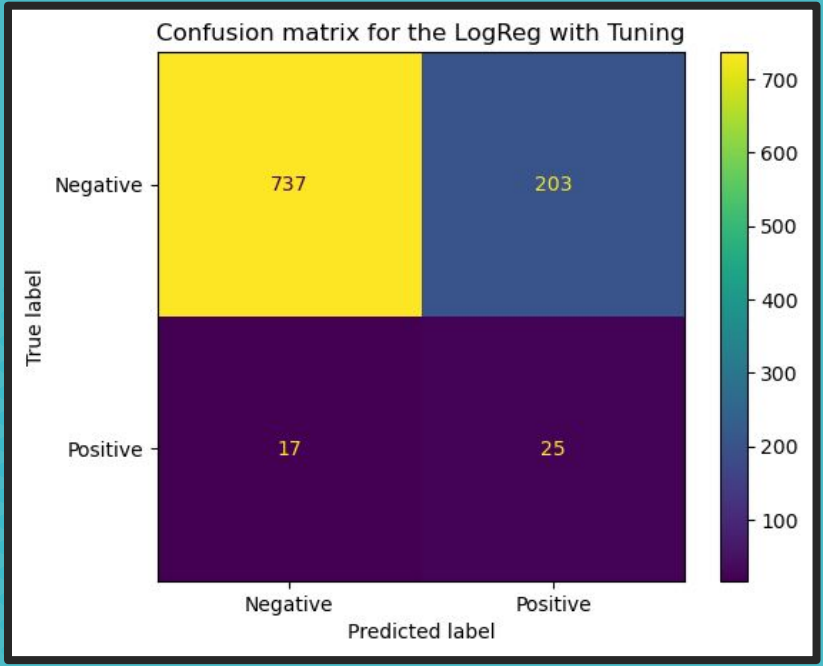
Model	Processing Time	Train Score	Test Score	Sensitivity
Log Reg	21 ms	0.8120	0.7760	0.5714
KNN	26 ms	0.9317	0.8004	0.3095
Decision Tree	44 ms	1.0000	0.8625	0.1429
Bagging	199 ms	0.9968	0.8819	0.1667
Random Forest	723 ms	1.0000	0.8890	0.2143
AdaBoost	237 ms	0.8426	0.7709	0.4762
SVM	1.01 s	0.8616	0.7984	0.3571
XGBoost	135 ms	0.9941	0.9002	0.1667

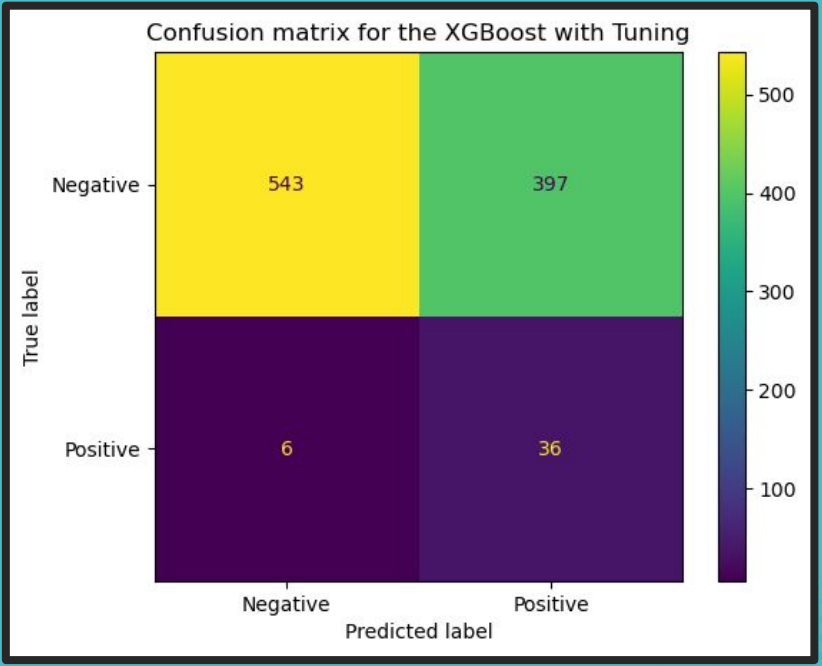
Hyperparameter Tuning

- Despite the dataset being balanced, there's only so much it can provide the model to train for it to be accurate.
- Decided to prioritize **Sensitivity**.
- Crucial for conditions like strokes, where **failing** to predict an **actual positive case** can have serious consequences. High sensitivity means the model is **effective** at **catching most positive cases**.

Model	Processing Time	Train Score	Test Score	Sensitivity
Log Reg	21 ms	0.8120	0.7760	0.5714
Log Reg Tuned	2.86 s	0.8107	0.7759	0.5952
XGBoost	135 ms	0.9941	0.9002	0.1667
XGBoost Tuned	2.9 s	0.8107	0.5896	0.8571

Tuned Confusion Matrix

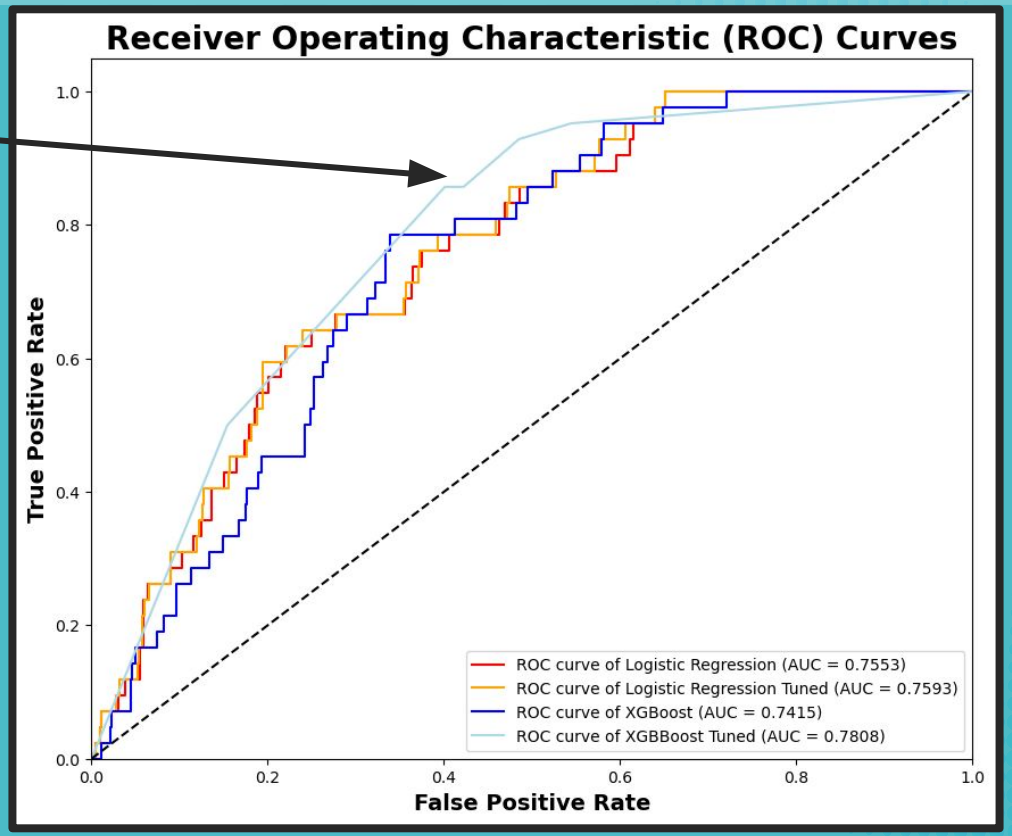




Sensitivity (recall): The proportion of true positives predictions out of all actual positive instances.

ROC AUC

Best performing Model **(XGBoost)**



Hyperparameter Tuning

Model	Processing Time	Train Score	Test Score	Sensitivity
Log Reg	21 ms	0.8120	0.7760	0.5714
Log Reg Tuned	2.86 s	0.8107	0.7759	0.5952
XGBoost	135 ms	0.9941	0.9002	0.1667
XGBoost Tuned	2.9 s	0.8107	0.5896	0.8571 ✅

Sensitivity (recall): The proportion of true positives predictions out of all actual positive instances.

Bernard

Obesity Classification Model

Obesity Class 2

Stroke Detection Model

Stroke Probability : 76.5%

CT Scan Image Stroke Detector

George

Bernard

Obesity Classification Model

Obesity Class 2

Stroke Detection Model

Stroke Probability : 76.5%

CT Scan Image Stroke Detector

George

04
CT Scan
Stroke
Detector

CT Scan images Model

Model used for CT scan images

Convolutional Neural Network (CNN)

It's a type of deep learning architecture well-suited for image classification tasks

How does CNN works?

Raw CT scan images

- Obtained from Kaggle
- Labelled with Stroke or non-stroke

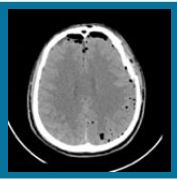


Image pre-processing

- Steps like resizing and augmentation

Model training

- Train based on processed images
- Learn based on the patterns in the image

Output as a classifier for stroke / nonstroke

CT scan images

- Images were taken from Kaggle.
- Training set have around 1,843 images
- 41% is stroke and 59% is non-stroke

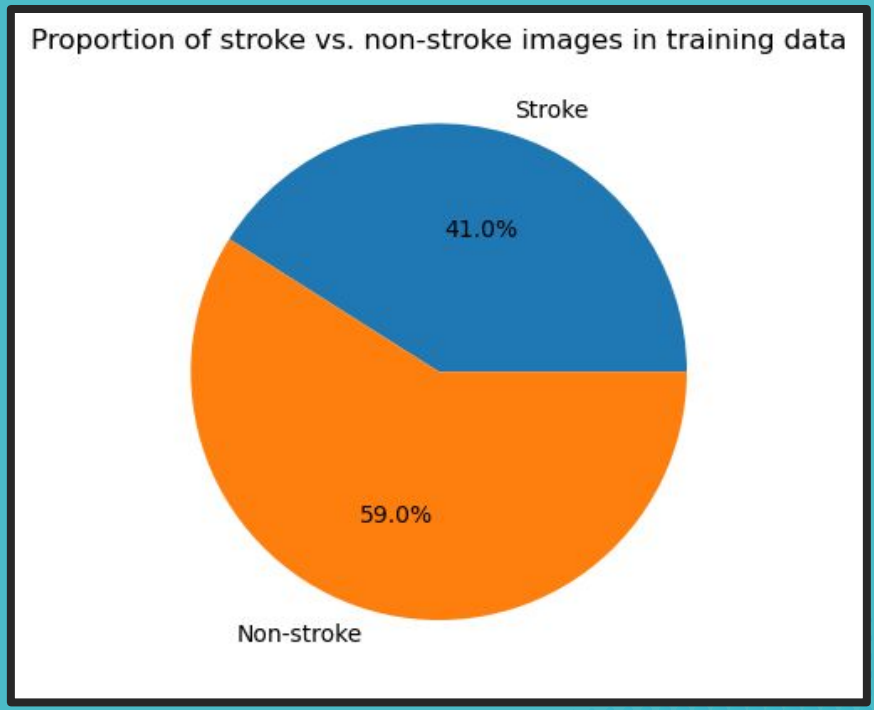
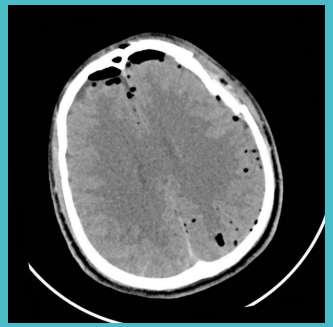
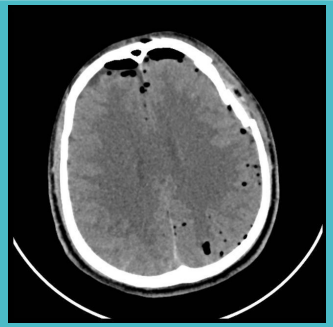


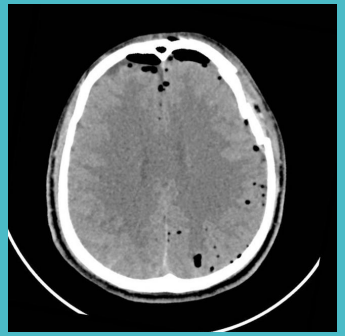
Image pre-processing (Augmentation)



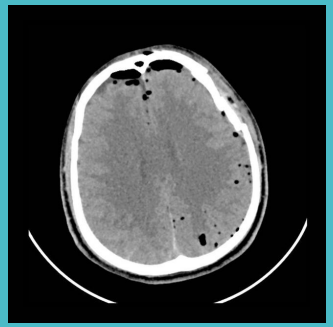
Rotate -10°



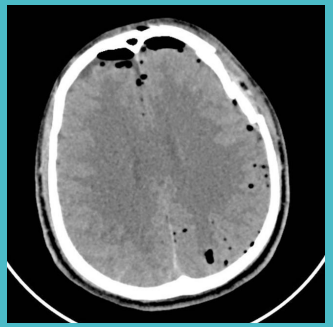
Original



Rotate 10°



Scale -10%

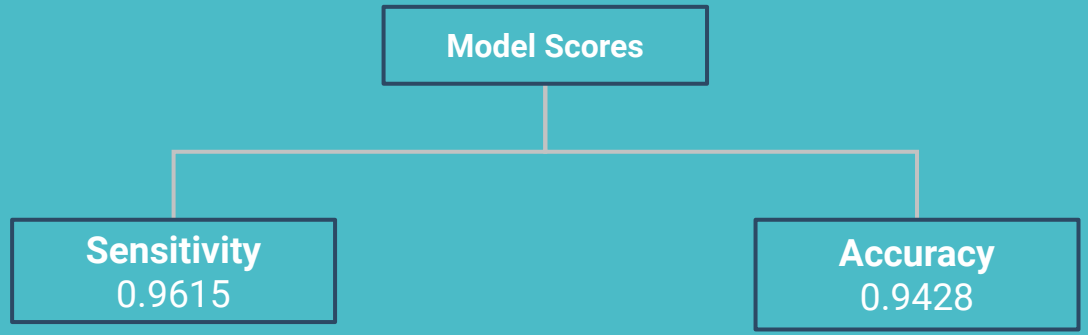


Scale +10%

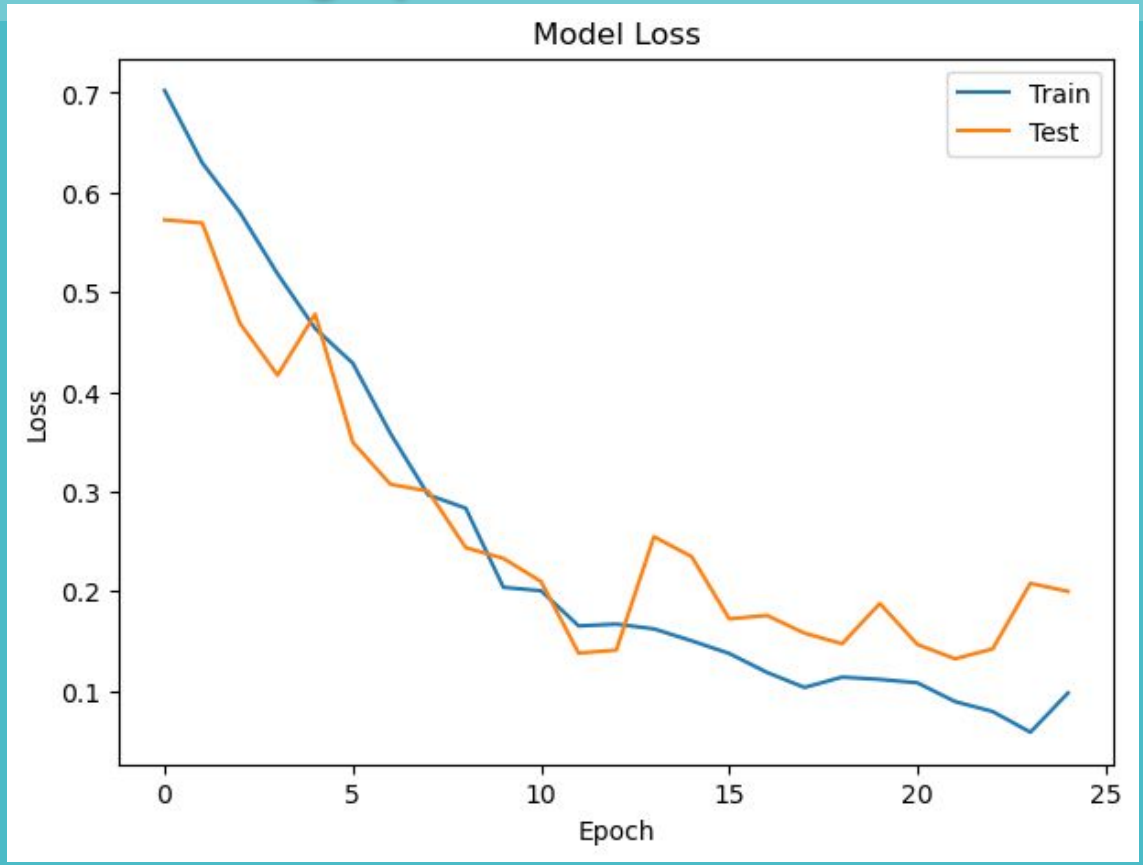
Model Training

Model Architecture	Sequential
No. of layers	9
Total parameters	7,466,177
Optimizer	Adam (Learning rate 0.001)
Loss function	Binary Cross-entropy
Metrics	Sensitivity

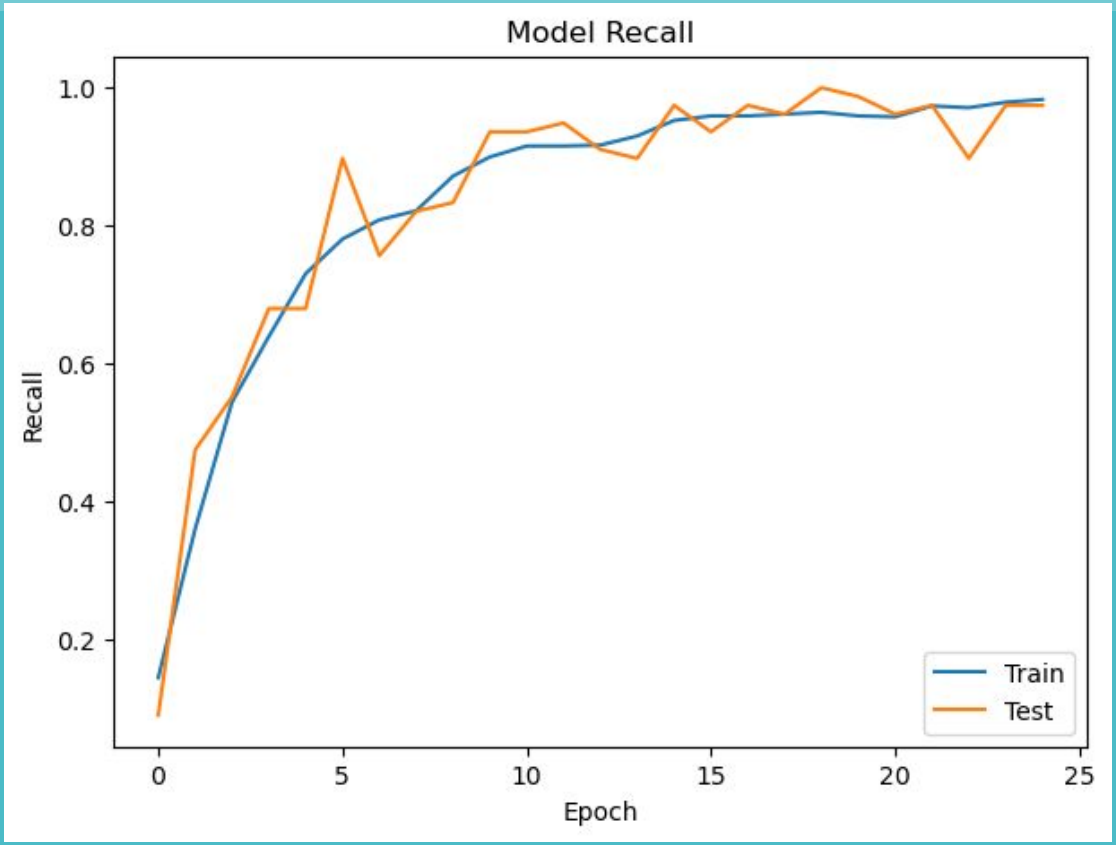
Model Score



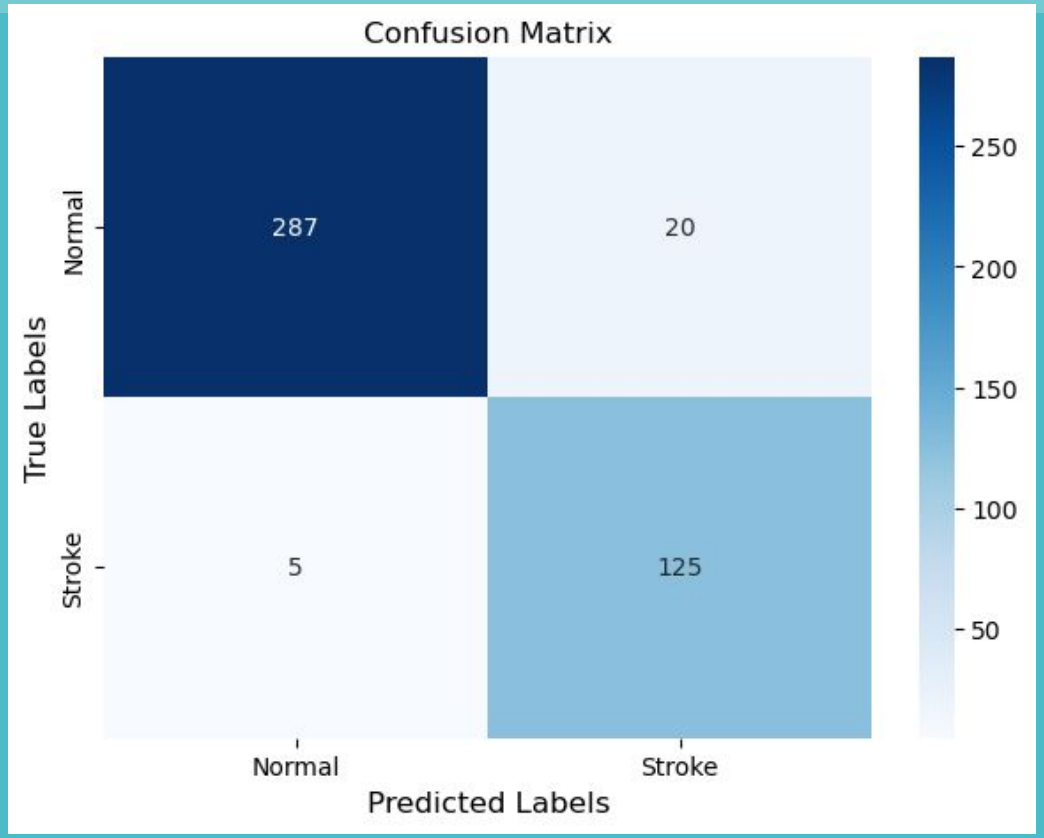
Loss over training epochs



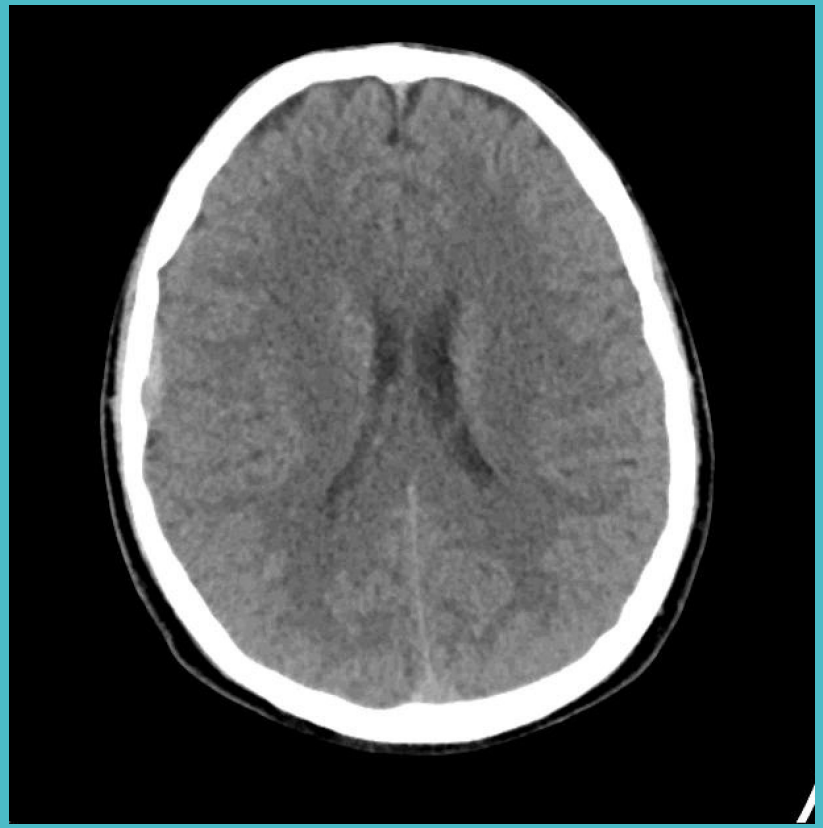
Recall over training epochs



Confusion Matrix

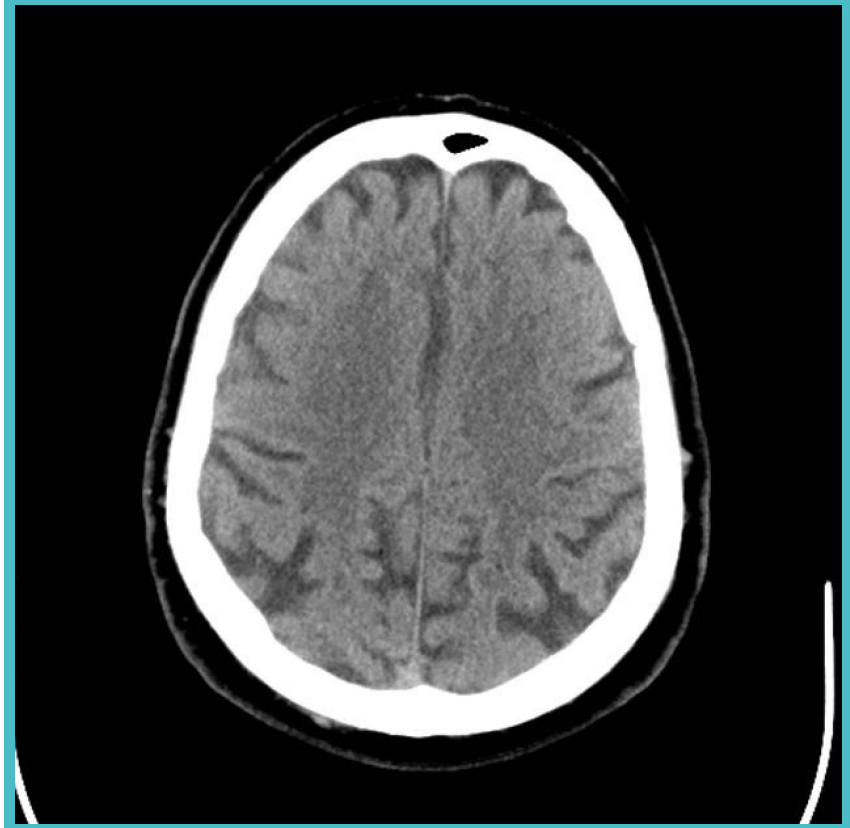


Test 1 - Risk of Stroke



Model prediction	Actual	
98.41%	*Stroke*	✅

Test 2 - Risk of Stroke



Model prediction	Actual	
0.04%	*Non-stroke*	✅

Bernard

Obesity Classification Model

Obesity Class 2

Stroke Detection Model

Stroke Probability : **76.5%**

CT Scan Image Stroke Detector

Stroke Probability : **98.41%**

George

Cost Benefits Analysis

Implementation Cost

Development cost

- Data collection and cleaning.
- Personnel costs for data scientists and engineers.
- Computational resources for training.

Implementation cost

- Integration into existing healthcare systems
- Cost of obtaining new CT scan images for risk prediction.
- Training for staff using the model.

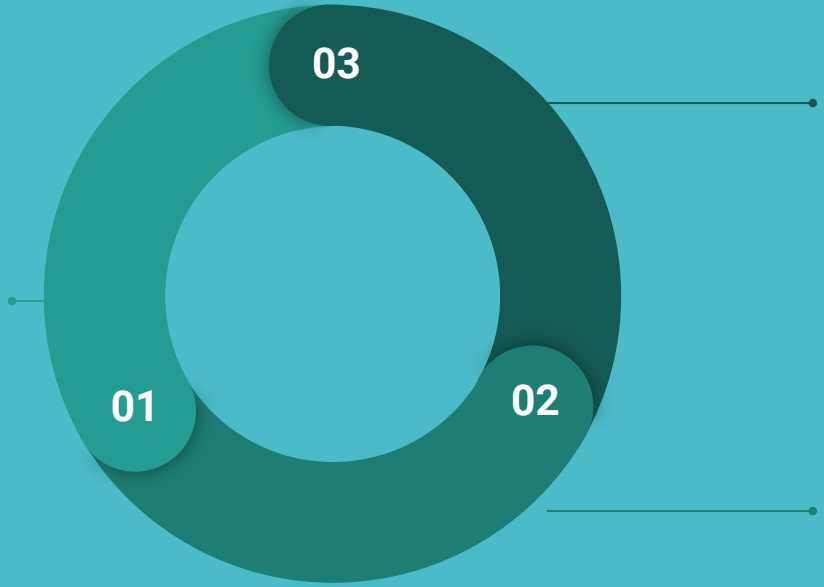
Maintenance

- Regular model updates to ensure accuracy with new data.
- Technical support and troubleshooting.

Hospital Implementation Benefit

Early detection benefits for patients

Potential cost savings by treating stroke early, for example, minimization of long term care.



Saving costs associated with doctor's time.

The app will be able predict the risk of getting stroke without requiring a doctor's diagnosis, allowing nurses or clinical assistants to use it and avoid occupying doctors' time.

Save on operational cost for hospital healthcare facilities

Including equipment cost, manpower cost and etc.

Patient Cost & Benefits

Without App Implementation

- Hospital cost $5,000 - $15,000
- Rehabilitation cost $6,000 - $12,000 for 6 months
- If the patient gets stroke, he will lose his income for 6 months or longer, $30,000 or more

With App Implementation

- Save hospitalization and rehabilitation cost.
- Cholesterol and high blood pressure medicine $1,000 - $1,200 a year.
- Will be able to continue working and have income.

Bernard

Streamlit

George

Limitation

Limitation of the models

- **Limited Training Data**
  Due to limited data, the model might be overly specific to the training data (overfitting) or lack the complexity to capture the data's patterns (underfitting), hindering its ability to generalize well to unseen examples.

- **Black box models**
  Complex models like neural networks can be highly effective but often lack clear explanations for their decisions. This can make them unsuitable for domains where understanding the reasoning behind a prediction is crucial.

Recommendation

Recommendation

We need to find more training data to improve our models. Here are some ways to address dataset limitations in specific areas:

- Obesity: For the obesity dataset, we could find more data, for example waist circumference and fat mass. This would make the model even more accurate.
- Stroke: The current stroke dataset has a minority of stroke-positive examples. Finding a bigger dataset with a more balanced distribution could improve the model's performance.
- CT Image scan: We can improve the accuracy of our model by finding more CT image scans for training

Conclusion

Recap: Problem Statement

The healthcare system faces a challenge in quickly identifying patients at risk of stroke upon arrival at the emergency department (A&E), despite standardized triage protocols. Swiftly discerning stroke risk based on initial health particulars such as **obesity** is difficult, leading to delays in referrals to specialists for preventive measures.

How might we help A&E identify patients at risk of stroke easier?

George

Bernard

- George will now be able to quickly diagnose potential stroke patients like Bernard at A&E.
- Get a quick general prediction from the CT scan of the brain.

George

Bernard

Obesity Classification Model

Stroke Detection Model

CT Scan Image Stroke Detector

Thank you

Q&A