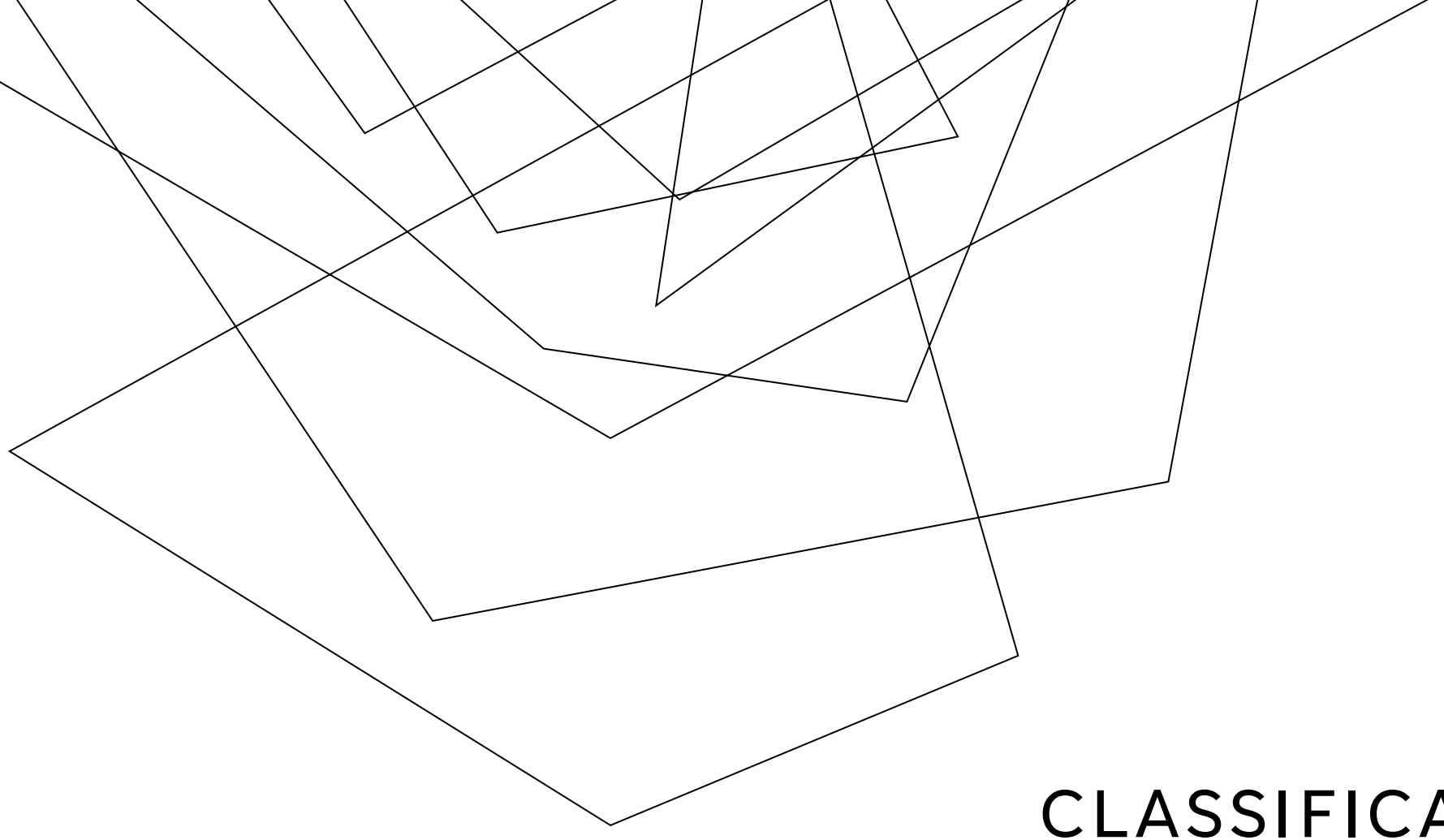


# CLASSIFICATION OF OBESITY LEVELS

Sunku Bhanu Kedhaar Nath

# AGENDA

Introduction

Data Exploration and Preprocessing

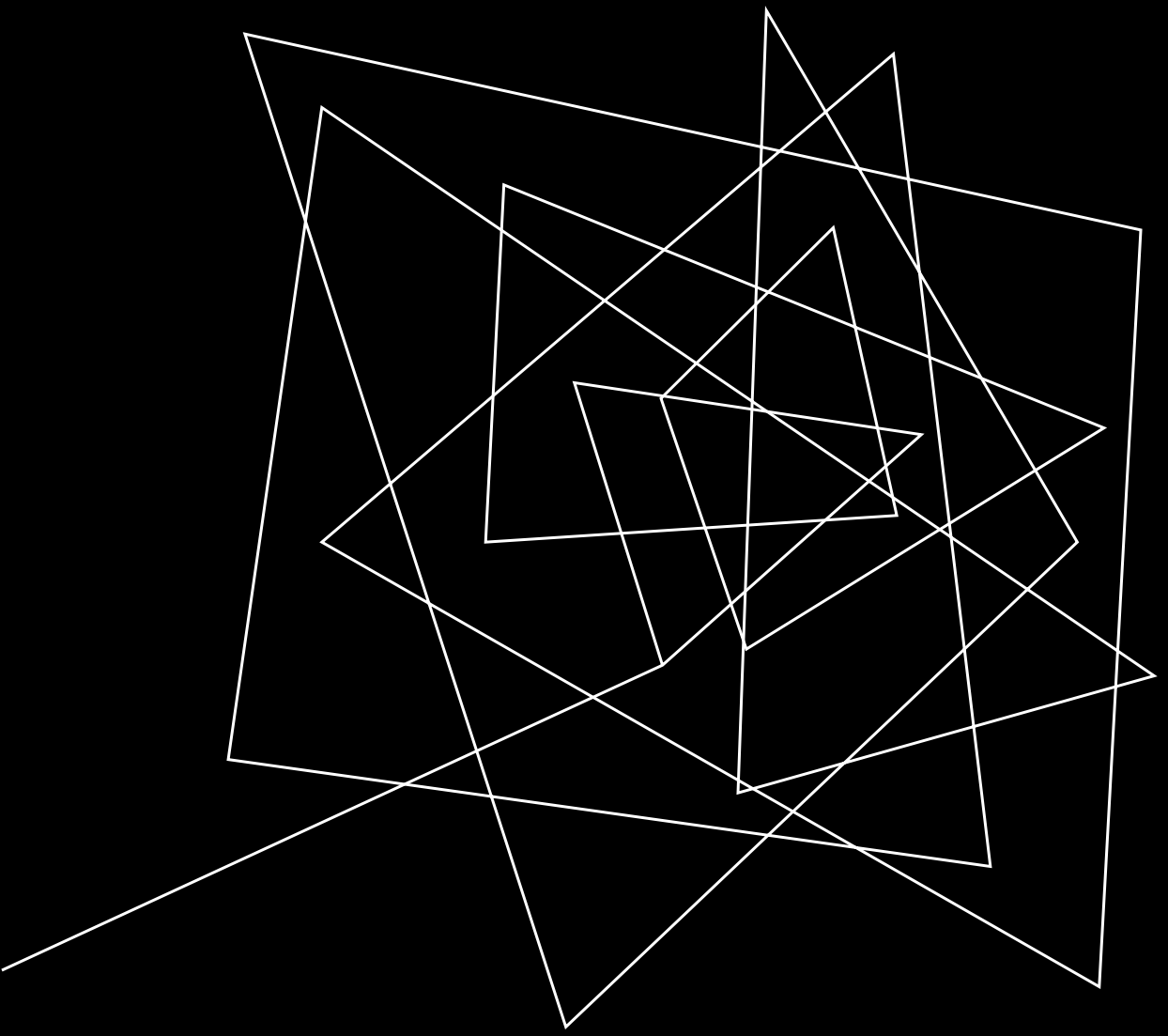
Traditional Algorithms

ChatGPT 3.5

Conclusion

# INTRODUCTION

Our project focuses on predicting obesity levels using a diverse set of machine learning algorithms, combining traditional models and advanced artificial intelligence (AI) capabilities. The dataset encompasses various patient features, and our goal is to build a powerful ensemble model that leverages the strengths of both traditional machine learning and AI models.



# PROBLEM STATEMENT

# PROBLEM STATEMENT

Task of Classifying health data is a big task, especially when it comes to predicting issues like liver disease and obesity Levels. The regular methods I use for this sometimes struggle to get things right, even after I put a lot of effort into fixing them. I've tried adjusting the features I look at, and fine-tuning our models, but it's still a tough nut to crack. Dealing with the mix of things that can lead to obesity adds another layer of difficulty. The goal is to figure out how to build models that really get what's going on and can predict health problems accurately. It's not just a ML project; it's about making a real impact on how we understand and deal with health issues using data and technology.

# DATA SET

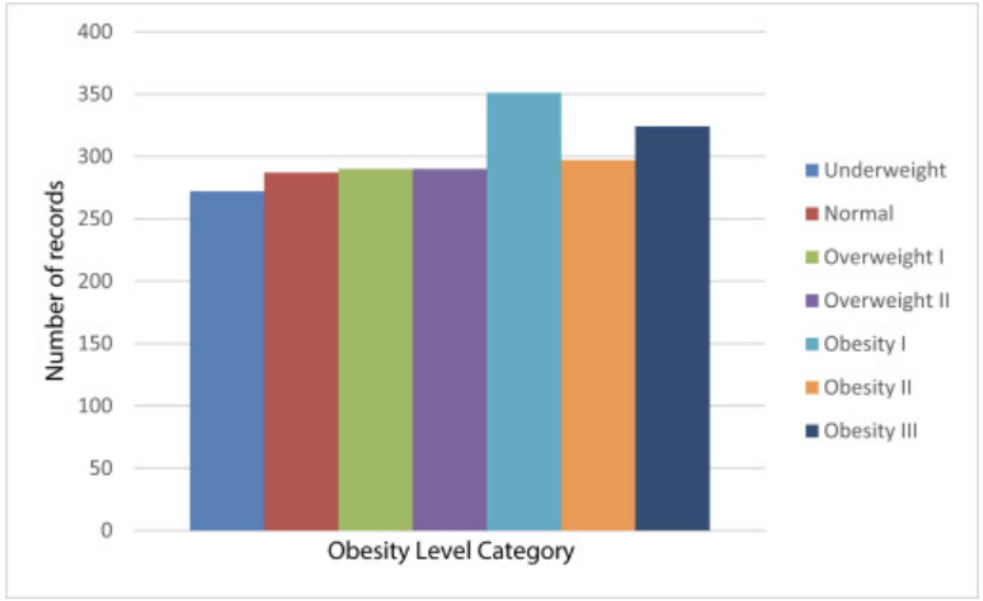
- **Source** : UCI
- **Size : 2111** rows with labels and 17 features
- **Description** : This dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), that allows classification of the data using the values of Insufficient weight, Normal weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

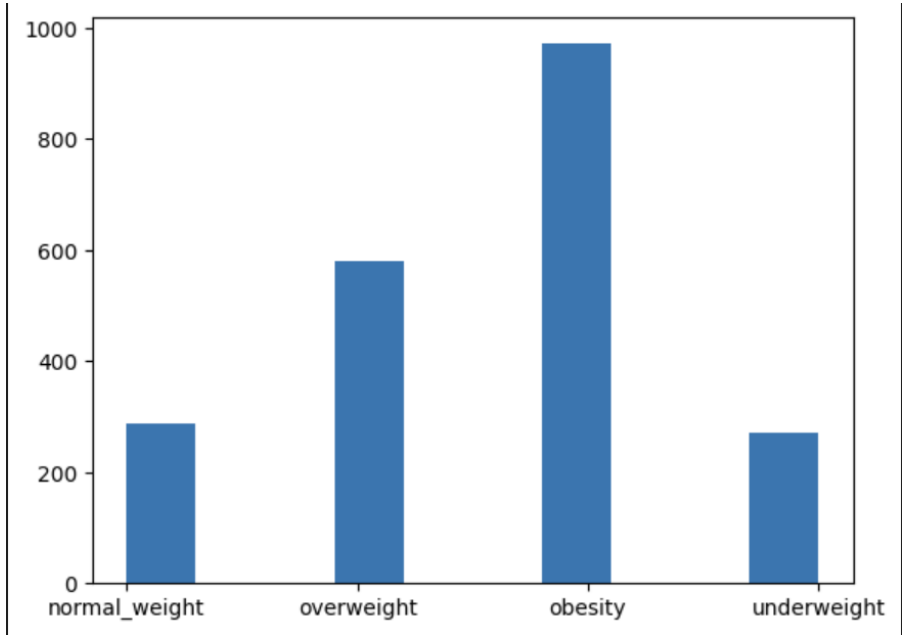
**Link**: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

# DATA EXPLORATION AND PREPROCESSING

# DATA EXPLORATION

- *Initially, there Ire 7 class labels "Obesity_Type_I", "Obesity_Type_II", "Obesity_Type_III", "Overweight_Level_I", "Overweight_Level_II", "Normal_weight", "Insufficient_weight". I reduced them into 4 class labels by grouping all the obesity types into "obesity" and all overweight types into "overweight".*
- *This created class imbalance with classes "normal_weight" and "underweight" having lesser number of rows.*
- *Class distribution before and after grouping class labels:*



# DATA EXPLORATION

- I built correlation matrix which can be used to determine how much variables are related to one another.
- Through the correlation matrix, the attributes “family_history_with_overweight”, “Frequent consumption of high caloric food”, “weight”, “Transportation used” and “Consumption of food between meals” are found to have greater correlations with the target variable.

| | Category_type_corelation_score |
|---|---|
| Age | -0.102718 |
| Height | -0.011333 |
| Weight | -0.367647 |
| FCVC | -0.033275 |
| NCP | 0.001036 |
| CH2O | -0.008295 |
| FAF | 0.038602 |
| TUE | 0.081648 |
| Category_type_encoded | 1.000000 |
| Gender | 0.123498 |
| family_history_with_overweight | 0.522162 |
| FAVC | 0.297247 |
| CAEC | 0.336019 |
| SMOKE | 0.081150 |
| SCC | 0.194236 |
| CALC | 0.119429 |
| MTRANS | 0.153978 |

# DATA EXPLORATION

- Males and females are having equal in numbers when it comes to obesity.
- When it comes to family history with overweight, more than 85 percent of them are expected to have either overweight or obesity problems.
- Frequent consumption of high caloric food and Consumption of food between meals are directly proportional to overweight and obesity.
- When it comes to transportation impact on obesity, it looks like people using public transportation are having more obesity when compared to other transportation types which has obesity.

# DATA PREPROCESSING

- The dataset has categorical variables such as 'Gender', 'family_history_with_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS'. I cannot use label encoding because the model can assume there is some kind of ordering among the data. Hence, I have used Onehot encoding.
- The dataset has numerical variables such as 'Age', 'Height', weIight', 'FCVC', 'NCP', 'CH2O', 'FAF', 'TUE' and all of them are in different scales.
- I have normalized all the numerical variables using Standard Scaler which would bring down the domain of each numerical variable into a certain window like [-2,2]

# TRADITIONAL MACHINE LEARNING MODELS

# ALGORITHMS SELECTED

## Decision Tree

- Developed a decision tree model to decode the intricacies of obesity prediction from patient data.
- Employed a robust 10-fold stratified cross-validation method with hyper-parameter tuning to ensure the model's adaptability to diverse data patterns. explores options Max depth, Min Samples split and min samples Leaf using grid search

## Random Forest

- Build an Initial model orchestrating an ensemble of predictors for obesity level prognosis.
- Ensured model robustness with a 10-fold stratified cross-validation, preventing overfitting and validating across diverse data patterns by conducting Hyper parameter tuning at every iteration finding the with best combination of n estimaters, max depth, min split and min leaf samples.

## SVC

- Support Vector Classifier to unravel the complexities of predicting obesity levels, opting for the linear kernel's simplicity.
- Maintained model stability with a 10-fold stratified cross-validation, ensuring consistent performance across diverse data subsets and parameters tuned at every split by changing kernel type, coefficient and regularization parameter.

## KNN

- Trained a KNN model to predict obesity levels, leaning on the collective wisdom of nearby neighbors.
- scrutinizing the model's performance through a 5-fold stratified cross-validation, ensuring a fair representation of each class with parameter tuning neighbors and weight.

# TRADITIONAL MODELS AND THEIR ACCURACIES

| Model Name | Accuracy | weighted F1-Score |
|---|---|---|
| Decision Trees | 94 | 95 |
| Random Forest | 94 | 95 |
| Support Vector Machines (SVM) | 93 | 94 |
| K-Nearest Neighbors (KNN) | 88 | 87 |

# CROSS VALIDATION WITH HYPER - PARAMETER TUNING

| Model Name | Accuracy | weighted F1-Score |
|---|---|---|
| Decision Trees | 95 | 95 |
| Random Forest | 94 | 94 |
| Support Vector Machines (SVM) | 98 | 98 |
| K-Nearest Neighbors (KNN) | 90 | 90 |

# CHATGPT 3.5 TURBO

I harnessed the power of GPT-3.5 Turbo, fine-tuning it to cater to the intricacies of predicting obesity levels from patient data. This involved customizing the model to comprehend and respond effectively to health-related inquiries. By tailoring GPT-3.5 Turbo to our specific context, I aimed to elevate its predictive capabilities, enabling it to generate insightful and contextually relevant predictions regarding obesity. This integration represents a synergy betIen traditional machine learning approaches and advanced AI, bringing forth a comprehensive solution for health predictions.

# FINE TUNING GPT 3.5

- The dataset is split into training, validation, and test sets.
- Data from the sets are converted into JSONL format for OpenAI model training. Separate JSONL files are created for training, validation, and testing.
- JSONL stores data in multiple lines in a text file, with each line containing a single JSON object.
- But JSONL files cannot be directly generated from .csv files. They should be converted into document format, and it can be done through CSVLoader.
- After training the model through train.jsonl file, I validate and adjust the model by passing validation.jsonl file and finally test it through test.jsonl.

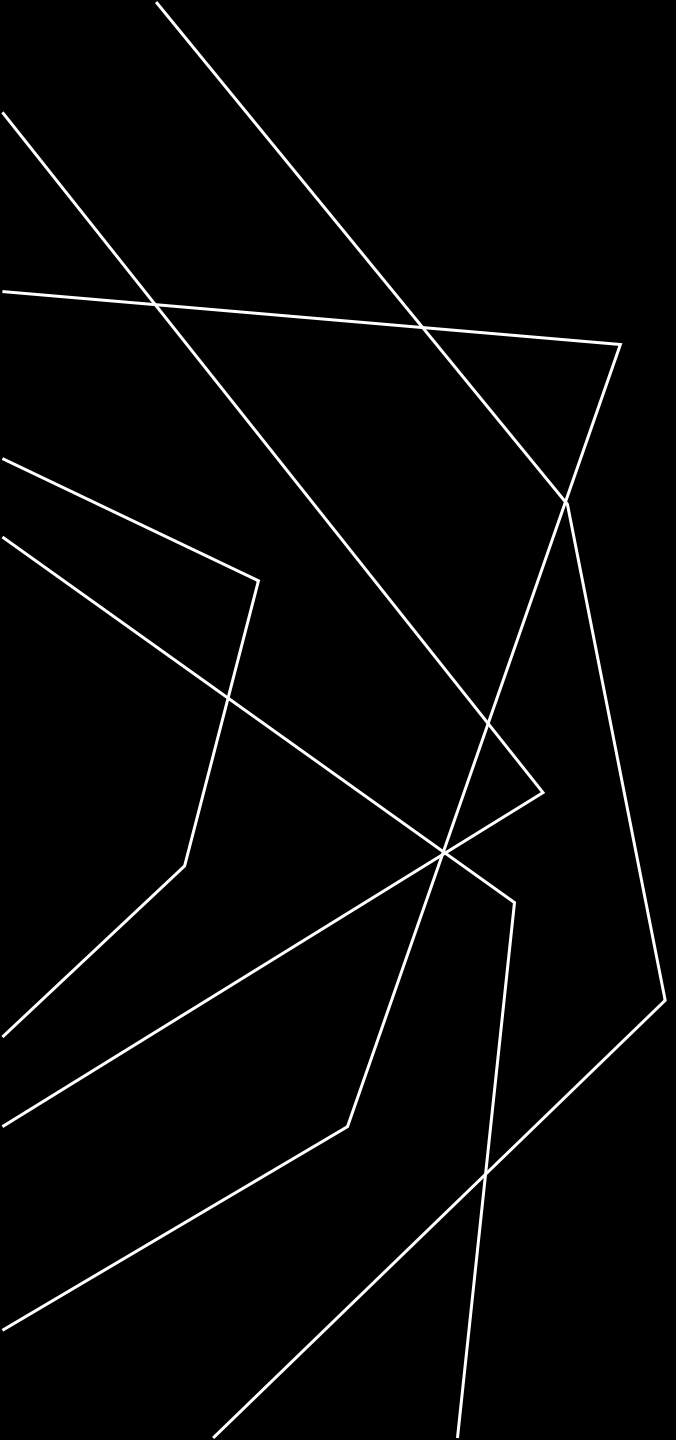
# ENSEMBLING METHOD

- As discussed before, I created four traditional models with the training data and then combine Cross validation and Hyperparameter tuning to get the best weighted F1 score for each. In cross validation, I divided the dataset into ten folds where nine folds would be used for training and one-fold for validation.
- The above process would be done for ten iterations where in each iteration the data would be shuffled and the model with the best weighted F1 score would be selected. This process is done for all the models.
- Now all the 4 traditional models along with the GPT 3.5 model are used in the ensemble process.
- For a given training set, the predictions of all the models are placed in a dataframe.
- Then voting is applied among all the models and the prediction with the highest vote count is selected for a particular test record.
- Ensembling eliminates the disadvantage of data imbalance in training set because even if couple of models from our ensemble model predict the class label incorrectly and three of them predict correctly, the ensemble output would be correct class label.
- Thus, it is better to use ensemble model when compared to single classifier when the data has imbalance classes.

# CODE WALK-THROUGH

# CONCLUSION

In conclusion, leveraging a diverse set of learning algorithms, including the robust GPT-3.5 model with external world knowledge, proves to be a sensible strategy for enhancing predictive accuracy. The adoption of an ensemble model, comprised of multiple classifiers, offers distinct advantages, particularly in scenarios characterized by imbalanced datasets. Through the combination of these models, I not only mitigate the limitations of relying on a singular approach but also attain superior overall performance, making ensemble models a preferred choice for predictive tasks in various applications.



# THANK YOU