SUGAR CLASSIFICATION

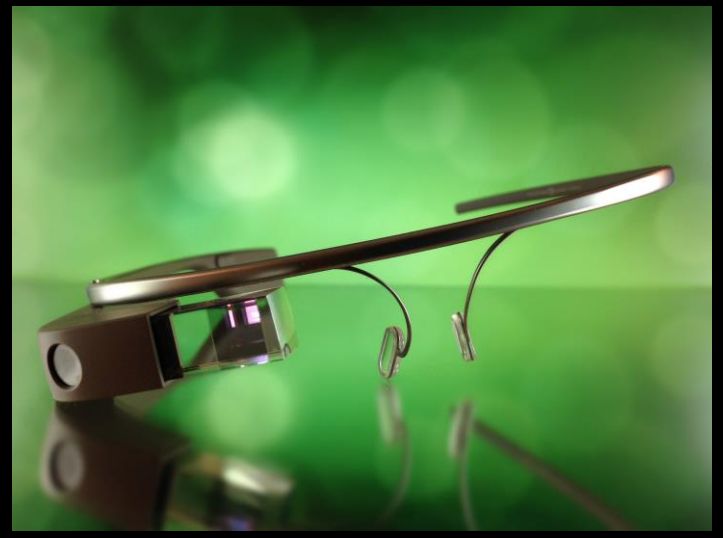
Predicting sugar content of dishes using menu text and description

Norman Jen

THE PROBLEM

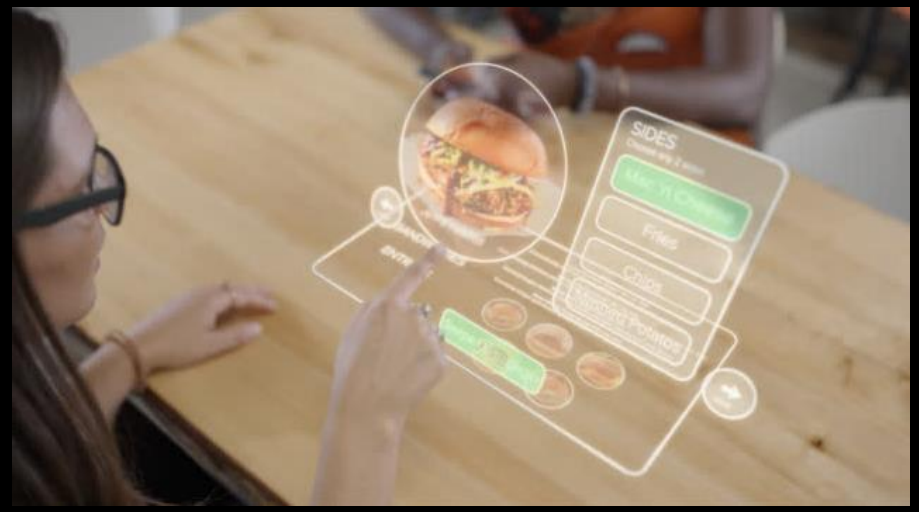
Purpose

- Google Glass is being revamped
- Use of new NLP and ML techniques



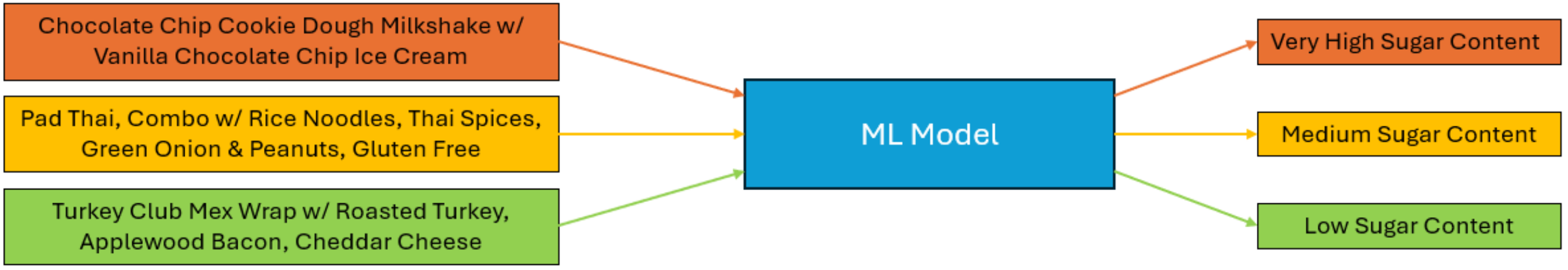
Users

- Diabetics
- Weight Loss



THE SOLUTION

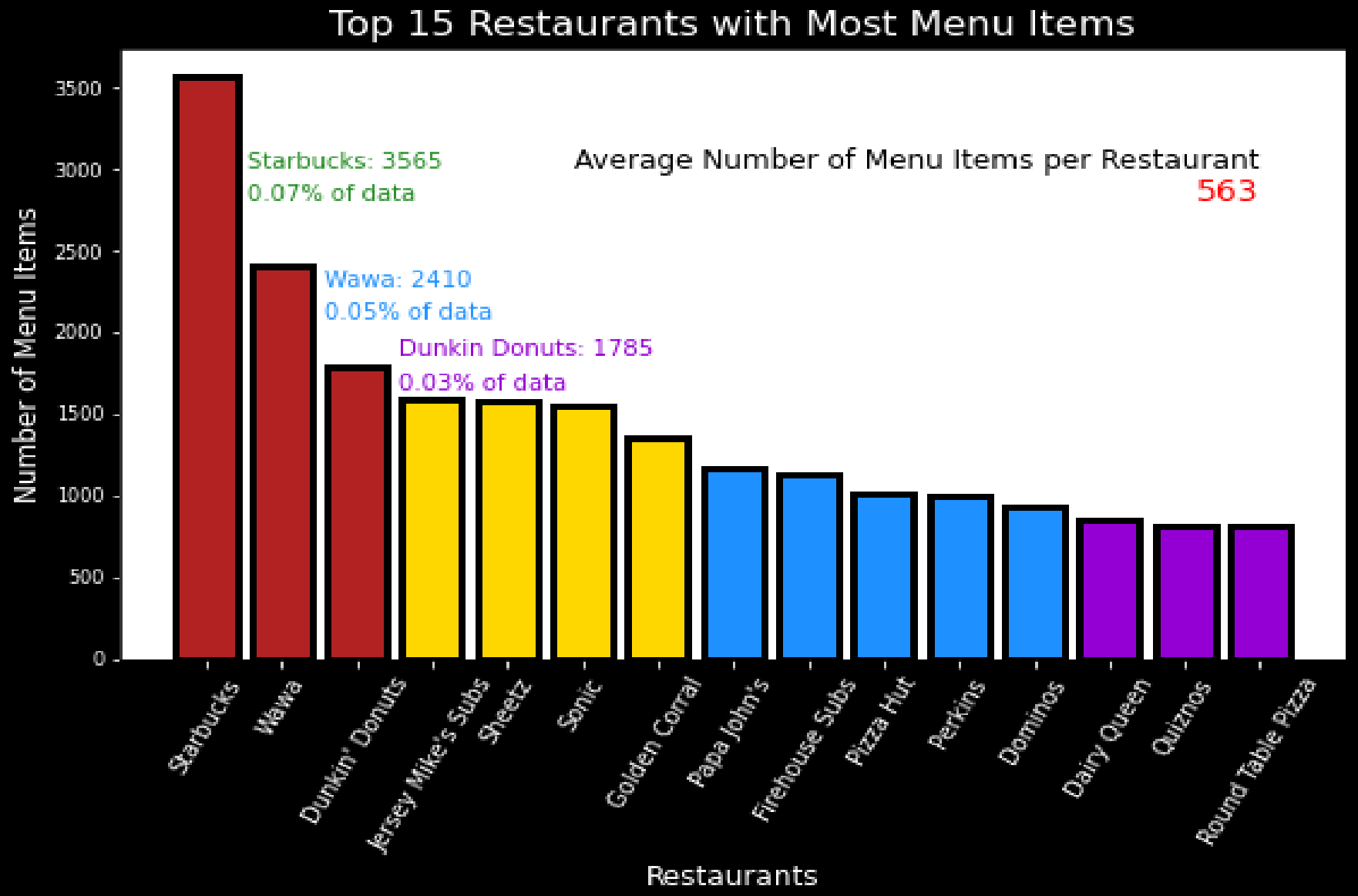
- Text Recognition
- ML Model
- Classification – Sugar Content



THE DATA

- NYC OpenData
- 2017-2018
- Chain Restaurants
- Menu Descriptions
- Nutrition Facts

Starbucks: 3565
Wawa: 2410
Dunkin' Donuts: 1785
Jersey Mike's Subs: 1585
Sheetz: 1584
Sonic: 1553
Golden Corral: 1351
Papa John's: 1165
Firehouse Subs: 1134
Pizza Hut: 1017
Perkins: 991
Dominos: 927
Dairy Queen: 847
Quiznos: 816
Round Table Pizza: 814



SUGAR CLASSES

- Very high sugar – >= 30g
- High sugar – 7 – 30g
- Medium sugar – 2 – 7g
- Low sugar – 0 – 2g
- Zero sugar – 0g

- 116 + calories
- 27 – 116 calories
- 8 – 27 calories
- 0 – 8 calories
- 0 calories

USDA – 10% of daily calories or less

- 58+%
- 14 – 58%
- 1 – 14%
- 0 – 1%
- 0%

AHA – 6% of daily calories or less

- 97+%
- 23 – 97%
- 2 – 23%
- 0 – 2%
- 0%

*Assuming 2,000 calorie diet

THE PROCESS

Data Preparation

Text Processing

Initial Modeling

Tuning and Selection

Evaluation

Ensemble

DATA PREPARATION

- Feature Selection
- Missing Data
- Bin Target Variable
- Prepare Text

```
0   Menu_Item_ID              65219 non-null  int64
1   Year                      65219 non-null  int64
2   Restaurant_Item_Name      65219 non-null  object
3   restaurant                65219 non-null  object
4   Restaurant_ID             65219 non-null  int64
5   Item_Name                 65219 non-null  object
6   Item_Description          65219 non-null  object
7   Food_Category             65219 non-null  object
8   Serving_Size              26899 non-null  float64
9   Serving_Size_text         39 non-null     object
10  Serving_Size_Unit         26927 non-null  object
11  Serving_Size_household    15238 non-null  object
12  Calories                  55315 non-null  float64
13  Total_Fat                 54846 non-null  float64
14  Saturated_Fat             54143 non-null  float64
15  Trans_Fat                 51503 non-null  float64
16  Cholesterol               53219 non-null  float64
17  Sodium                    54991 non-null  float64
18  Potassium                 1098 non-null   float64
19  Carbohydrates             54288 non-null  float64
20  Protein                   54233 non-null  float64
21  Sugar                     52931 non-null  float64
22  Dietary_Fiber             53440 non-null  float64
23  Calories_100g             25878 non-null  float64
24  Total_Fat_100g            25686 non-null  float64
25  Saturated_Fat_100g        25285 non-null  float64
26  Trans_Fat_100g            23828 non-null  float64
27  Cholesterol_100g          25126 non-null  float64
28  Sodium_100g               25853 non-null  float64
29  Potassium_100g            623 non-null    float64
30  Carbohydrates_100g        25592 non-null  float64
31  Protein_100g              25531 non-null  float64
32  Sugar_100g                25207 non-null  float64
33  Dietary_Fiber_100g        25391 non-null  float64
34  Calories_text             303 non-null    object
35  Total_Fat_text            69 non-null     object
36  Saturated_Fat_text        50 non-null     object
37  Trans_Fat_text            10 non-null     object
38  Cholesterol_text          358 non-null    object
39  Sodium_text               93 non-null     object
40  Potassium_text            2 non-null      object
41  Carbohydrates_text        634 non-null    object
42  Protein_text              467 non-null    object
43  Sugar_text                591 non-null    object
44  Dietary_Fiber_text        811 non-null    object
45  Kids_Meal                 65219 non-null  int64
46  Limited_Time_Offer        65219 non-null  int64
47  Regional                  65219 non-null  int64
48  Shareable                 65219 non-null  int64
```

```
Restaurant_Item_Name        0
restaurant                  0
Item_Name                   0
Item_Description            0
Food_Category               0
Sugar                   12288
dtype: int64
```

```
0   Restaurant_Item_Name  52931 non-null  object
1   restaurant            52931 non-null  object
2   Item_Name             52931 non-null  object
3   Item_Description      52931 non-null  object
4   Food_Category         52931 non-null  object
5   Sugar                 52931 non-null  float64
```

	sugar_class	text
21688	4	Deconstructed Breakfast Taco Deconstructed Br...
42983	2	Lay's Kettle Cooked 40% Less Fat Original La...
26327	4	Chicken Maui Zaui w/ Polynesian Sauce, Pan,...
31076	3	All Meat Pizza on Gluten Free Crust, Medium,...
13655	4	El Nino Margarita El Nino Margarita El Nino M...

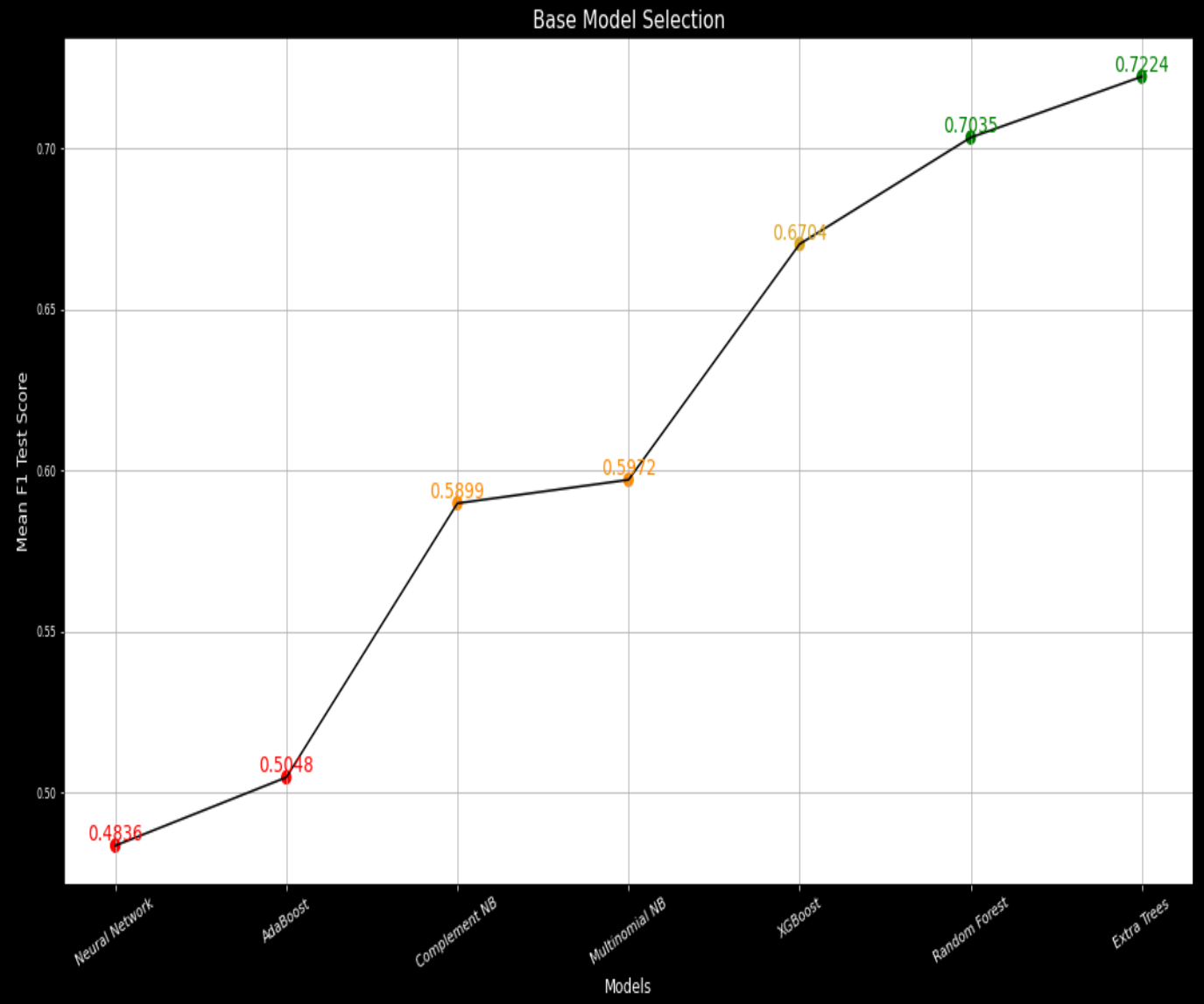
TEXT PROCESSING

- Remove special characters/ numbers
- Lower case
- Tokenization
- Stopwords
- Stemming

```python
from sklearn.base import BaseEstimator, TransformerMixin
import nltk
from nltk.stem import SnowballStemmer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

stop_words = stopwords.words('english')

class Preprocessor(BaseEstimator, TransformerMixin):
    def __init__(self):
        pass

    def fit(self, data, y=None):
        return self

    def transform(self, data, y=None):
        preprocessed_data = [self.stem_doc(doc) for doc in data]
        return preprocessed_data

    def stem_doc(self, doc):
        stemmer = SnowballStemmer('english')
        lower_doc = [token.lower() for token in word_tokenize(doc) if token.isalpha()]
        filtered_doc = [token for token in lower_doc if token not in stop_words]
        stemmed_doc = [stemmer.stem(token) for token in filtered_doc]
        tokenized_doc = " ".join(set(stemmed_doc))
        return tokenized_doc
```

INITIAL MODELING

- Neural Network (Sequential, Dense)
- AdaBoost
- Naïve-Bayes (Multinomial, Complement)
- XGBoost
- Random Forest Classifier
- Extra Trees Classifier



EVALUATION

F1 Score

- False Negative not more important than False Positive
- Interpretable
- Model Selection
- Resistant to Class Imbalance

Log Loss

- Takes class probabilities into account
- Penalizes consistently incorrect labeling
- Differentiable

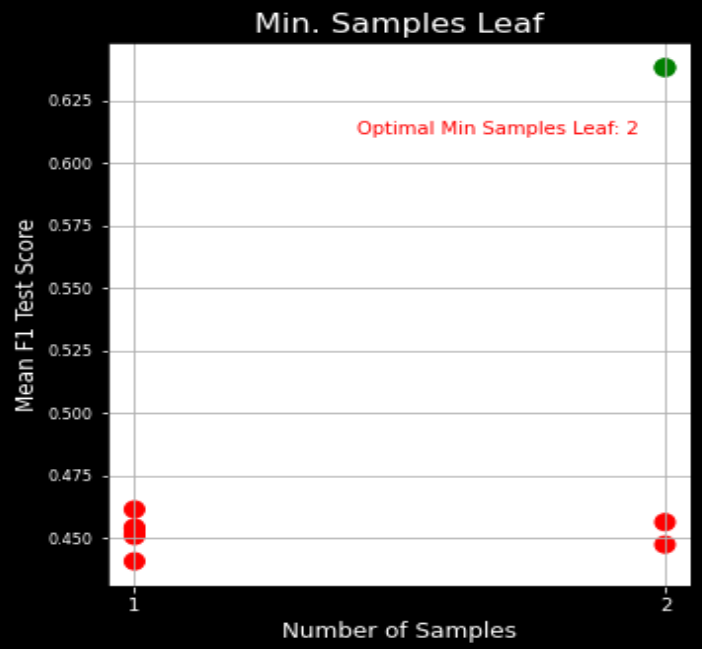
TUNING AND SELECTION

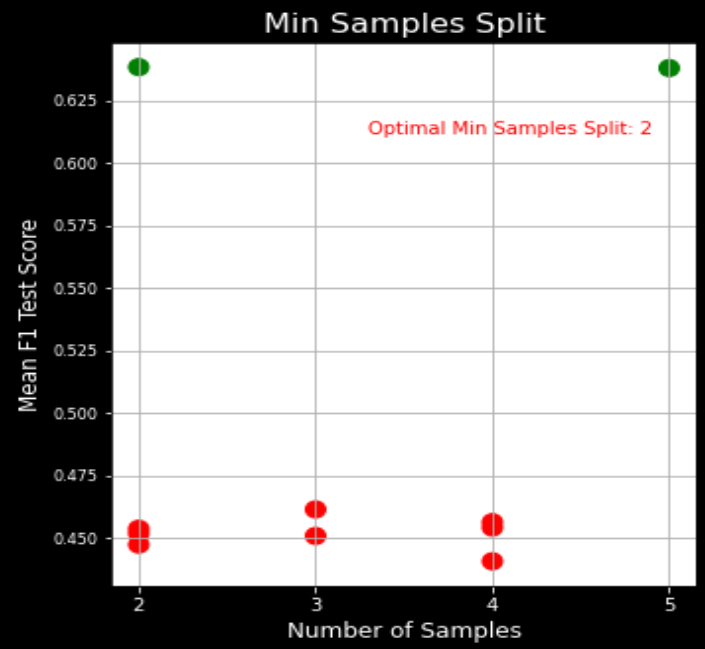
- ~~Neural Network (Sequential, Dense)~~
- ~~AdaBoost~~
- ~~Naïve-Bayes (Multinomial, Complement)~~
- XGBoost
- Random Forest Classifier
- Extra Trees Classifier

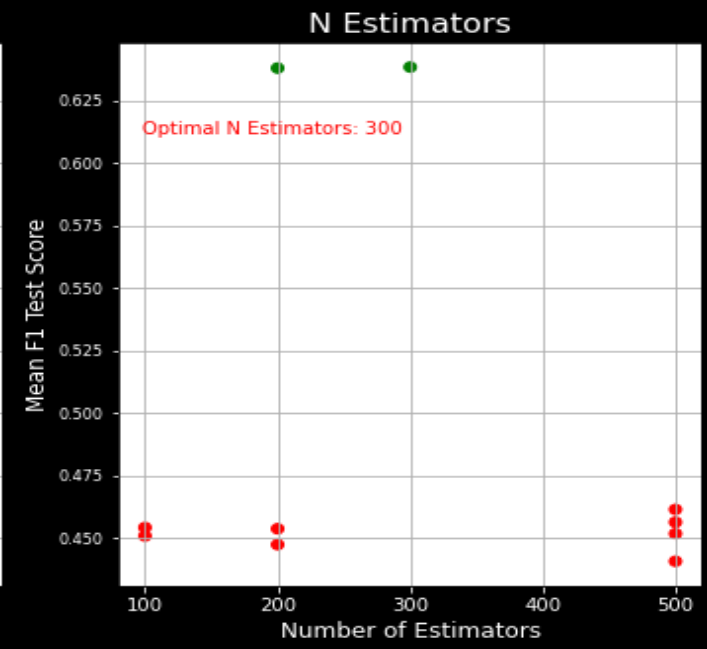
- XGBoost
- Random Forest Classifier
- Extra Trees Classifier

- Hyperparameter Tuning
- RandomizedSearchCV
- Iteration

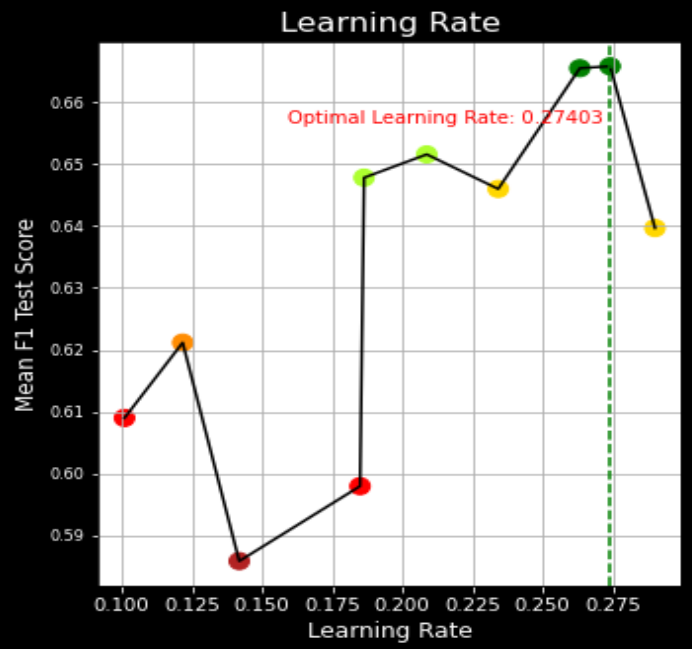
RFC

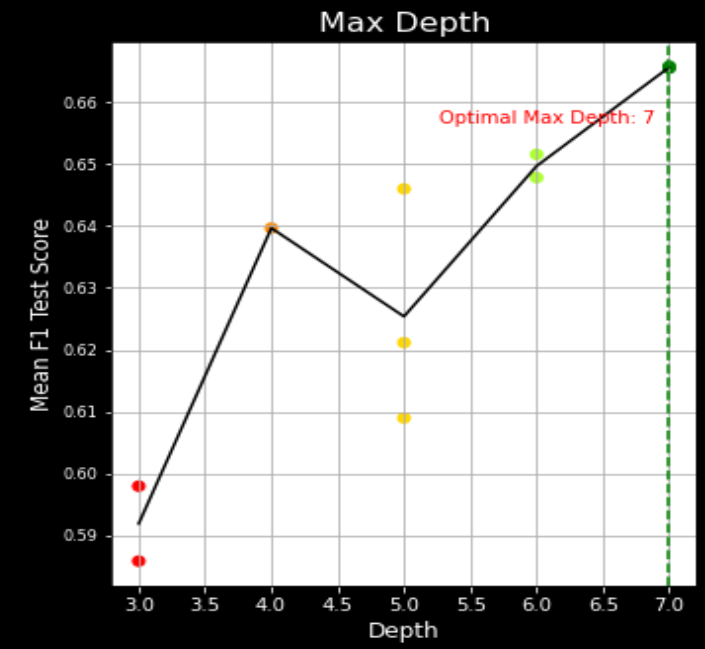

Min. Samples Leaf
Optimal Min Samples Leaf: 2
Mean F1 Test Score
Number of Samples

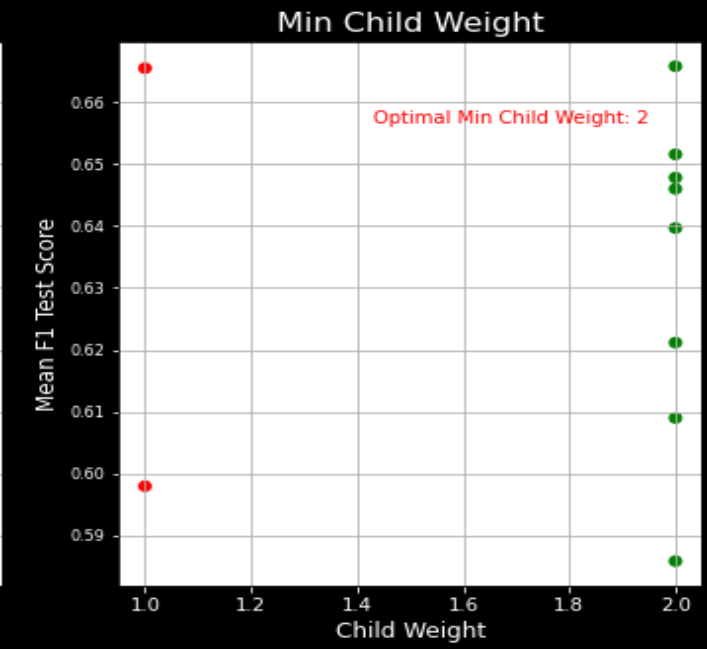

Min Samples Split
Optimal Min Samples Split: 2
Mean F1 Test Score
Number of Samples


N Estimators
Optimal N Estimators: 300
Mean F1 Test Score
Number of Estimators

XGB


Learning Rate
Optimal Learning Rate: 0.27403
Mean F1 Test Score
Learning Rate

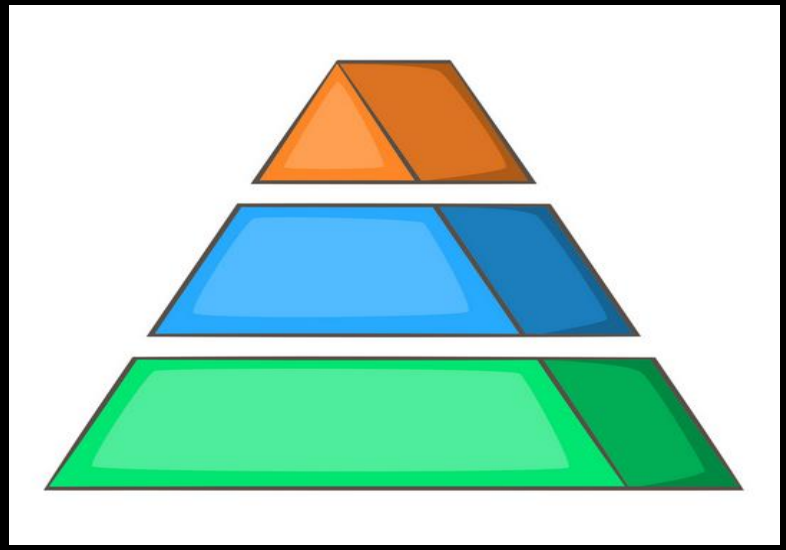

Max Depth
Optimal Max Depth: 7
Mean F1 Test Score
Depth

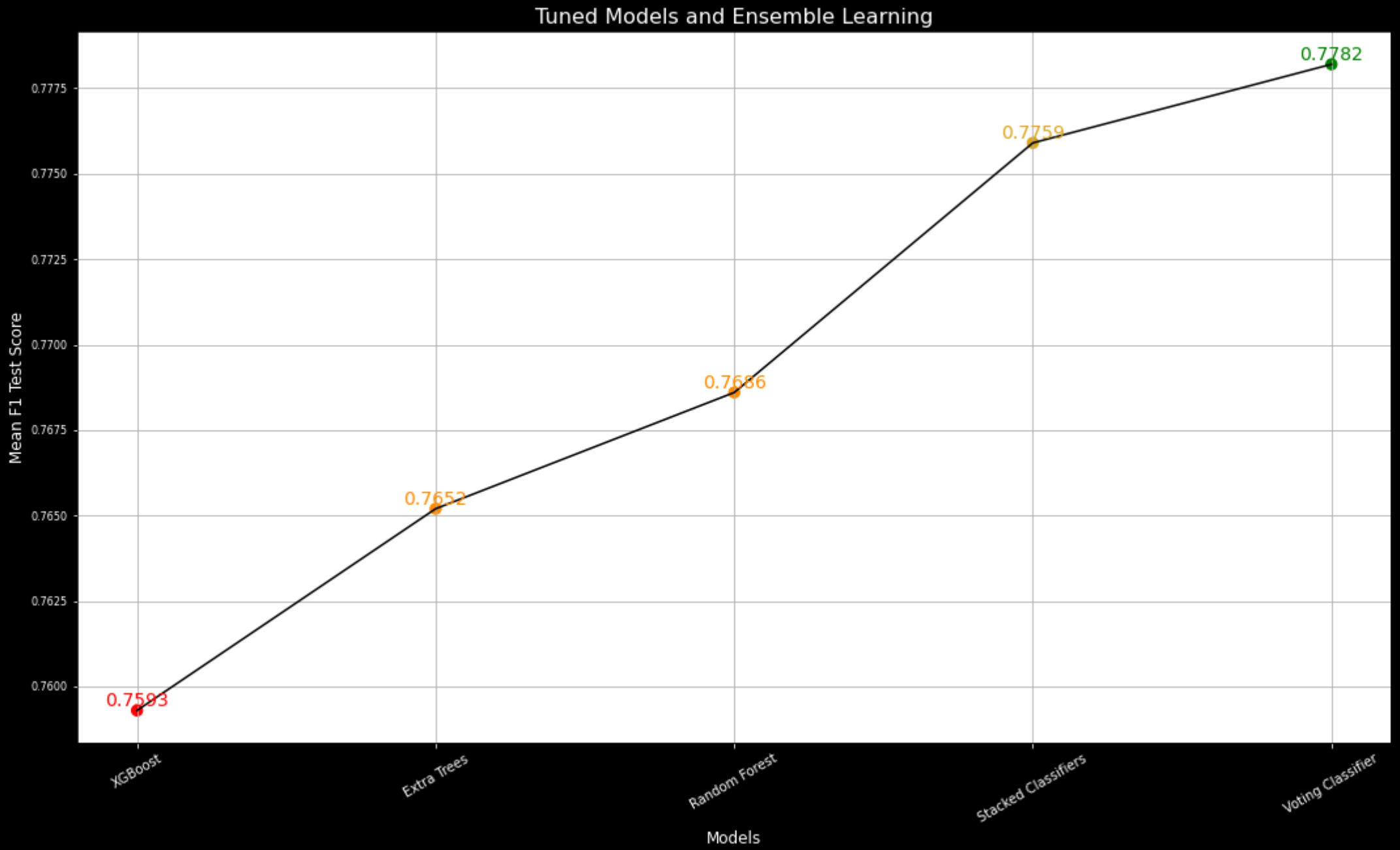

Min Child Weight
Optimal Min Child Weight: 2
Mean F1 Test Score
Child Weight

ENSEMBLE METHODS

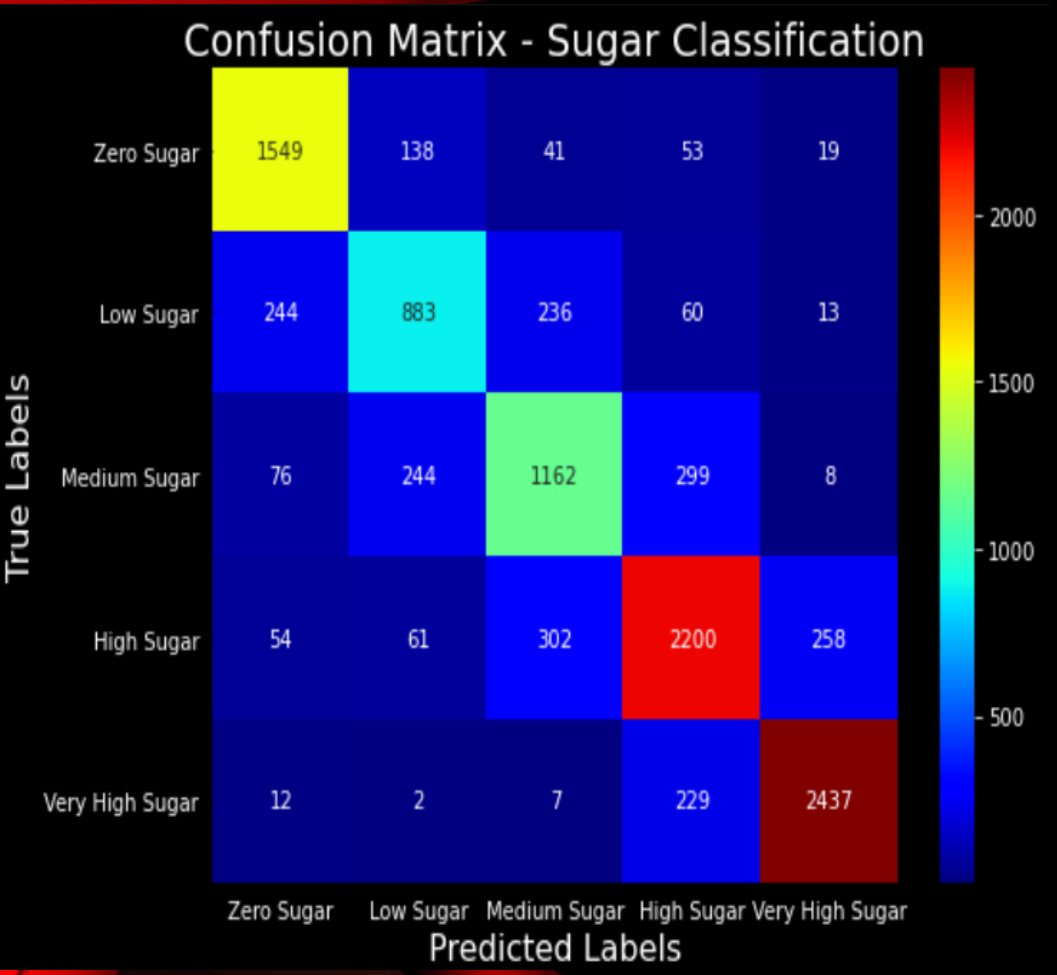
Voting Classifier

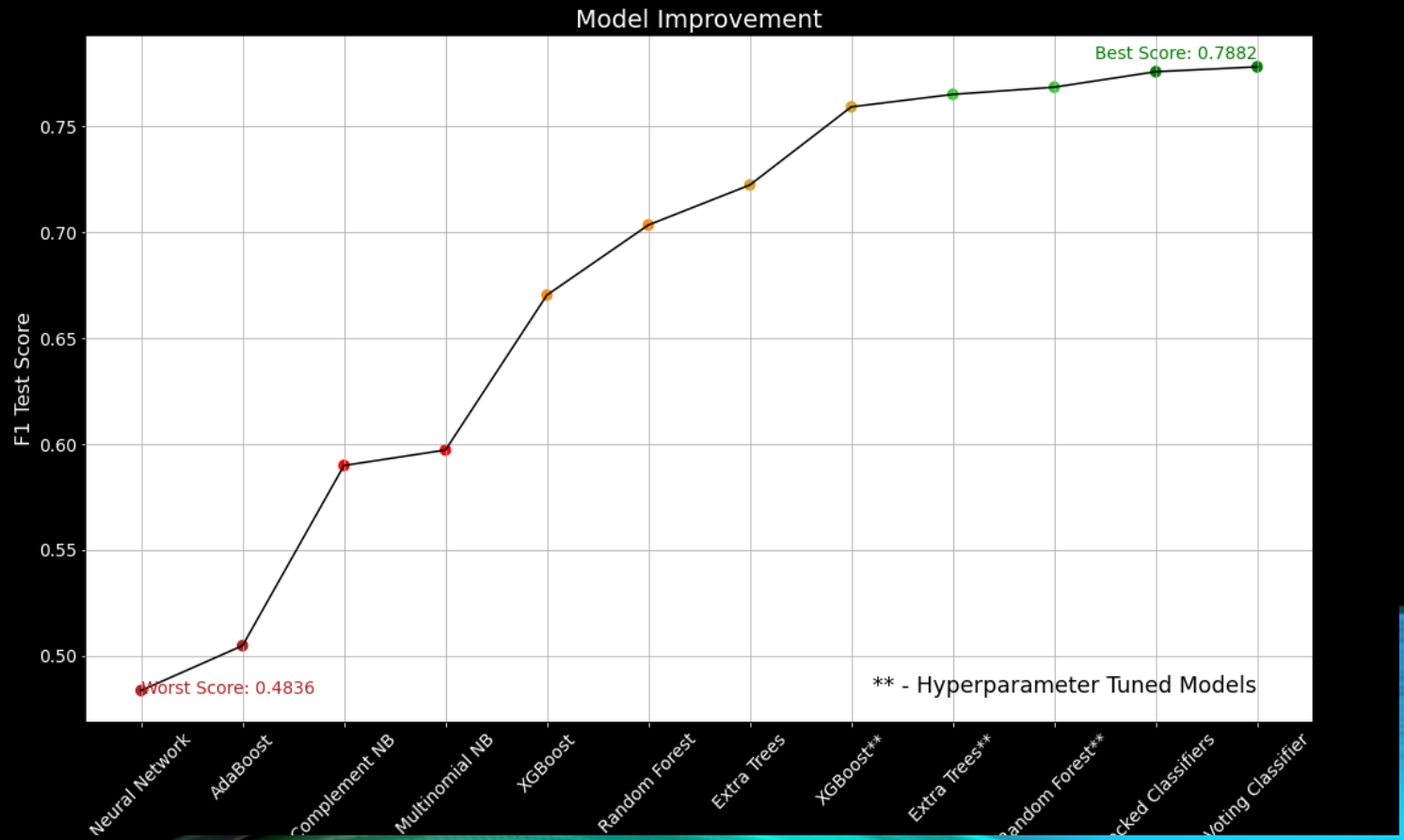
Stacking




Tuned Models and Ensemble Learning
Mean F1 Test Score
0.7775
0.7750
0.7725
0.7700
0.7675
0.7650
0.7625
0.7600
0.7593
0.7652
0.7686
0.7759
0.7782
XGBoost
Extra Trees
Random Forest
Stacked Classifiers
Voting Classifier
Models

RESULTS


Confusion Matrix - Sugar Classification
Zero Sugar
Low Sugar
Medium Sugar
High Sugar
Very High Sugar
1549 138 41 53 19
244 883 236 60 13
76 244 1162 299 8
54 61 302 2200 258
12 2 7 229 2437
2000
1500
1000
500
True Labels
Zero Sugar
Low Sugar
Medium Sugar
High Sugar
Very High Sugar
Predicted Labels


Model Improvement
Best Score: 0.7882
Worst Score: 0.4836
** - Hyperparameter Tuned Models
F1 Test Score
0.75
0.70
0.65
0.60
0.55
0.50
Neural Network
AdaBoost
Complement NB
Multinomial NB
XGBoost
Random Forest
Extra Trees
XGBoost**
Extra Trees**
Random Forest**
cked Classifiers
Voting Classifier

NEXT STEPS

More Data

Neural Network

Re-visiting Text Processing

Further Tuning

QUESTIONS?

Norman Jen

normcjen3@gmail.com

https://www.linkedin.com/in/normanjen/