MULTI CLASS PREDICTION OF OBESITY RISK

COMPUTATIONAL LEARNING

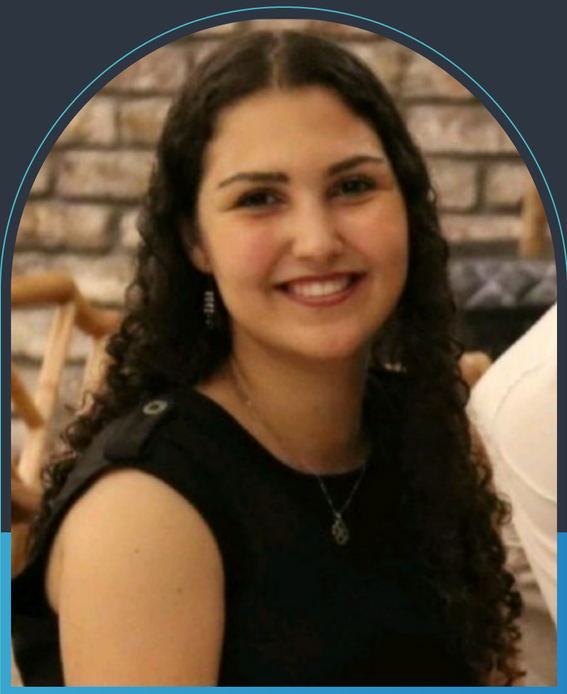
Kaggle compatition link:
multi-class prediction of obesity risk

lecturer:Yehudit Aperstein

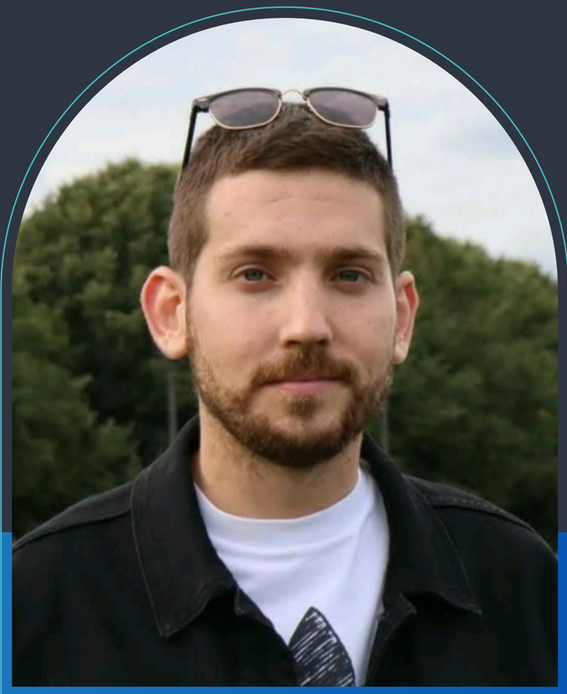
FINAL PROJECT

OUR TEAM

Shahar Felman



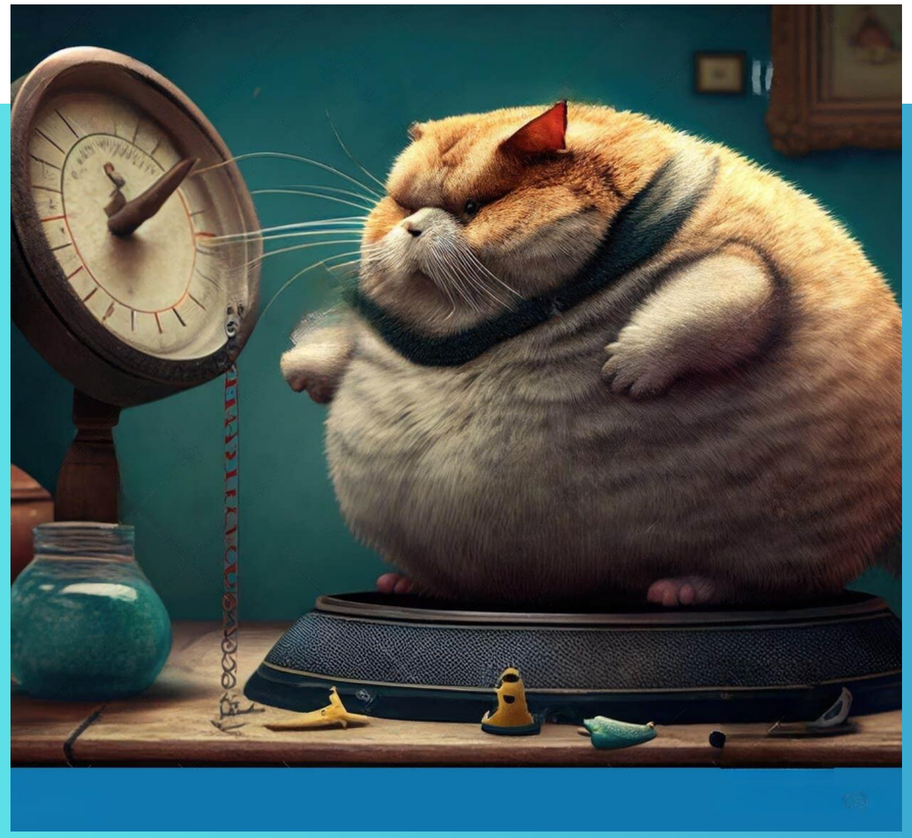
Noy Tsafrir



Tom Mandel

INTRUODUCTION

Obesity is a chronic complex disease defined by excessive fat deposits. It can lead to health problems and often comes from eating too much junk food and not moving enough.



Obesity is a rising global health issue with significant impacts on health and economy.

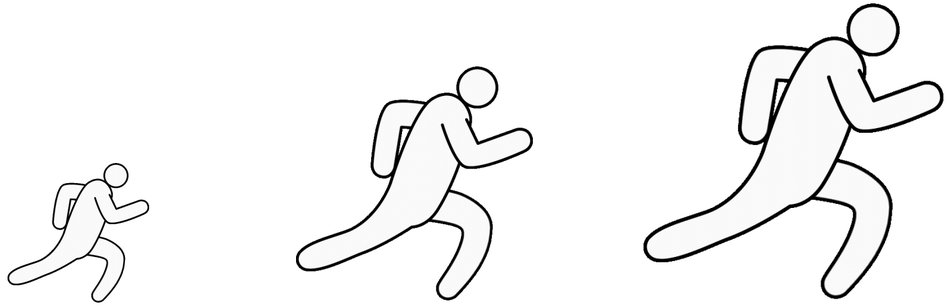
WHO (World Health Organization) estimates over 40% of the global population will be overweight by 2030.

Current tools like BMI calculators are limited, not considering crucial factors like family history, physical activity, and genetics.

OUR PURPOSE

To predict the risk of obesity among individuals based on different parameters, categorizing them into multiple classes (underweight, normal weight, overweight, etc.).

Evaluated by the **accuracy** score.



OUR DATA

- our dataset was generated from a deep learning model trained on the Obesity risk dataset.
- Feature distributions are close to, but not exactly the same, as the original.

	original dataset	our dataset
number of observations	2,111	20,758

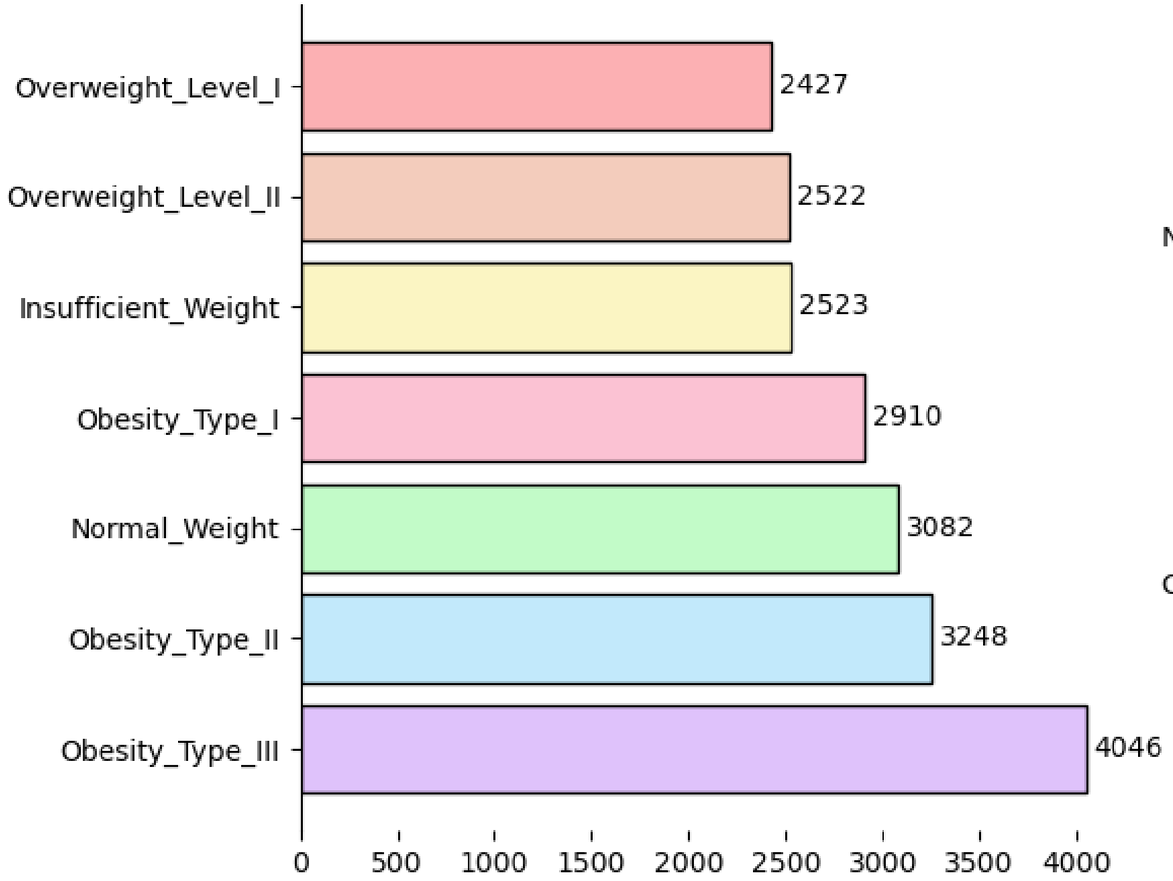
categorical

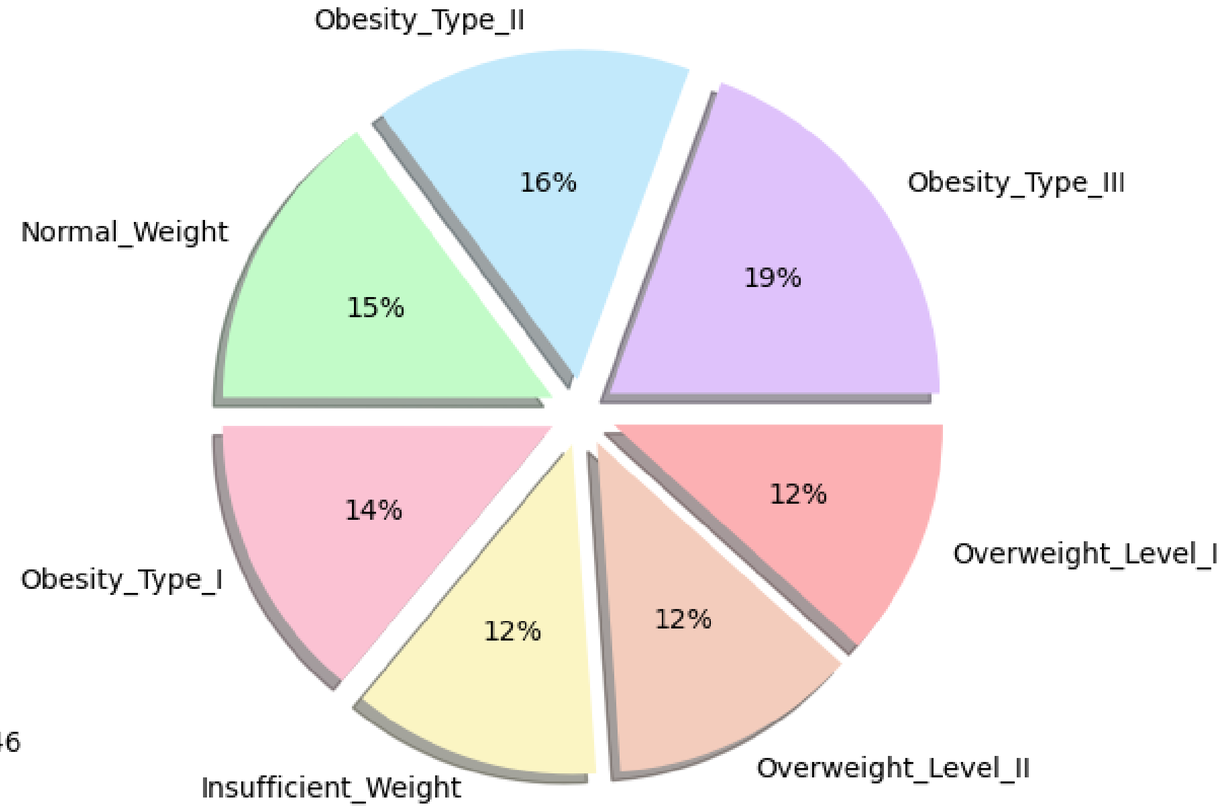
numerical

PREDICTORS

Gender	male or female
Age	age in years (14-61 years)
Height	height in meters (1.45-1.98 m)
Weight	weight in kilograms (39-165 kg)
FAVC	Frequently consumed high-calorie food (yes/no)
FCVC	Frequency of consumption of vegetables (1-3)
NCP	Number of main meals (1-4)
CAEC	Consumption of food between meals (no / sometimes / frequently / always)
Smoke	yes or no
CH2O	Consumption of water daily in Liters (1-3 L)
SCC	The individual keeps track of their caloric intake (yes or no).
FAF	Frequency of physical activity in week (0-3)
TUE	Time using electronic devices in a day in hours (0-3 H)
CALC	Consumption of alcohol (no / sometimes / frequently / always)
MTRANS	Type of transportation used (automobile / motorbike / bike / public transportation / walking)
Family History Of Overweight	The individual has a family member who is overweight or obese (yes / no)

RESPONSE VALUSE





DATA PRE-PROCESSING

- Checked for Missing Information
- Handling Categorical Predictors: (“get dummies”, “label encoder”)
- Removed Irrelevant Columns
- Feature Engineering (BMI)
- Train-Test Split (80-20)

DECISION TREE MODEL:

Flowchart-like tree structure

- Internal node ⟶ Feature (or Predictor)
- Branch ⟶ Decision rule
- Each leaf node ⟶ Outcome.

+ Easy to understand and interpret.

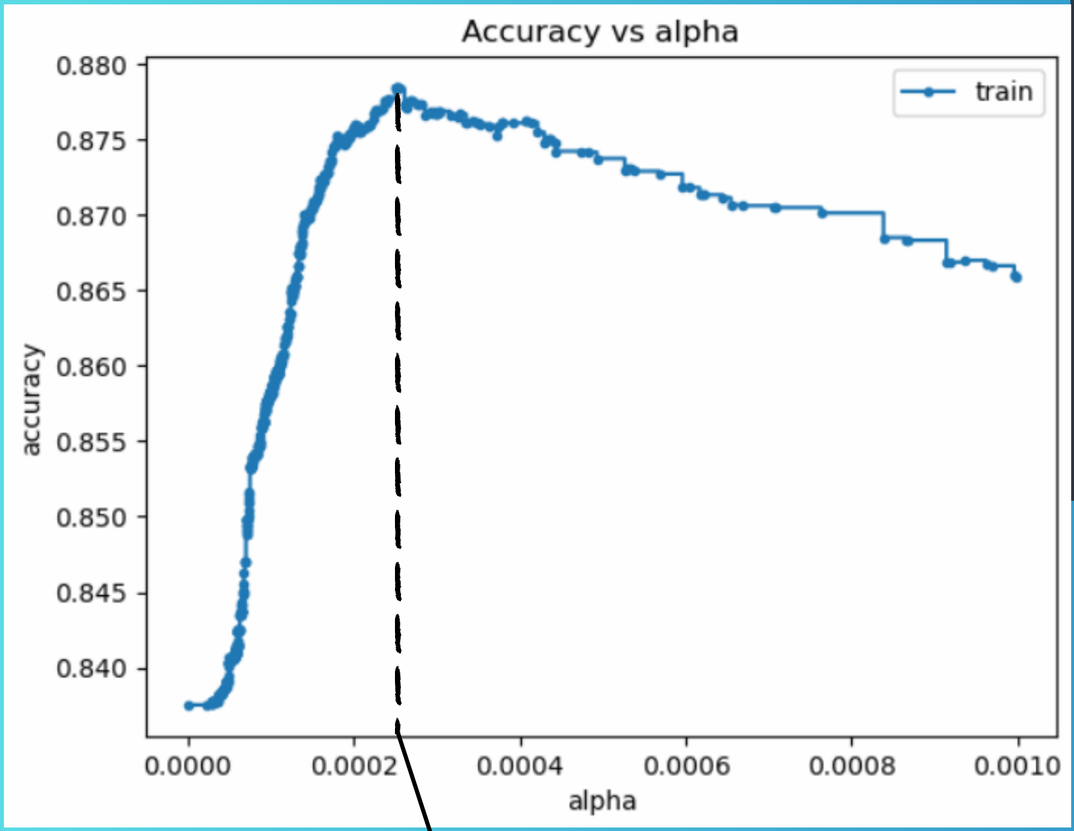
+ Little data preprocessing needed.

+ Handles both numerical and categorical data.

− Overfitting Risk (prunning can make it better)

− Minor data changes can result in different trees.

DECISION TREE IN OUR PROJECT



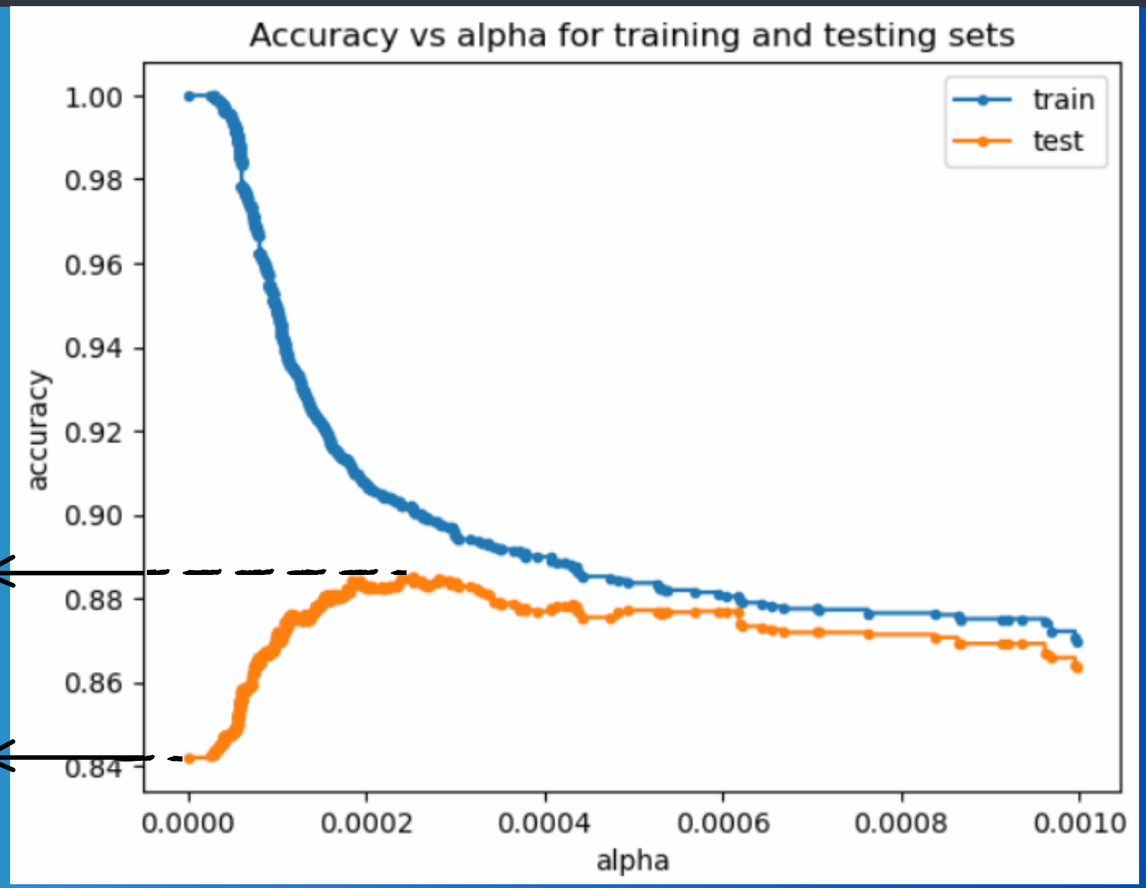
α = 0.00025

Default

↳ **cv --> best alpha**

↳ **prunning**

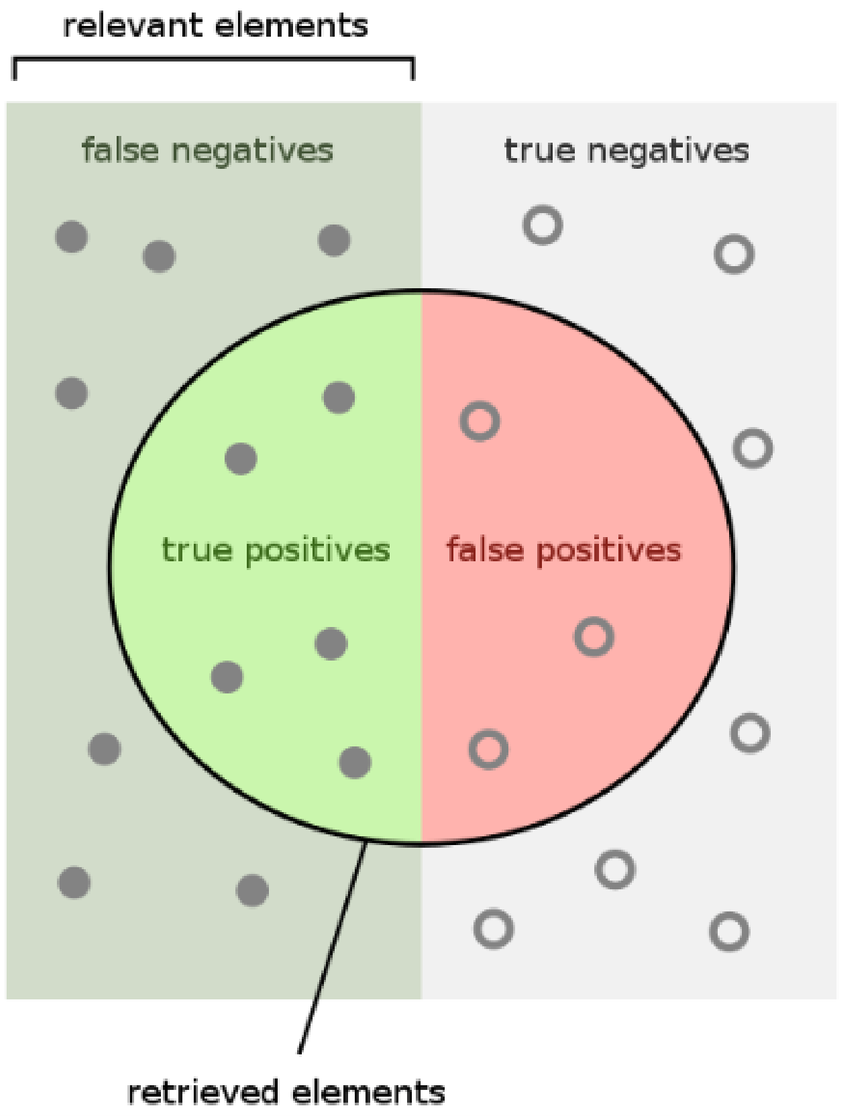
↳ **best accuracy**



0.8829

0.8420

SCORE METRICS



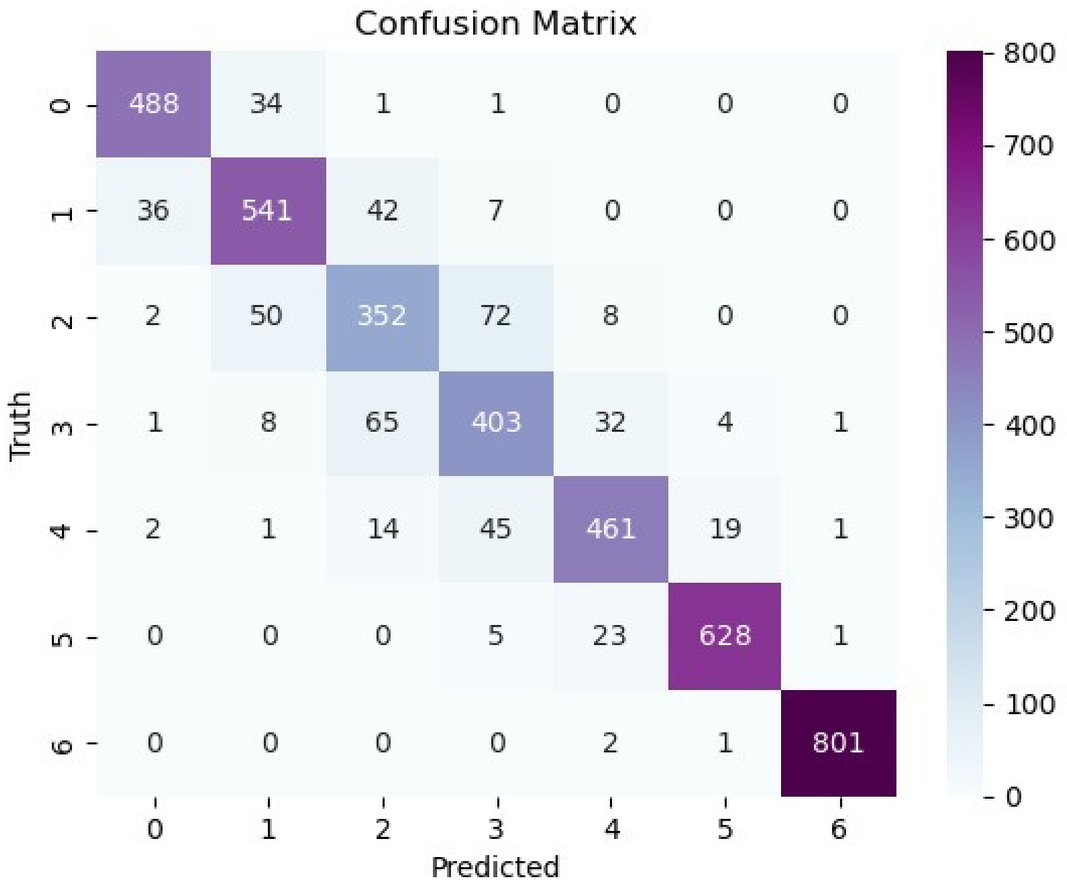
Precision = $\frac{TP}{TP + FP}$

Recall = $\frac{TP}{TP + FN}$

F1 = $\frac{2 * Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$

Accuracy = $\frac{TP+TN}{Total}$

DECISION TREE RESULTS



Accuracy	0.885
Presicion	0.873
Recall	0.873
F1-Score	0.873

LOGISTIC REGRESSION

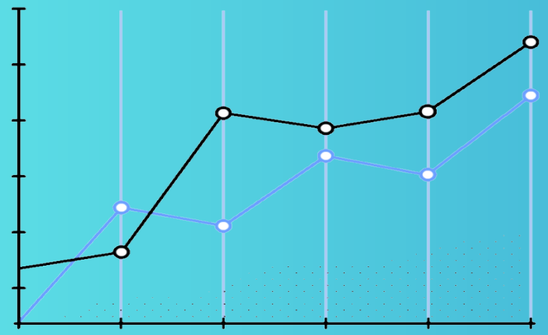
Logistic regression is a statistical method used for binary classification, predicting the probability of an event occurrence by fitting data to a logistic curve.

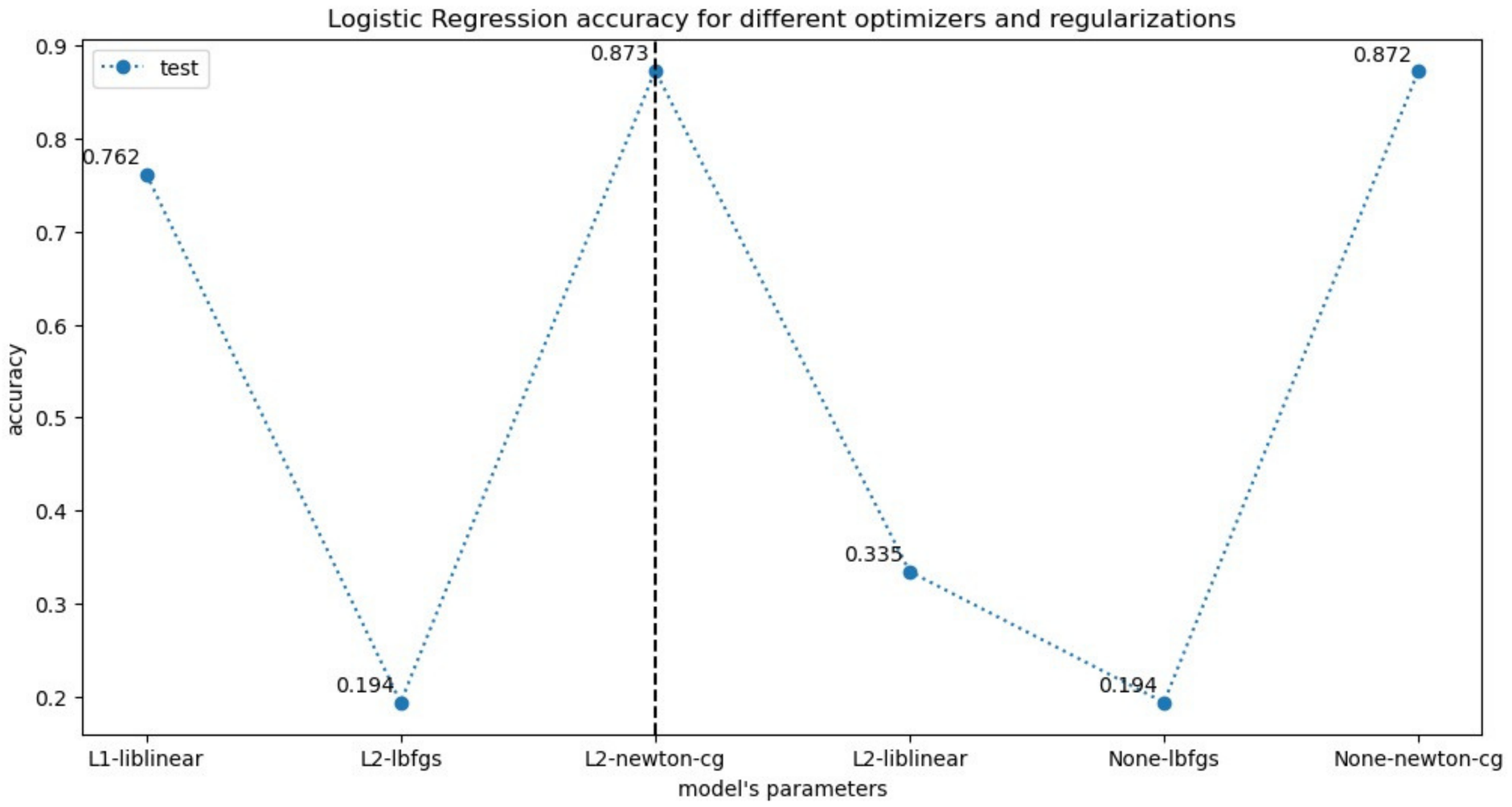
- **+** Easy to understand and interpret
- **+** Efficient for small datasets and quickly trained
- **+** provides probabilistic outputs between 0 and 1
- **−** Struggles with complex data patterns
- **−** Challenges in multi-class classification
- **−** Performance heavily depends on quality of features

LOGISTIC REGRESSION IN OUR PROJECT

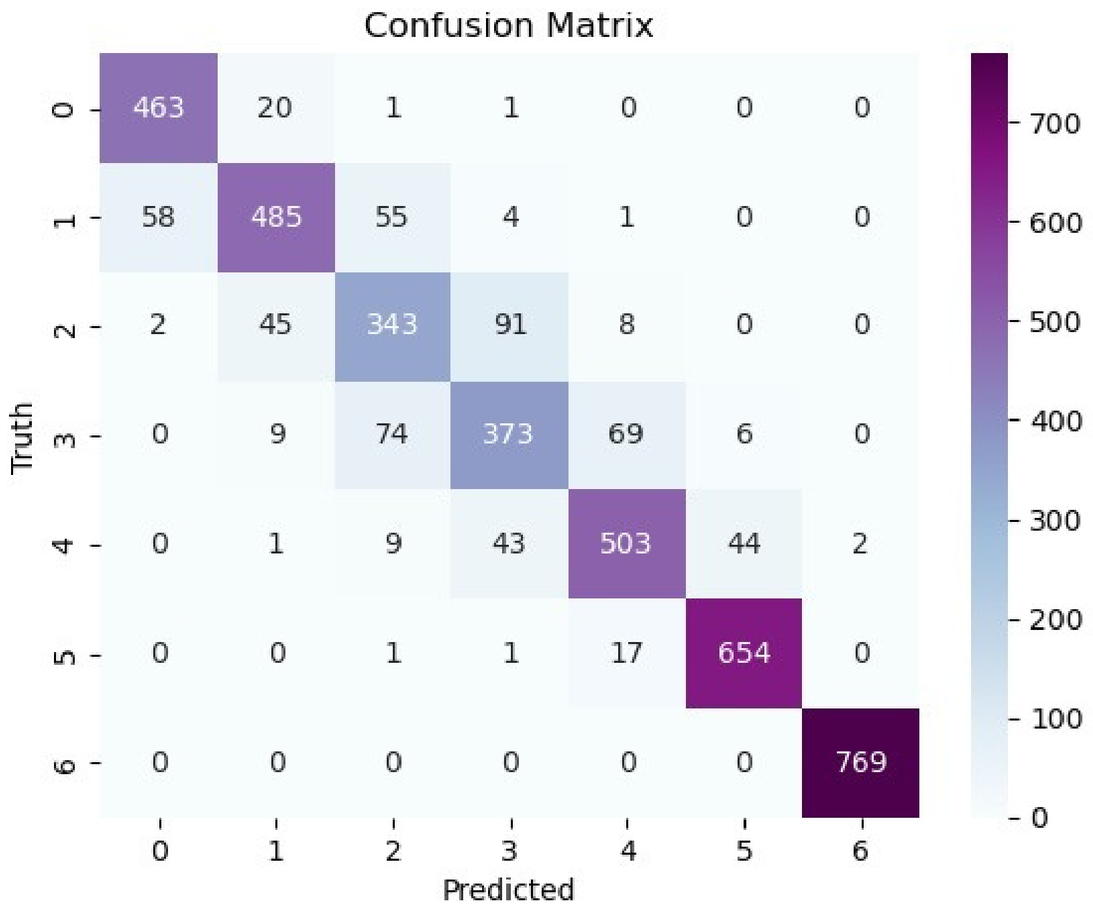
penalty = l2 (Ridge)

solver =newton-cg





LOGISTIC REGRESSION RESULTS



Accuracy	**0.873**
Presicion	**0.859**
Recall	**0.860**
F1-Score	**0.859**

XGBOOST MODEL:

XGBoost stands for “eXtreme Gradient Boosting”.

Provide a high-performance implementation of gradient boosting

Designed for speed and performance.

Handles complex datasets

Helps in reducing overfitting

Maintains fast processing speeds

OPTUNA:

A New Generation Hyperparameter Optimization Framework

BENEFITS

- Streamlines model optimization
- Enhances model performance
- Supports visualization for analysis.

FEATURE

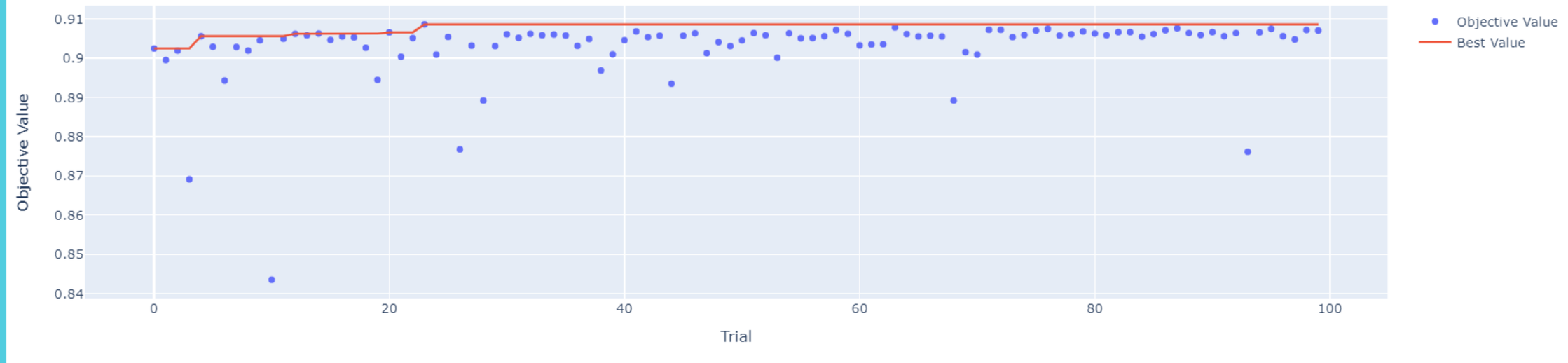
Offers efficient and flexible architecture for optimization

- Automated search for the best hyperparameters
- Visualization tools
- Easy parallelization.

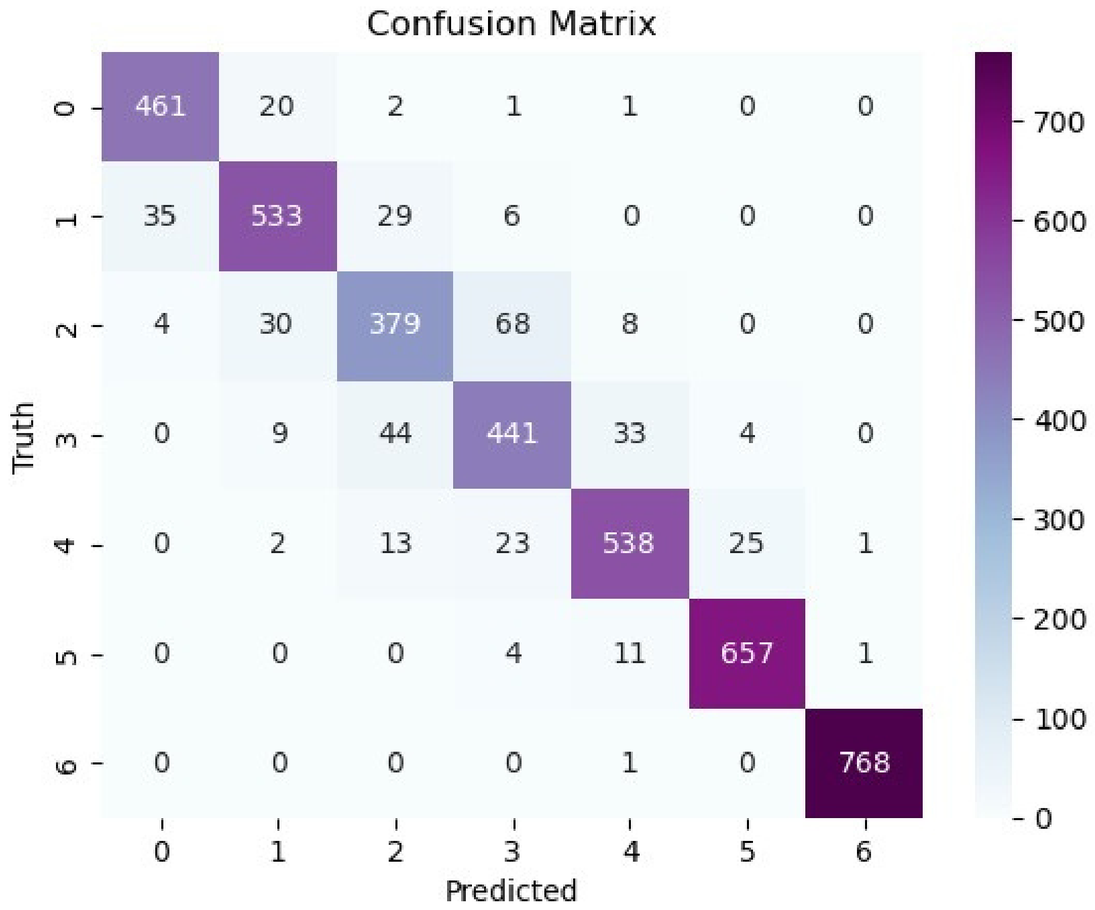
```
booster= 'gbtree',
objective= 'multi:softmax',
num_class= 7,
n_estimators= 249,
eval_metric= 'merror',
min_split_loss= 0.07562076566063626,
learning_rate= 0.1324886494002871,
max_depth= 6,
subsample= 0.695696838500120 8,
colsample_bytree= 0.639717584811432 1,
min_child_weight= 4,
reg_lambda= 3.399309838330835 4e-05,
reg_alpha= 3.387151632412493 7
```

XGBOOST IN OUR PROJECT

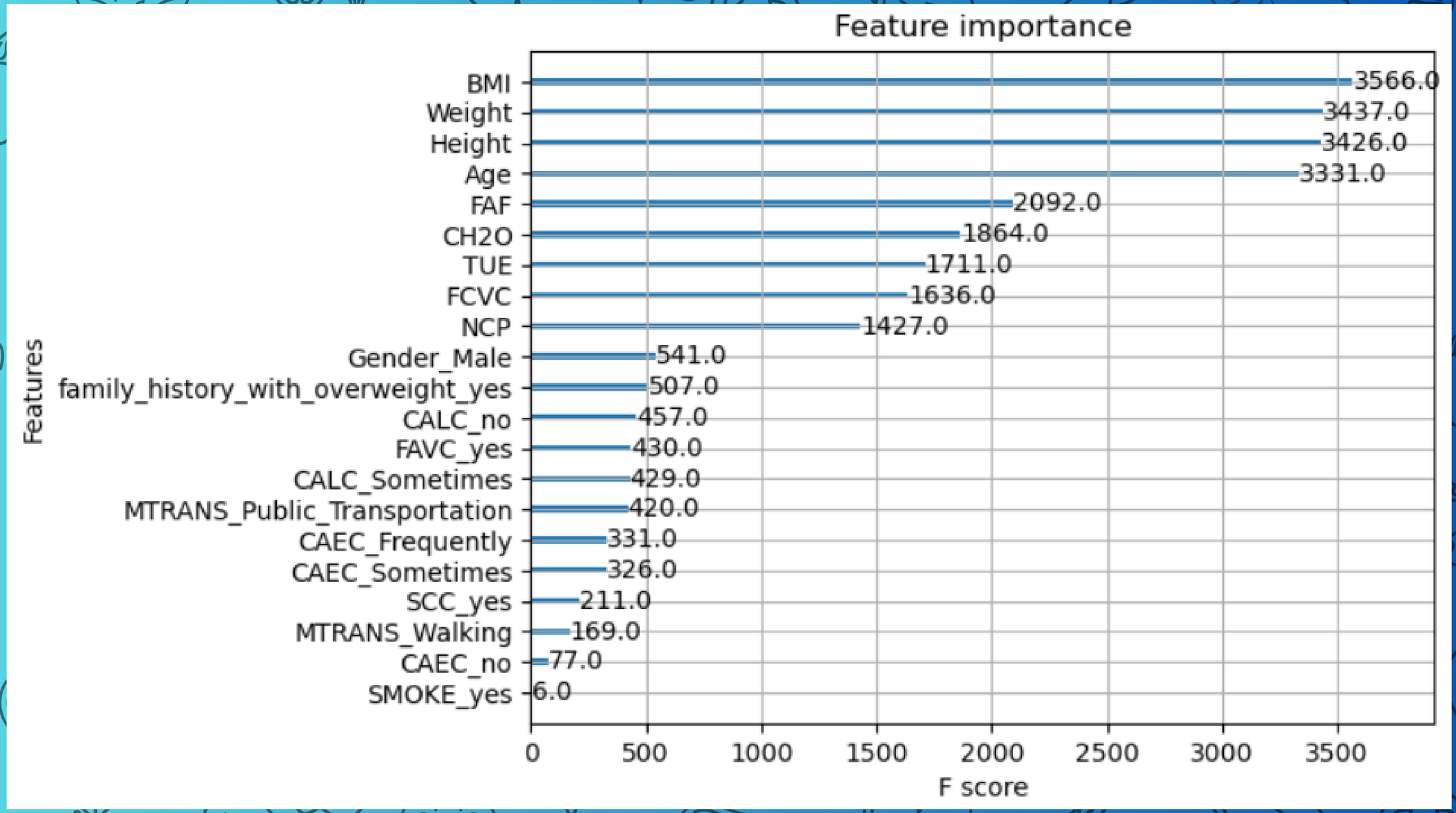
The best hyperparameters



XGBOOST RESULTS

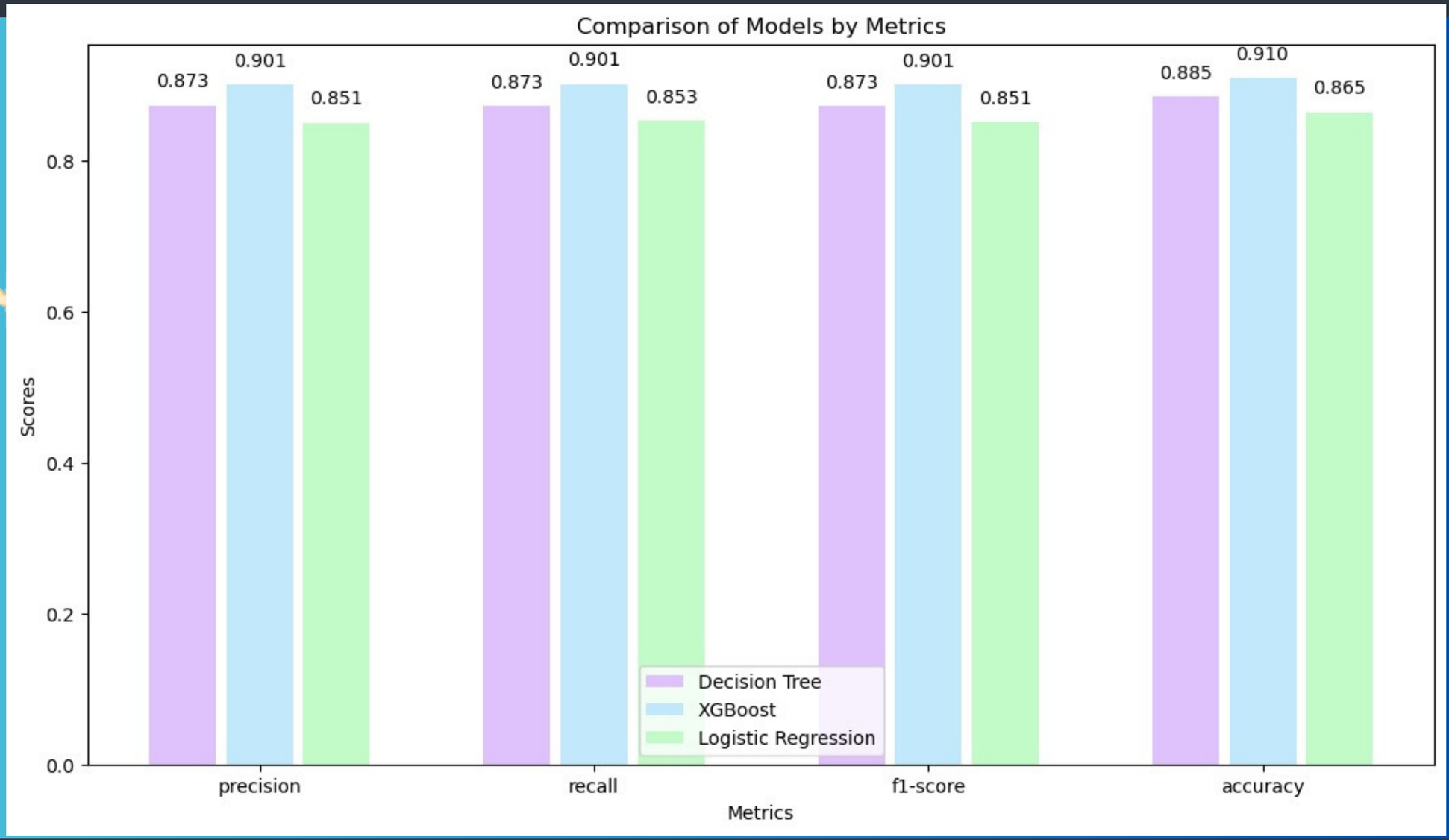


Accuracy	**0.910**
Presicion	**0.901**
Recall	**0.900**
F1-Score	**0.901**


Feature importance
BMI
Weight
Height
Age
FAF
CH2O
TUE
FCVC
NCP
Gender_Male
family_history_with_overweight_yes
CALC_no
FAVC_yes
CALC_Sometimes
MTRANS_Public_Transportation
CAEC_Frequently
CAEC_Sometimes
SCC_yes
MTRANS_Walking
CAEC_no
SMOKE_yes
3566.0
3437.0
3426.0
3331.0
2092.0
1864.0
1711.0
1636.0
1427.0
541.0
507.0
457.0
430.0
429.0
420.0
331.0
326.0
211.0
169.0
77.0
6.0
Features
0
500
1000
1500
2000
2500
3000
3500
F score

FINAL RESULTS

MODELS COMPARISION



LITERATURE REVIEW

1. **"Estimation of Obesity Levels Based on Dietary Habits and Physical Condition Using Computational Intelligence", 2022:**
- Analyzed data pre-simulation
- Used LGBM, XGB, RF, DT, ET, LR
- LGBM most effective with 97.45% accuracy

kaggle:

best accuracy: 92.341%

avg LR: 88%

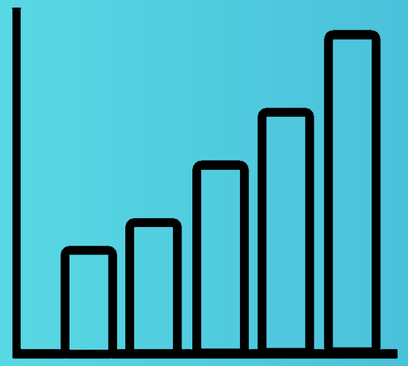
2. "A Machine Learning Approach for Obesity Risk Prediction", 2021:
- Targets Bangladesh population
- Used K-NN, SVM, LR, NB, RF, DT, AdaB, MLP, GB
- LR most accurate with 97.09% accuracy

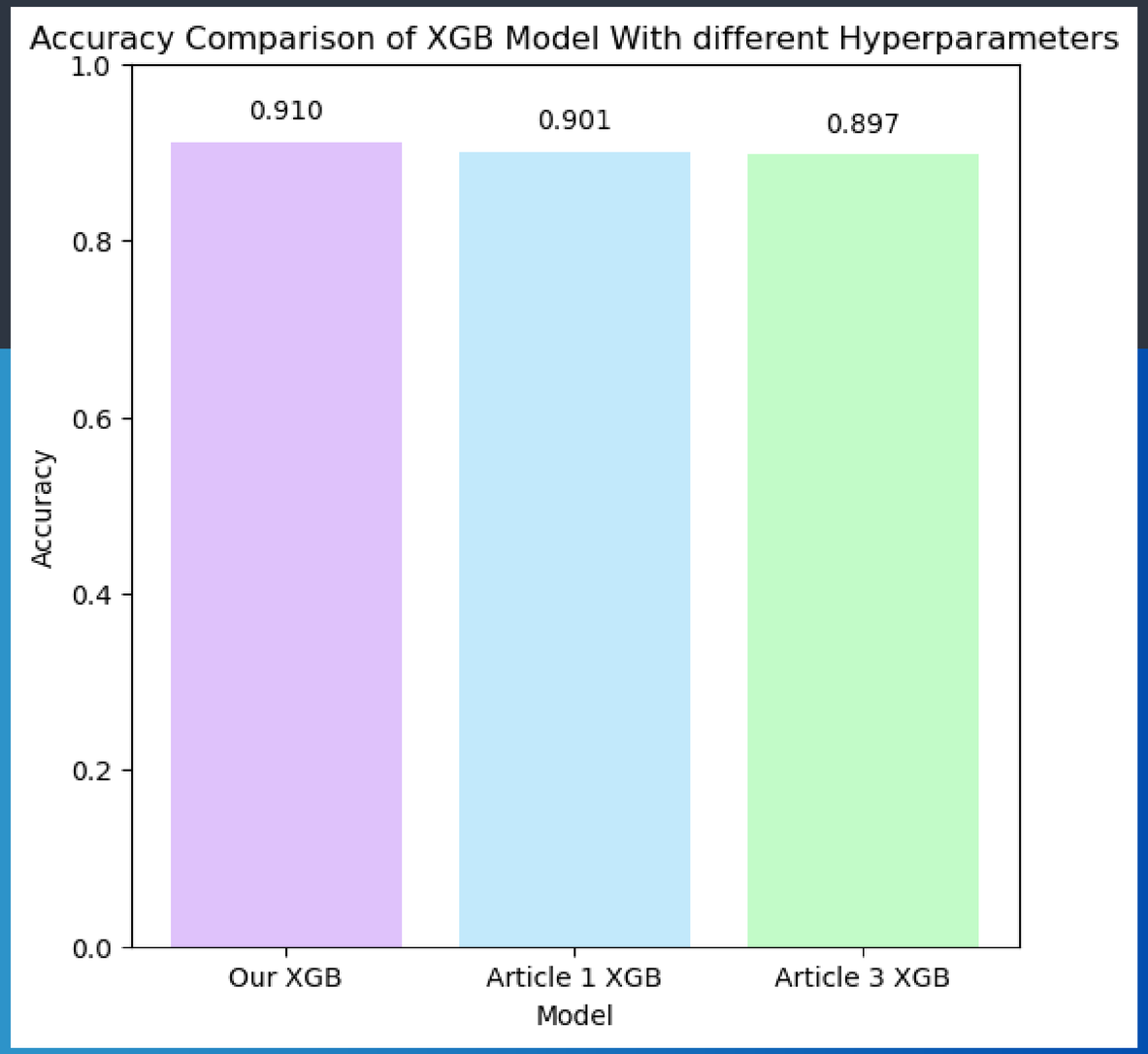
3. **"Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits", 2023:**
- Utilized data from 498 participants aged 14-61
- Explored LR, RF, XGB
- LR achieved highest accuracy at 98.79%

COMPARISON OF ARTICLES

Comparing XGB model across the reviewed articles and our project

Regarding XGB model, our hyper-parameters selection was the best fit for our dataset





CONCLUSION

- XGBoost model showed superior accuracy at 91%
- Demonstrates effectiveness of ensemble learning
- Significant correlations found between obesity and variables
- Highlights importance of considering multiple factors

REFERENCES

article 1

Santisteban Quiroz, J. P. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. Informatics in Medicine Unlocked, 29, 100901.

article 2

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. Current Research in Behavioral Sciences, 2, 100053.

article 3

Gozukara Bag, H. G., Yagin, F. H., Gormez, Y., González, P. P., Colak, C., Gülü, M., Badicu, G., & Ardigò, L. P. (2023). Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits. Diagnostics (Basel), 13(18), 2949.

Thank you