AFEKA The Academic College of Engineering in Tel Aviv

Software Engineering

Project Name:

Multi-class prediction of obesity risk

Final Project Report

Student Name: Noy Tsafrir
ID Number 208013755

Student Name: Shahar Felman
ID Number 206345282

Student Name: Tom Mandel
ID Number 205633688

Kaggle Competition: Multi-Class Prediction of Obesity Risk

Course: Computational Learning, 10230

Supervisor's Name: Dr. Yehudit Aperstein

Submission Date: 17/04/2024

1. Introduction

Obesity is a chronic complex disease defined by excessive fat deposits that can impair health. Obesity can grow in both adults and children, men, and women. Obesity is a major cause of health problems and is correlated with various diseases and conditions, particularly cardiovascular diseases, type 2 diabetes, obstructive sleep apnea, certain types of cancer, and osteoarthritis. People are classified as obese when their body mass index (BMI), a person's weight divided by the square of the person's height, is over 30 kg/m2.

There are many reasons why some people have trouble losing weight. Often, obesity results from inherited, physiological, and environmental factors, combined with diet, physical activity, and exercise choices. 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.

The aim of this project is to raise awareness about the risk of obesity through key lifestyle factors. To achieve this, we will use Decision Tree, XGBoost, and Linear Regression models to analyze our dataset, which comes from a deep learning model trained on data related to Obesity or CVD risk. Our simulated dataset contains 16 features and 20,758 records while the original dataset, contains only 2111 records, focuses on estimating obesity levels among individuals aged 14 to 61 from Mexico, Peru, and Colombia, based on their eating habits and physical conditions, collected through a web-based survey.

2. Literature Review

We summarize three academic articles that utilize computational intelligence and machine learning to predict or understand obesity-related factors.

"Estimation of Obesity Levels Based on Dietary Habits and Physical Condition Using Computational Intelligence" explores obesity prediction using machine learning models. The study, which analyzed the same data as our original dataset (prior to the data simulation process in our dataset), applied several machine learning techniques including Light Gradient Boosting Machine (Light GBM), XG Boost, random forest (RF), decision tree (DT), Extremely Randomized Trees (ET), and logistic regression (LR). The study identified the Light GBM model as most effective, achieving an impressive accuracy rate of 97.45%.

"A Machine Learning Approach for Obesity Risk Prediction" targets the population of Bangladesh, employing over 1100 records dataset to predict obesity risk with machine learning algorithms such as K-NN, SVM, Logistic Regression, Naïve Bayes, Random Forest, Decision Tree, ADA Boost, MLP and Gradient Boost. Logistic regression was the most accurate model, with 97.09% accuracy rate.

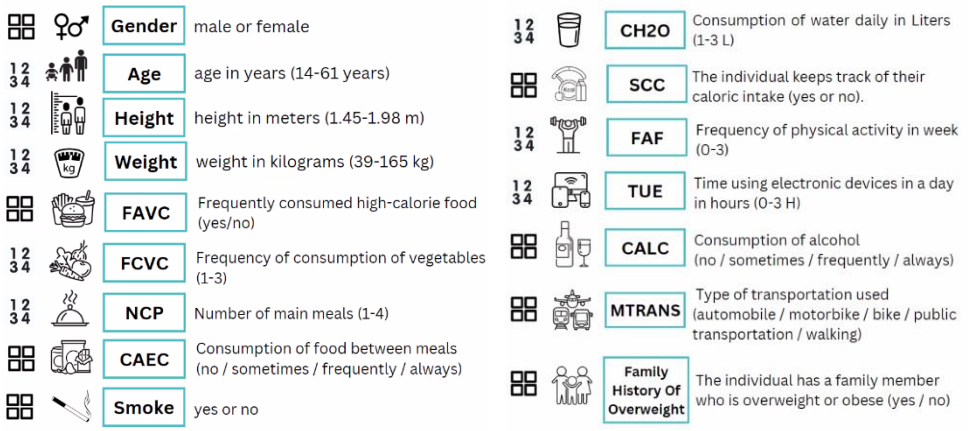
"Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits" utilized encompassed information from 498 participants aged between 14 and 61 years, focusing on their eating habits and physical activity levels to estimate obesity levels. The research explored various machine learning models like the logistic regression (LR) model, Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The LR achieved the highest accuracy rate of 98.79%.

3. Methodology

3.1 Data Exploration

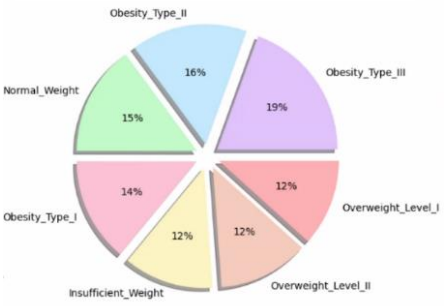
First, we explored the data to get first understanding of the predictors, the response, and their distribution. The predictors in our dataset as described in Figure 1.

Figure 1: Predictors included in the dataset



The Response value named “NObeyesdad” is a multi-class value divided to 7 classes. Furthermore, Figure 2 illustrates the balanced distribution of classes.

Figure 2: Classes Description



3.2 Data Preprocessing

First, we made sure that our dataset did not have any missing information, making sure everything was complete. Next, we worked on the categorical predictors, by using the ‘get_dummies’ function and by labeling the response variable after ordering the classes. These changed them into a format that our machine learning models could use easily. We also removed the id column because it didn’t add relevant information related to the subject. We added a new variable for the BMI because it's important for understanding obesity levels. This made it easier to see how the numbers in our data match up with different obesity categories.

We used 80% of the dataset for training and the remaining 20% for testing.

3.3 Models

3.3.1 Decision Tree Model:

Decision trees are hierarchical structures resembling flowcharts. Their advantages in multi-classes classification include simplicity, interpretability, efficient handling of numerical and categorical data. We start by using a simple DT loaded with the default parameters to receive a base line of prediction accuracy. Then we tried to avoid overfit with pruning functions and different values of alpha. The best alpha parameter was selected using cross validation.

3.3.2 XGBoost Model:

XGBoost stands for "Extreme Gradient Boosting." It reflects the model's foundation in gradient boosting, an ensemble learning technique, fine-tuned with various optimizations. Its key features include achieving high accuracy, preventing overfitting, handling complex datasets, and maintaining fast processing speeds throughout. Based on previous works (both on Kaggle competition’s notebooks and referenced articles), We selected this model as we saw the high accuracy scores and decided to make an extra effort to optimize its hyper-

parameters to try achieving a higher prediction accuracy. The hyper-parameters were selected using Optuna which is a framework for hyper-parameters optimization (with cross-validation integrated).

3.3.3 Logistic Regression Model:

Logistic regression is used for predicting the probability of different classes in a classification problem. For multi-class classification, A softmax functionis used to handle multiple classes. Its simplicity, flexibility, interpretability through probability outputs, and efficient performance were key reasons for selecting this model along with the wide use of it in other works. We fitted the model using standard hyper-parameters and then we preform cross-validation to ensure the accuracy score of the test.

The performance of the models was evaluated using multiple classification metrics: Accuracy, Precision, F1-score, and Recall. After fitting the models and testing them as previously described, we proceeded to fit some of them using hyperparameters outlined in one of the referenced articles (described in Literature Review) for the purpose of comparing results.

4. Results

4.1 Hyper-parameters

4.1.1 Decision Tree Model:

Using the default hyperparameters, the DTC model achieved an accuracy score of 84.2%. Figure 3 illustrates the accuracy as a function of alpha (the pruning parameter) following the application of the cost complexity path function. Figure 4 offers insight into the improvements in overfitting by comparing test accuracy to train accuracy. A value of α=0.00025 was determined to be the best alpha, resulting in an accuracy of 88.29%.

Figure 3: Accuracy per alpha pruning param

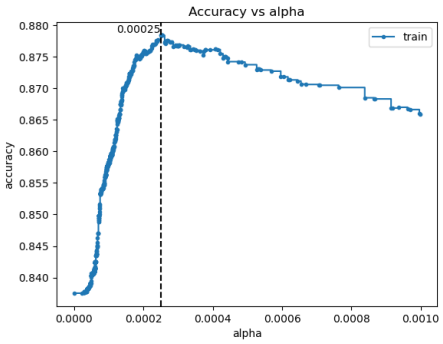
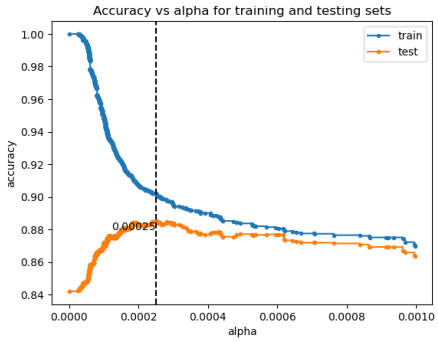


Figure 4: Accuracy of train vs test



4.1.2 XGBoost Model:

A study using 100 Optuna trials was conducted, with the best hyperparameters outlined in Figure 6. Figure 5 illustrates the scores throughout the process.

Figure 5: Optuna trials accuracy

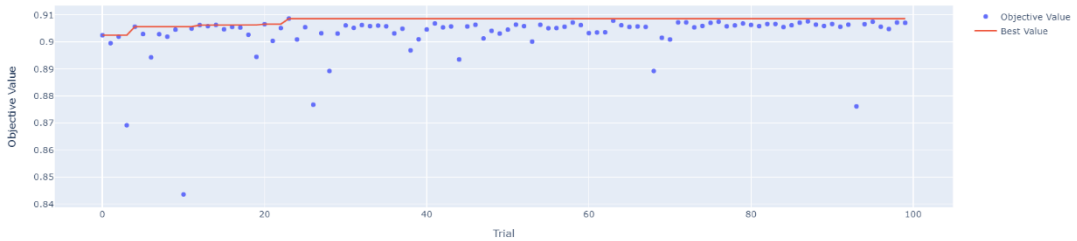


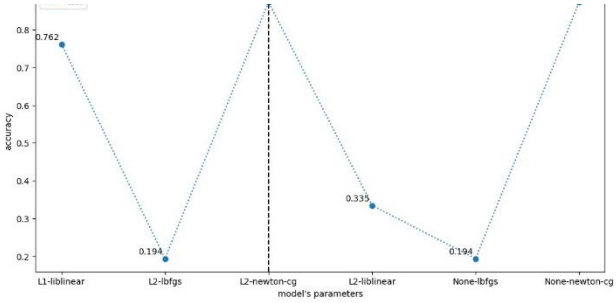
Figure 6: XGBoost hyper-parameters

```
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= 0.6956968385001208,
colsample_bytree= 0.6397175848114321,
min_child_weight= 4,
reg_lambda= 3.3993098383308354e-05,
reg_alpha= 3.3871516324124937
```

4.1.3 Logistic Regression Model:

We examined and evaluated several combinations of optimizers (solvers) and regularizations (penalties) using cross-validation. Figure 7 illustrates the differences observed. The selected solver is 'newton-cg' with penalty 'L2' (Ridge).

Figure 7 : LR accuracy with different hyperparameters



4.2 Results comparison

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	**0.908**	**0.899**	**0.898**	**0.898**
Decision Tree	0.885	0.873	0.873	0.873
Logistic Regression	0.873	0.859	0.860	0.859

Table 1: Model performance evaluation

Table 1 displays the performance metrics for each model as outlined in our methodology. It is evident from both Table and Figure 8: Modesl performance bar chart that XGBoost consistently outperformed the other models across all metrics. Conversely, the Logistic regression technique yielded the lowest performance among the three models assessed.

Figure 8: Modesl performance bar chart

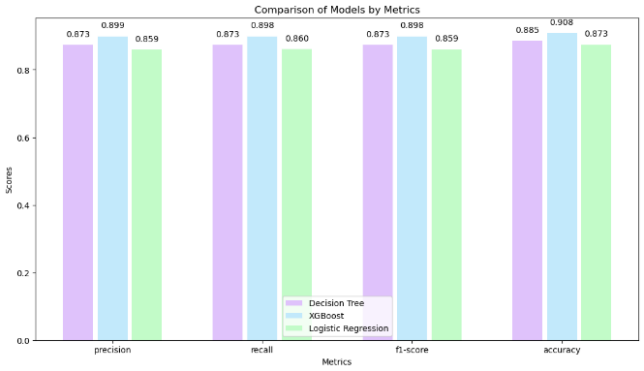


Figure 9: XGBoost parameters comparison

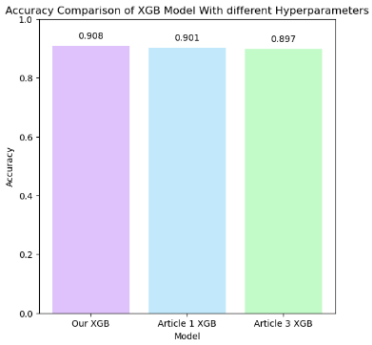


Figure 9 illustrates our comparison of hyperparameters for the XGBoost model with those from other articles. It is evident that the hyperparameters we identified were the most suitable for our dataset.

4.3 Feature Importance

Figure 10 presents the feature importance rankings from our XGBoost model.

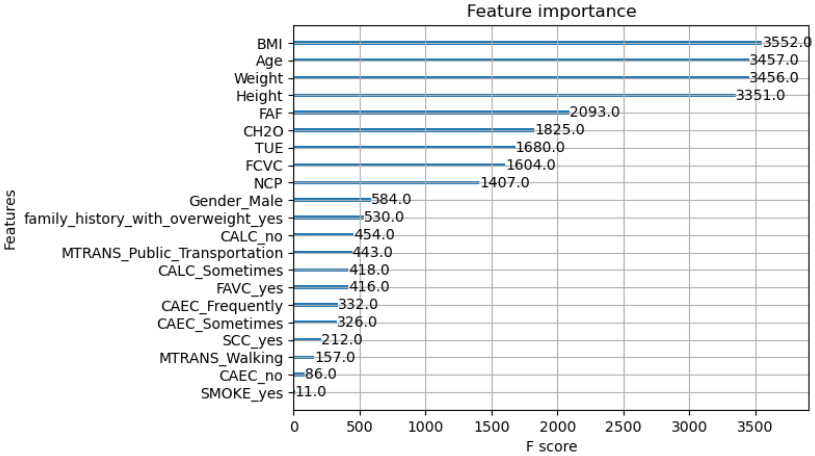


Figure 10: XGBoost feature importance

The top determinants of obesity risk were BMI, Age, Weight, and Height, with F scores notably higher than other variables. BMI was the leading indicator, emphasizing the model's prioritization of physiological factors in obesity risk assessment. Lesser yet significant features included physical activity and daily water intake. Gender and family history also contributed to the model's predictive accuracy, reflecting the multifaceted nature of obesity.

5. Discussion

Our study addresses the global health concern of rising overweight populations by utilizing machine learning to predict obesity risk. We tested various models, including Decision Tree, Logistic Regression, and XGBoost, on an enriched dataset. Comparing our results with prominent research articles reveals varying accuracies, emphasizing methodological nuances.

Optimizing hyperparameters was pivotal in refining model performance. Through rigorous optimization, we identified hyperparameters that ensured optimal predictive accuracy, particularly in our XGBoost model, which emerged as the most effective.

Our findings were benchmarked against the Kaggle "Multi-Class Prediction of Obesity Risk" competition, contributing to its objectives, and leveraging its collaborative environment for model refinement. Utilizing the Kaggle dataset, our robust approach to model tuning led to noteworthy results. Despite achieving an accuracy slightly below the best leaderboard entry, at 90.8%, our hyperparameter optimization and model selection strategies proved effective.

These results underscore machine learning potential in enhancing obesity prediction and facilitating better health management, advocating for its continued evolution in public health applications.

6. Conclusion

Our findings reveal that XGBoost emerged as the superior model, boasting an impressive accuracy of 90.8%. This underscores the effectiveness of ensemble learning in handling complex relationships within the dataset.

Through our analysis, we observed significant correlations between obesity and certain variables, emphasizing the importance of considering multiple factors in understanding and predicting obesity.

Moving forward, to further enhance our model's performance, we recommend experimenting with the Light Gradient Boosting Machine (LGBM). This model has demonstrated high performance in our Kaggle competition, indicating its potential suitability for our dataset. Additionally, considering different loss functions could provide new insights into the learning process and improve predictive accuracy. Finally, expanding our feature engineering efforts will likely uncover more intricate patterns and relationships within the data.

7. Appendices

- 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.

- 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.

- 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.