Kaggle Competitions: Obesity Risk Prediction

Machine Learning Project Report

June 5, 2024

Contents

Authors

Urszula Szczęsna, Mateusz Deptuch and Gaspar Sekula

1 Introduction

This report presents our approach to predicting obesity risk using various machine learning models. The complete project, including all data, code, and additional resources, is available in our GitHub repository.

2 Business objective

The aim is to predict the obesity risk of patients basing on their general health and lifestyle features. Our model may be used by doctors in hospitals and clinics as well as regular people caring about their well-being.

3 Data

3.1 Source

The data comes from Multi-Class Prediction of Obesity Risk, Kaggle. The data consists of 2 files:

- `train.csv` - the training dataset; NObeyesdad is the categorical target,
- `test.csv` - the test dataset; the objective is to predict the class of NObeyesdad for each row.

3.2 Train, test and validation data frames

The model will be built, tested and validated on `train.csv` data frame, which consists of total 20758 rows. The table was split as follows:

- 70% - train data,
- 20% - validation data,
- 10% - test data.

Finally, the model will be tested on `test.csv` data frame.

4 Target class

The target class is a type of obesity (status or level). Possible classes are:

- Obesity type I,
- Obesity type II,
- Obesity type III,
- Insufficient weight,
- Normal weight,
- Overweight level I,
- Overweight level II.

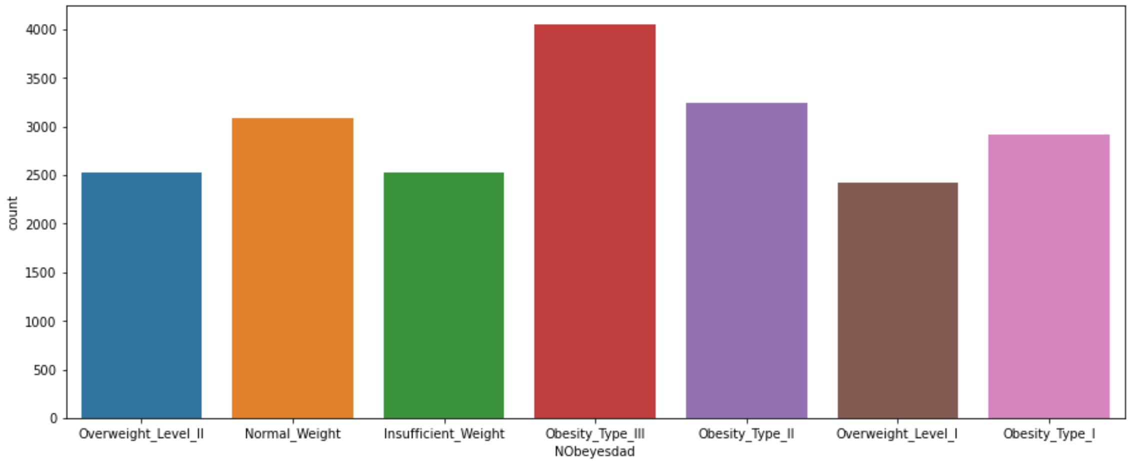


Figure 1: Number of target class types in training data.

The most popular value is Obesity type III (19%) while the least popular is Overweight level I (12%). The class sizes in the dataset show slight differences but are overall fairly balanced.

5 Features

Every observation is described by 17 feature columns and 1 class column which identifies the type or level of obesity:

- ID,
- Gender (male, female),
- Age,
- Height (in meters),
- Weight (in kilograms),
- Family history with overweight (whether the individual has a family history of overweight or obesity),
- FAVC - Frequent consumption of high-caloric food,
- FCVC - Frequency of consumption of vegetables,
- NCP - Number of main meals,
- CAEC - Consumption of food between meals,
- SMOKE - whether the individual is a smoker or not,
- CH20 - the amount of daily water consumption,
- SCC - Calories consumption monitoring,
- FAF - Physical activity frequency,
- TUE - Time using technology devices,
- CALC - Consumption of alcohol,
- MTRANS - mode of transportation used by the individual (Public transportation, Automobile, Walking, Motorbike, Bike)

6 New features

During the feature engineering process, we created a new feature called BMI, which stands for Body Mass Index. BMI is calculated as the weight in kilograms divided by the square of the height in meters (kg/m^2). This new feature combines height and weight, which are pretty correlated in our dataset, to provide a single metric that quantifies the body fat of individuals. In the real world, BMI is a widely used index that effectively indicates levels of obesity.
We decided to drop only the 'ID' column, as it does not provide any meaningful information. For the rest of the columns, we chose to retain them, as their variance is equal to 0.

7 Modelling

7.1 Standard and advanced models

We decided to evaluate 9 standard and advanced models and then select the top performers for our Stacking classifier. The models we trained included: Random Forest, Decision Tree, SVC, Logistic Regression, Naive Bayes, XGBoost, SGD Classifier, K-Nearest Neighbors, and Ada Boost. All models have been hyperparameterised and afterwards cross-validated. The accuracy scores are summarized in Table 1. The scores are calculated for the models with the best parameters. Column 'Train accuracy' presents the accuracy scores on train data, 'Test accuracy' - on test data.

Model	**Train Accuracy**	**Test Accuracy**
Random Forest	0.89	0.88
Decision Tree	0.86	0.86
SVC	0.86	0.86
Logistic Regression	0.85	0.85
Naive Bayes	0.74	0.75
XGBoost	0.89	0.89
SGD	0.73	0.75
K-Neighbors	0.73	0.74
AdaBoost	0.68	0.56
Stacking	0.90	0.88
TPOT	0.89	0.88

Table 1: Model performance.

Conlcusions:

- Our criterion of choice is accuracy score on test data as well as cross-validation mean accuracy score on test data. Hence, the best model is the XGBoost.
- It is worth mentioning, that cross-validation process has shown overfitting in only one of the models - AdaBoost.
- Only a few models performed well, this problem appeared to be difficult for K-Nearest Neighbors, SGD and AdaBoost.

7.2 AutoML

We have also tried out the AutoML model - TPOT Classifier. The results were not way better than in best standard and advanced models (see Table 1).
TPOT's best pipeline was: `XGBClassifier(input_matrix, learning_rate=0.5, max_depth=3, min_child_weight=5, n_estimators=100, n_jobs=1, subsample=1.0, verbosity=0)`.
The accuracy scores were as follows:

- Generation 1 - Current best internal CV score: 0.8771
- Generation 2 - Current best internal CV score: 0.8860

- Generation 3 - Current best internal CV score: 0.8860
- Generation 4 - Current best internal CV score: 0.8867
- Generation 5 - Current best internal CV score: 0.8873
- Generation 6 - Current best internal CV score: 0.8881
- Generation 7 - Current best internal CV score: 0.8881
- Generation 8 - Current best internal CV score: 0.8881

It was worthwhile checking the performance of AutoML model, however we will choose the best model from the previous paragraph as the one solving our business problem.

8 Chosen model

8.1 Details

The chosen model for OObesity Risk classification task is the XGBoost with the following parameters:

- subsample = 0.75,
- gamma = 0.5,
- n_estimators = 1000,
- max_depth = 7,
- learning_rate = 0.01,
- colsample_bytree = 0.5.

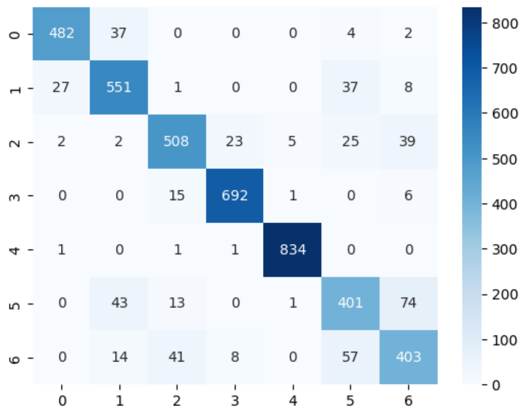
8.2 Accuracy

The model was verified on val data:

Class	**Precision**	**Recall**	**F1-Score**
0: Insufficient Weight	0.94	0.92	0.93
1: Normal Weight	0.85	0.88	0.87
2: Obesity Type I	0.88	0.84	0.86
3: Obesity Type II	0.96	0.97	0.96
4: Obesity Type III	0.99	1.00	0.99
5: Overweight Level I	0.77	0.75	0.76
6: Overweight Level II	0.76	0.77	0.76
Accuracy		0.89	
Macro avg	0.88	0.88	0.88
Weighted avg	0.89	0.89	0.89

Table 2: XGBoost Classification Report on val data.

8.3 Confusion Matrix



	0	1	2	3	4	5	6
0	482	37	0	0	0	4	2
1	27	551	1	0	0	37	8
2	2	2	508	23	5	25	39
3	0	0	15	692	1	0	6
4	1	0	1	1	834	0	0
5	0	43	13	0	1	401	74
6	0	14	41	8	0	57	403

Figure 2: Confusion Matrix for val data.

8.4 Cross-validation

The results were so good, that it was necessary to validate them in cross-validation process. We have decided to run cross-validation in 7 loops. Achieved CV scores on train data were as following: 0.88246628, 0.89547206, 0.87283237, 0.89643545, 0.88921002, 0.89445783 and 0.89590361. Therefore mean CV Train score equals to 0.8895.

8.5 ROC curves

Also, we constructed ROC curves for the model:

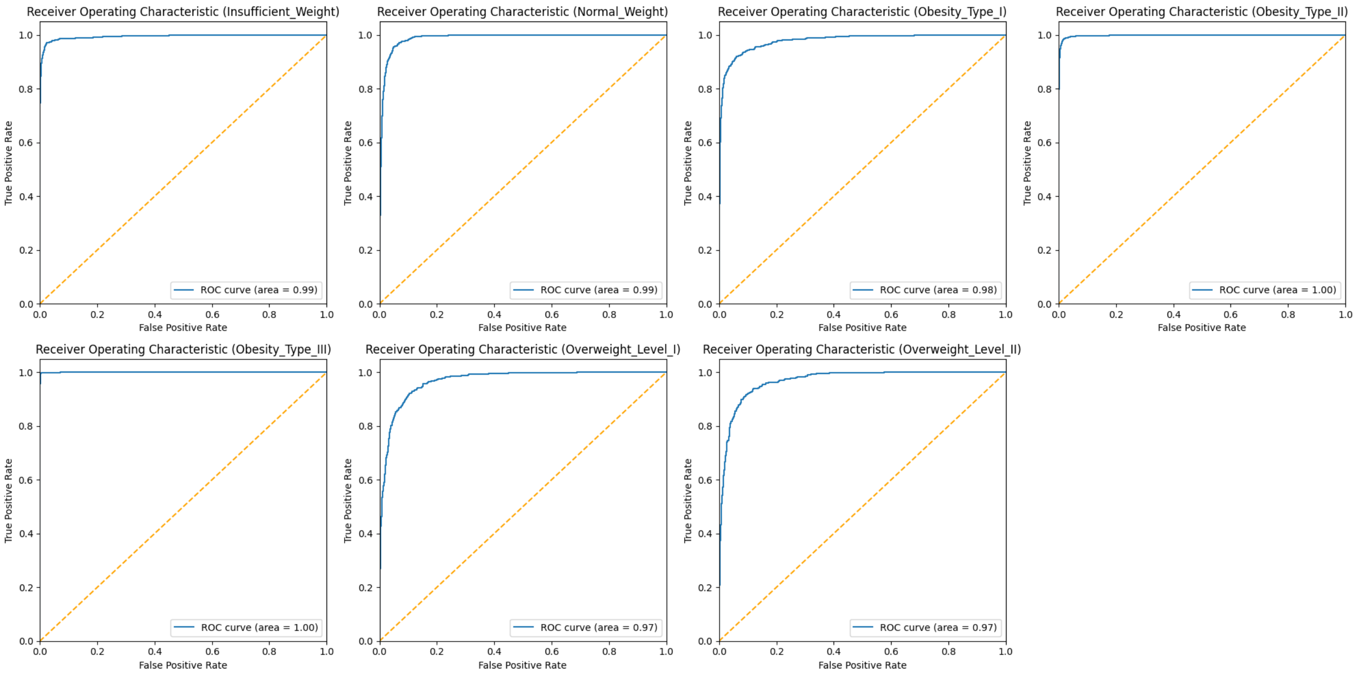


Figure 3: ROC curves for the XGBoost model.

A ROC (receiver operating characteristic) curve is a graphical plot that illustrates the performance of a classifier model at varying threshold values.

The area under the curves for each class, marked as AUC (area under curve) is treated as a measure of goodness and accuracy of the model. The value of the indicator ranges from 0 to 1. The higher the value,the better the model. For each class, the AUC score is at least 0.97, which means the model is very good.
Moreover, a perfect prediction method results in a point at the upper left corner of the ROC space, corresponding to coordinates $(0, 1)$, indicating 100% sensitivity (no false negatives) and 100% specificity (no false positives). This point, $(0, 1)$, is also referred to as perfect classification, which our model for *Obesity Type III* and is super close to it for *Obesity Type II* as well as *Insufficient Weight*.

8.6 SHAP Plots

Shap plots were generated (for each class separately).
SHAP (SHapley Additive exPlanations) calculates the impact specific values have on the outcome.
The plot illustrates the impact of features on predicting the value of each class. Colors indicate whether the variable has a high (purple) or low (green) value for a given observation. The larger the arrow, the greater the impact. The baseline represents the initial prediction value, and $f(x)$ denotes the correct prediction value for the class.

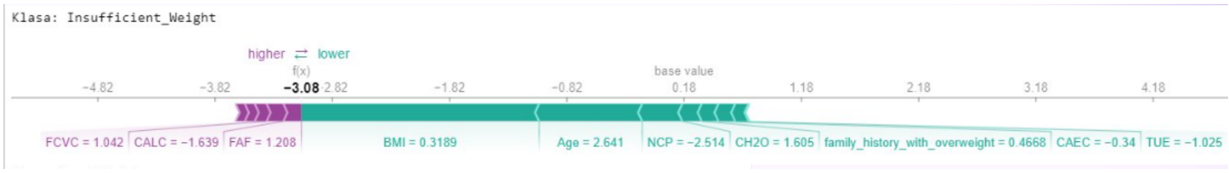


Figure 4: Shap plot for Class Insufficient Weight

Features such as BMI, Age, and number of main meals (NPC) increase the likelihood of being classified to this class. On the other hand frequency of physical activity (FAF) and lower alcohol consumption (CALC) decrease the likelihood.

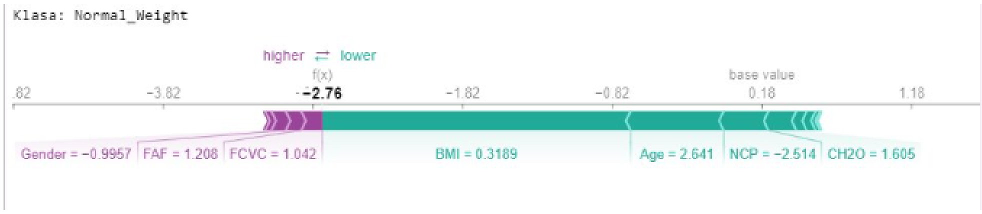


Figure 5: Shap plot for Class Normal Weight

Features such as higher BMI and older age increase the likelihood of being classified as normal weight. Conversely, gender, higher frequency of physical activity (FAF) and higher fast food consumption (FCVC) decrease the likelihood of being classified as normal weight.

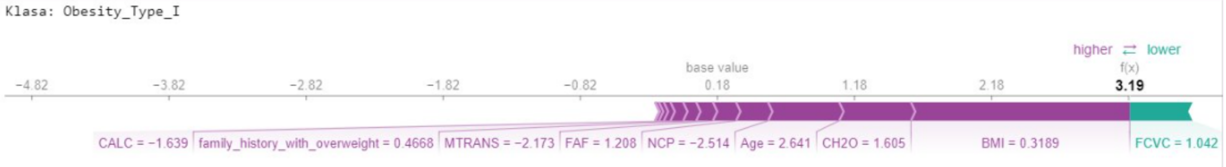


Figure 6: Shap plot for Class Obesity Type I

Higher fast food consumption (FCVC) decreases the likelihood of being classified as "Obesity_Type_I." On the other hand, several features increases this likelihood: lower alcohol consumption (CALC), lower calorie consumption (NCP), older age, higher water intake (CH2O) and higher BMI.

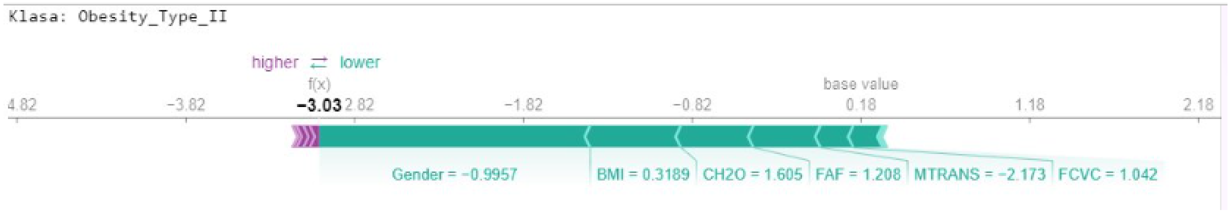


Figure 7: Shap plot for Class Obesity Type II

Gender (being female), higher BMI and higher water consumption (CH20) increase the likelihood of being classified as "Obesity_type_II"

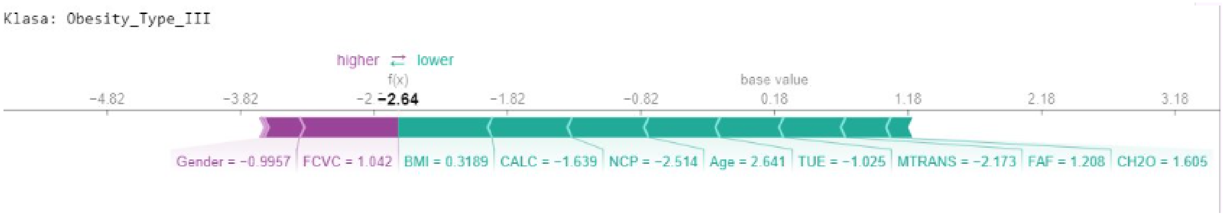


Figure 8: Shap plot for Class Obesity Type III

A higher BMI, older age, and lower alcohol consumption increase the likelihood of being classified as 'Obesity_Type_III.' Conversely, frequent consumption of vegetables (FCVC) and being female decrease the likelihood of being classified into this category.

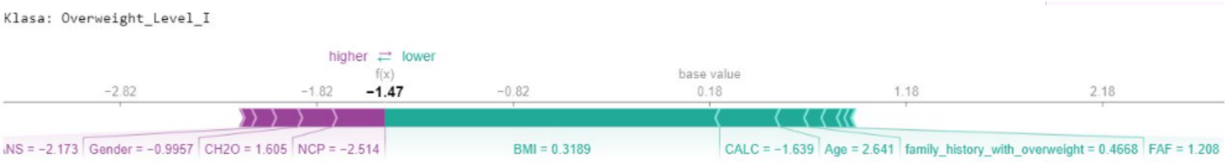


Figure 9: Shap plot for Class Overweight Level I

Having a higher BMI and being older, along with lower alcohol consumption, increase the likelihood of this classification. Conversely, a lower number of meals (NPC) and higher water consumption decrease the likelihood.

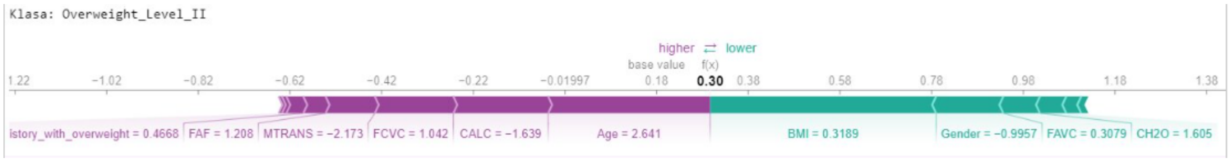


Figure 10: Shap plot for Class Overweight Level II

This example is intriguing because the base value closely aligns with the correct prediction value. A higher BMI and being female shift the prediction towards the left, while older age, lower alcohol consumption, and frequent vegetable consumption shift it towards the right.

9 Tests

The model was verified on test data (10%) and the results are as follows:

Class	Precision	Recall	F1-Score
0: Insufficient Weight	0.92	0.94	0.93
1: Normal Weight	0.88	0.89	0.89
2: Obesity Type I	0.89	0.86	0.88
3: Obesity Type II	0.97	0.97	0.97
4: Obesity Type III	1.00	1.00	1.00
5: Overweight Level I	0.78	0.77	0.77
6: Overweight Level II	0.79	0.80	0.79
Accuracy		0.90	
Macro avg	0.89	0.89	0.89
Weighted avg	0.90	0.90	0.90

Table 3: XGBoost Classification Report on test data.

And the confusion matrix looks as follows:

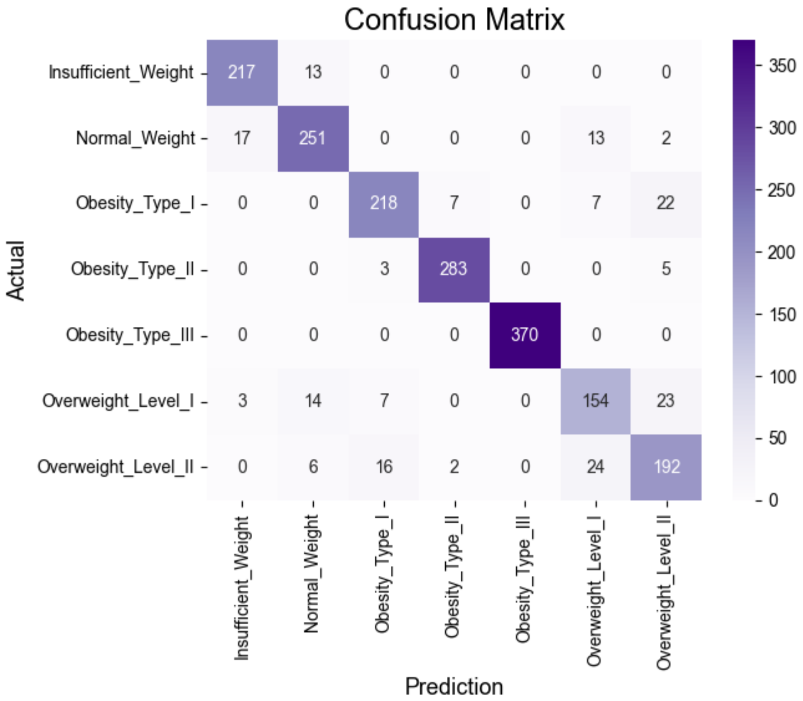


Figure 11: Confusion Matrix for val data.

10 Summary

We have managed to achieve the business objective of the project (to classify Obesity Risk types: build a machine learning model that will classify them). The best model achieves accuracy score around 0.89 points and 0.90 in final test data. The model was validated and works well.