Multi-Class Prediction of Obesity Risk

Doroch Michel 2024

Competition/data overview

- Multi-class classification problem
- Target: 7 classes
- Scoring: accuracy score
- Synthetically-Generated Datasets
- Train set shape: 20758 x 17 (at the beginning)
- Test set shape: 13840 x 17 (at the beginning)
- February 1, 2024 - February 29, 2024

kaggle

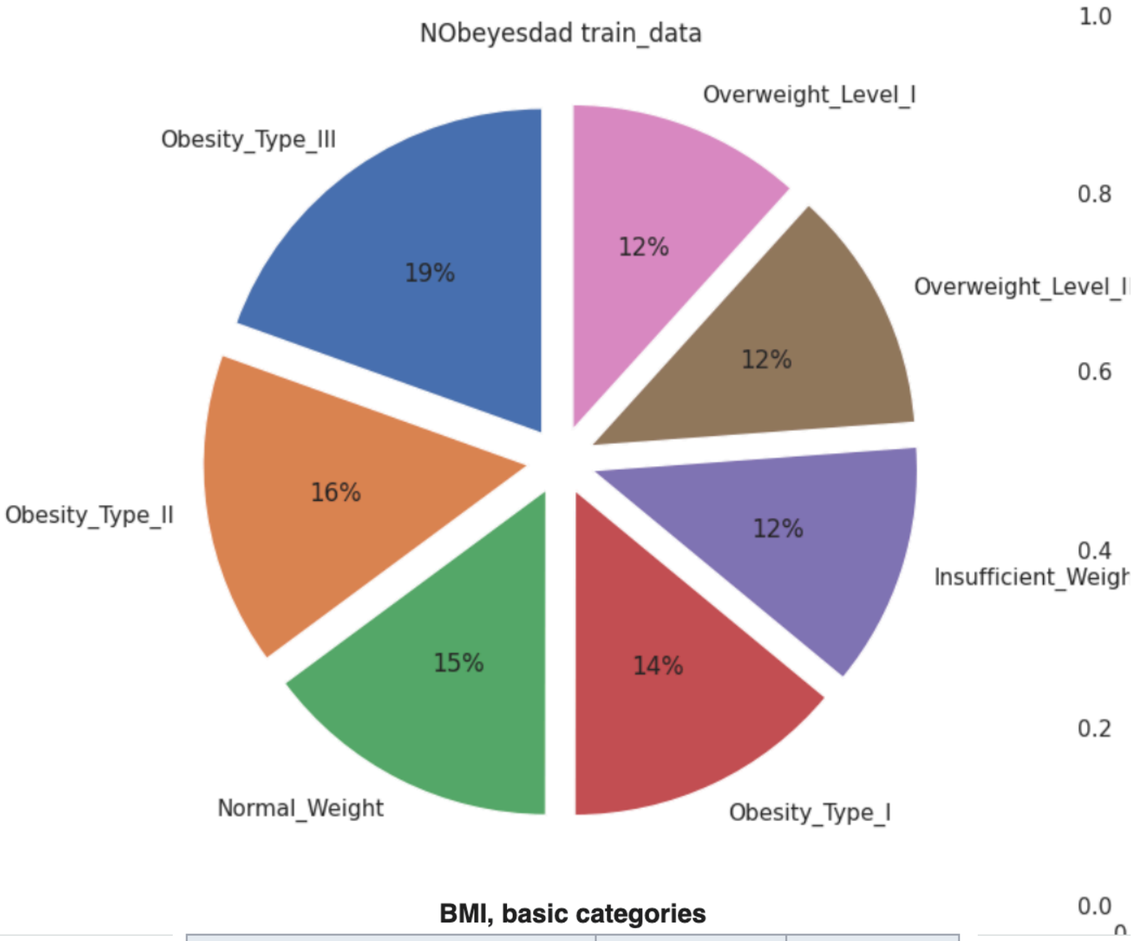
Features

- Gender
- Age
- Height
- Weight
- Frequent consumption of high caloric food (FAVC)
- Frequency of consumption of vegetables (FCVC)
- Consumption of food between meals (CAEC)
- Consumption of water daily (CH2O)
- family_history_with_overweight
- Smoke
- Consumption of alcohol (CALC)
- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Male	24.443011	1.699998	81.669950	yes	yes	2.000000	2.983297	Sometimes	no	2.763573	no	0.000000	0.976473	Sometimes	Public_Transportation	Overweight_Level_II
1	Female	18.000000	1.560000	57.000000	yes	yes	2.000000	3.000000	Frequently	no	2.000000	no	1.000000	1.000000	no	Automobile	Normal_Weight
2	Female	18.000000	1.711460	50.165754	yes	yes	1.880534	1.411685	Sometimes	no	1.910378	no	0.866045	1.673584	no	Public_Transportation	Insufficient_Weight
3	Female	20.952737	1.710730	131.274851	yes	yes	3.000000	3.000000	Sometimes	no	1.674061	no	1.467863	0.780199	Sometimes	Public_Transportation	Obesity_Type_III
4	Male	31.641081	1.914186	93.798055	yes	yes	2.679664	1.971472	Sometimes	no	1.979848	no	1.967973	0.931721	Sometimes	Public_Transportation	Overweight_Level_II

Target

- Underweight
- Normal
- Overweight I
- Overweight II
- Obesity I
- Obesity II
- Obesity III

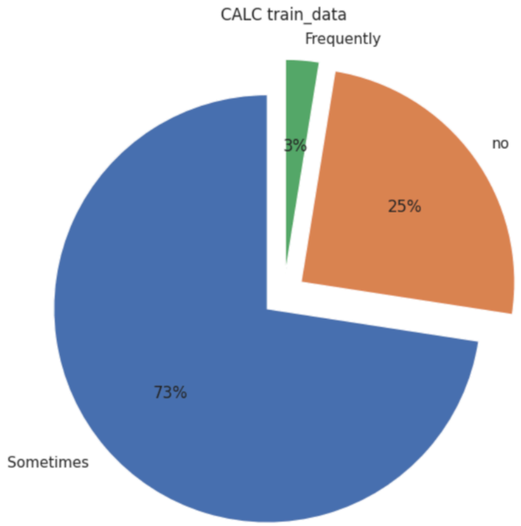


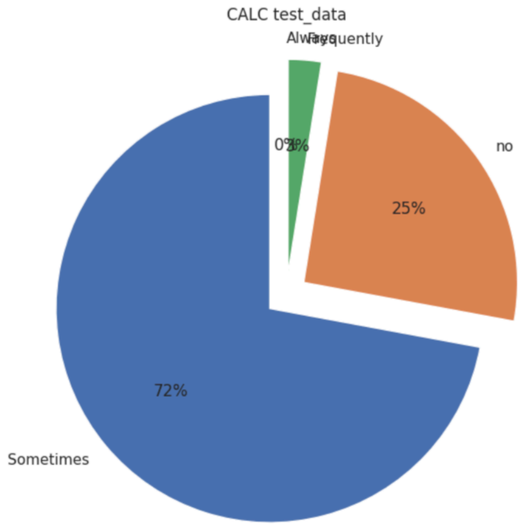
BMI, basic categories

Category	BMI (kg/m^2)[c]	BMI Prime[c]
Underweight (Severe thinness)	< 16.0	< 0.64
Underweight (Moderate thinness)	16.0 – 16.9	0.64 – 0.67
Underweight (Mild thinness)	17.0 – 18.4	0.68 – 0.73
Normal range	18.5 – 24.9	0.74 – 0.99
Overweight (Pre-obese)	25.0 – 29.9	1.00 – 1.19
Obese (Class I)	30.0 – 34.9	1.20 – 1.39
Obese (Class II)	35.0 – 39.9	1.40 – 1.59
Obese (Class III)	≥ 40.0	≥ 1.60

Cat features pie charts

- Consumption of alcohol (CALC)
- test set contains new value ‘Always’






```
[316]: train_data[train_data['CALC'] == 'Always']
```

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad

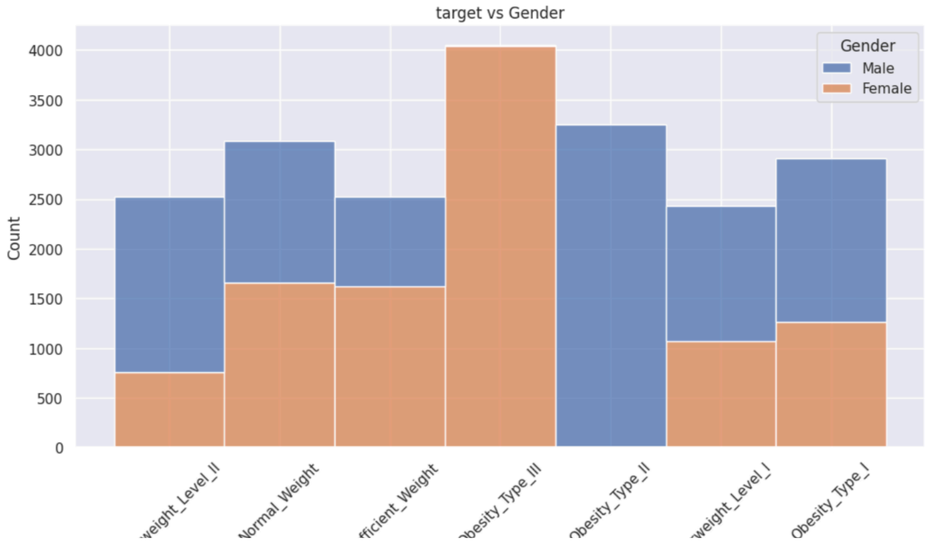
+ Code + Markdown

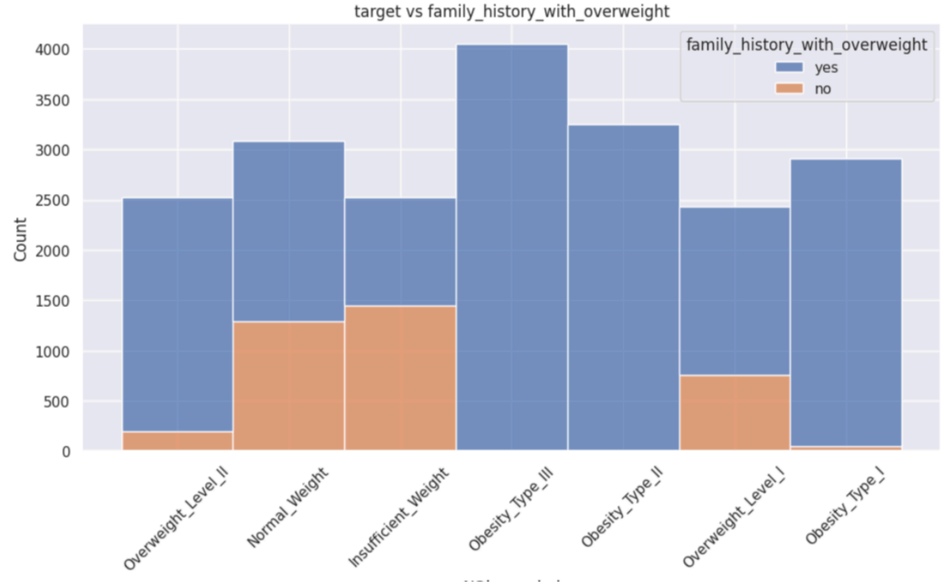
```
[317]: test_data[test_data['CALC'] == 'Always']
```

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS
4023	Female	20.0	1.66	60.0	yes	yes	3.0	3.0	Always	no	2.0	no	0.0	0.0	Always	Public_Transportation
7443	Male	21.0	1.67	66.5	no	yes	2.0	3.0	Frequently	no	2.0	no	1.0	0.0	Always	Public_Transportation

Hist plots cat features

- Target distribution grouped by gender
- Target distribution grouped by family_history_with_overweight





feature engineering

- +5 index features
- $BMI = \frac{Weight}{Height^2}$
- $index\_breightman = \frac{Weight}{Height * 100 * 0.7 - 50}$
- $index\_noorden = \frac{Weight}{Height * 100 * 0.42}$

BMI, basic categories

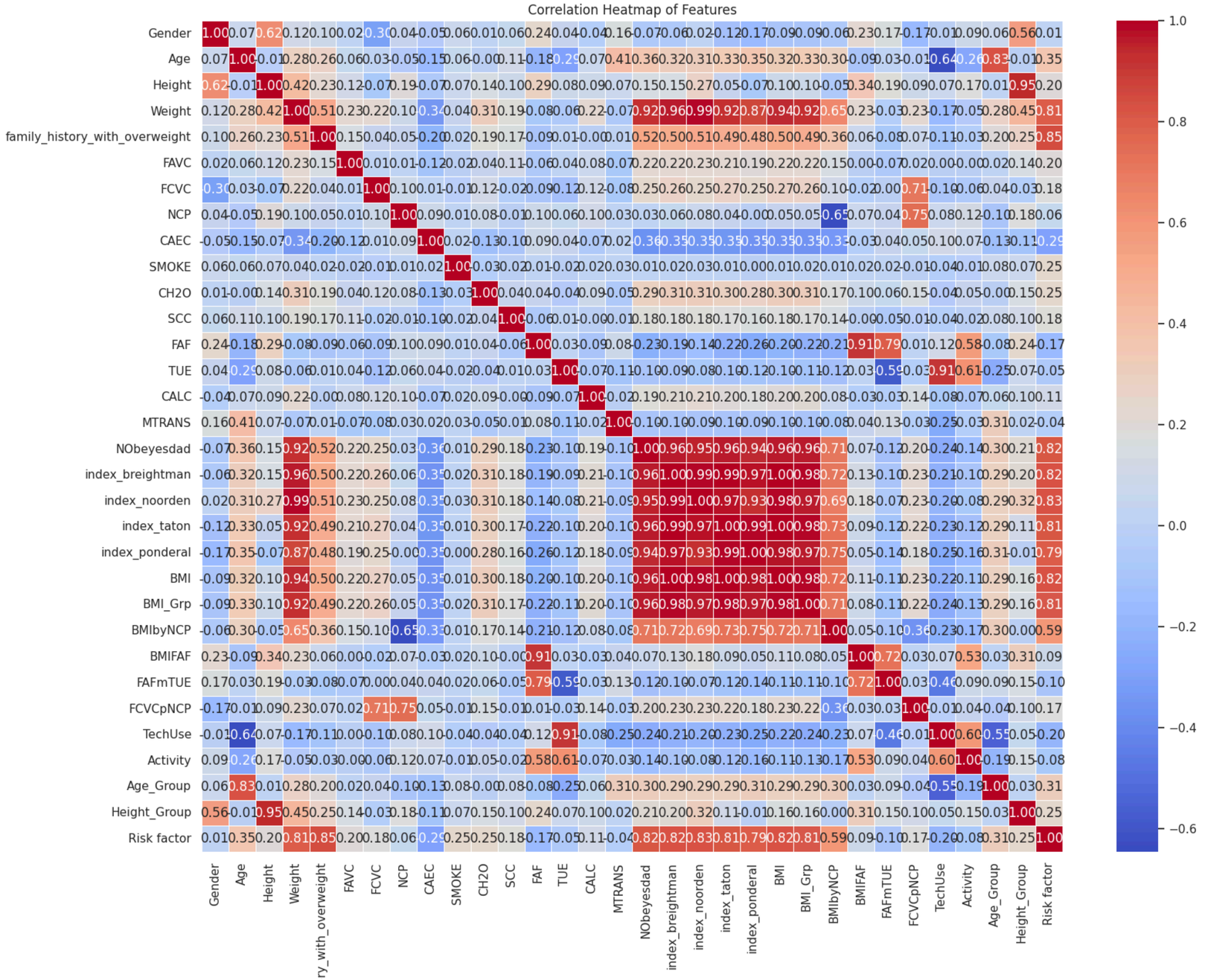
Category	BMI (kg/m²)[c]	BMI Prime[c]
Underweight (Severe thinness)	< 16.0	< 0.64
Underweight (Moderate thinness)	16.0 – 16.9	0.64 – 0.67
Underweight (Mild thinness)	17.0 – 18.4	0.68 – 0.73
Normal range	18.5 – 24.9	0.74 – 0.99
Overweight (Pre-obese)	25.0 – 29.9	1.00 – 1.19
Obese (Class I)	30.0 – 34.9	1.20 – 1.39
Obese (Class II)	35.0 – 39.9	1.40 – 1.59
Obese (Class III)	≥ 40.0	≥ 1.60

feature engineering (+~1% acc cv)

- +10 features
- BMI_Grp: BMI bins [18.5, 25, 30, 35, 40]
- Age_Group: Age bins [0, 18, 30, 45, ‘inf']
- Activity: FAF/TUE
- Risk factor: (BMI + Age_Group)*(family_history_with_overweight + SMOKE)

Correlation

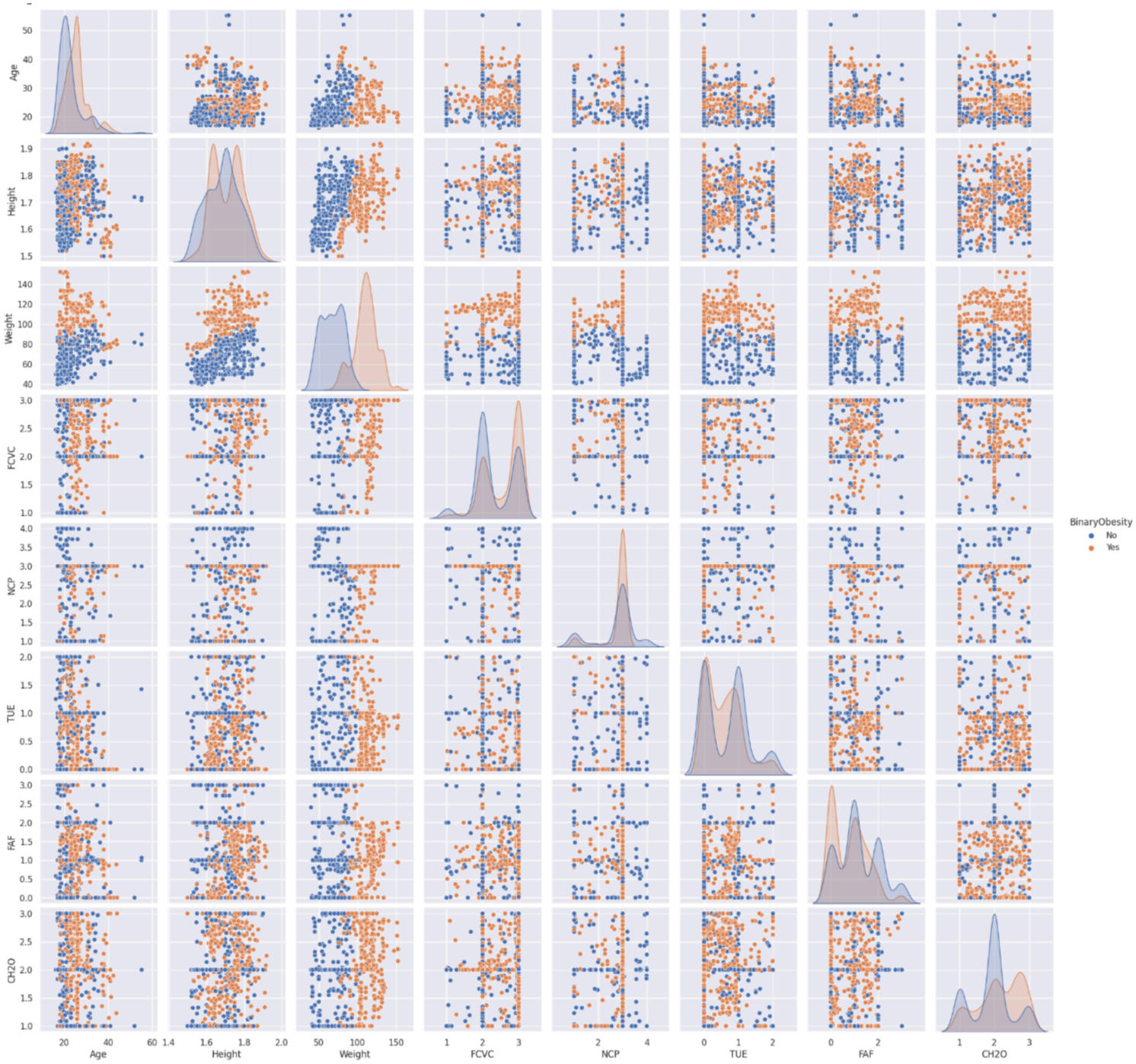
- Multicollinearity
- Index breightman (-0.5% acc)
- Boosted Trees are immune to multicollinearity



Distributions of num features according to BinaryObesity

- Outliers
- Age/Weight
- Height/Weight

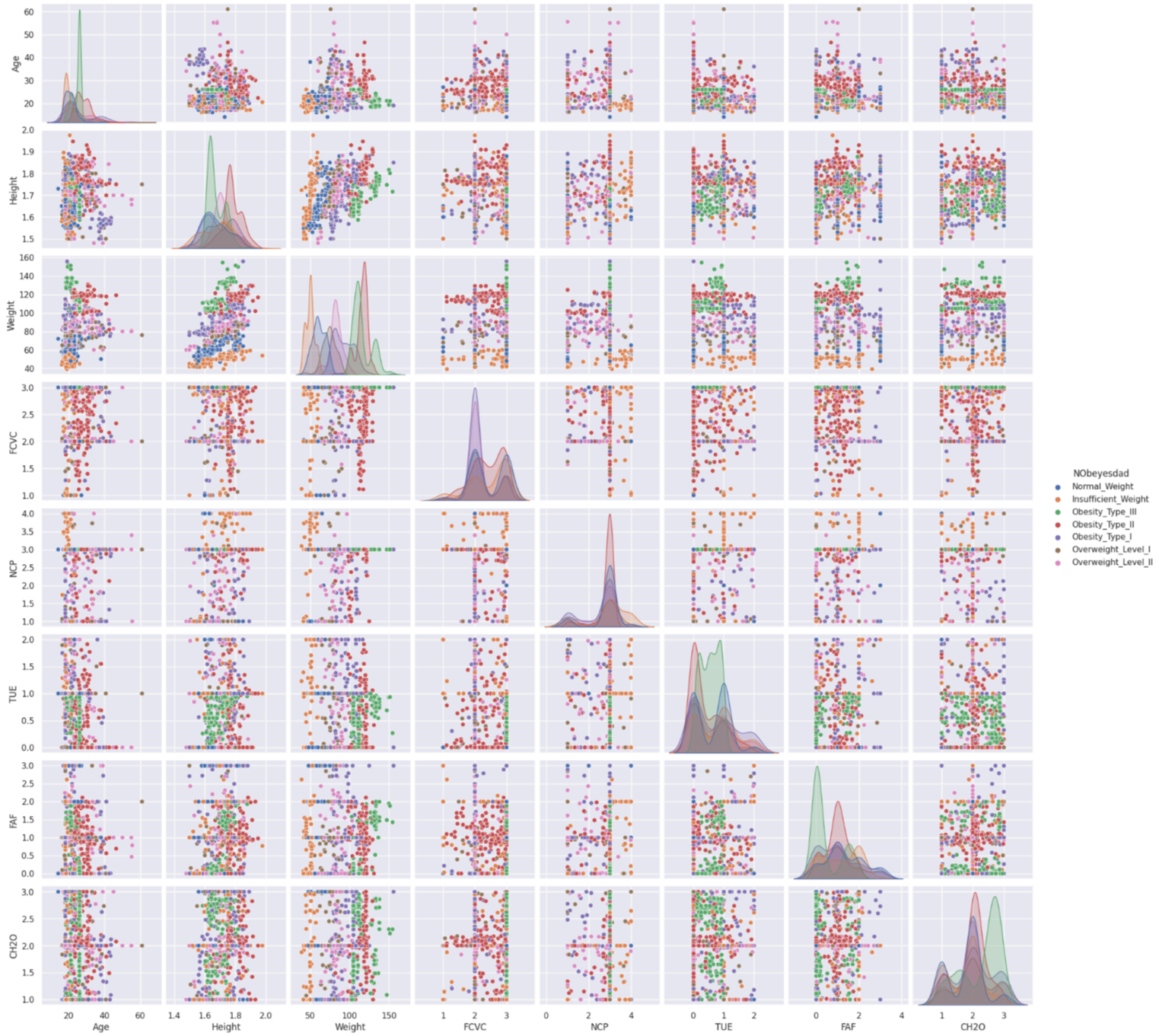
- Binary Classification
- Roc-Auc = 0.997
- Accuracy = 0.975



Distributions of num features according to NObeyesdad

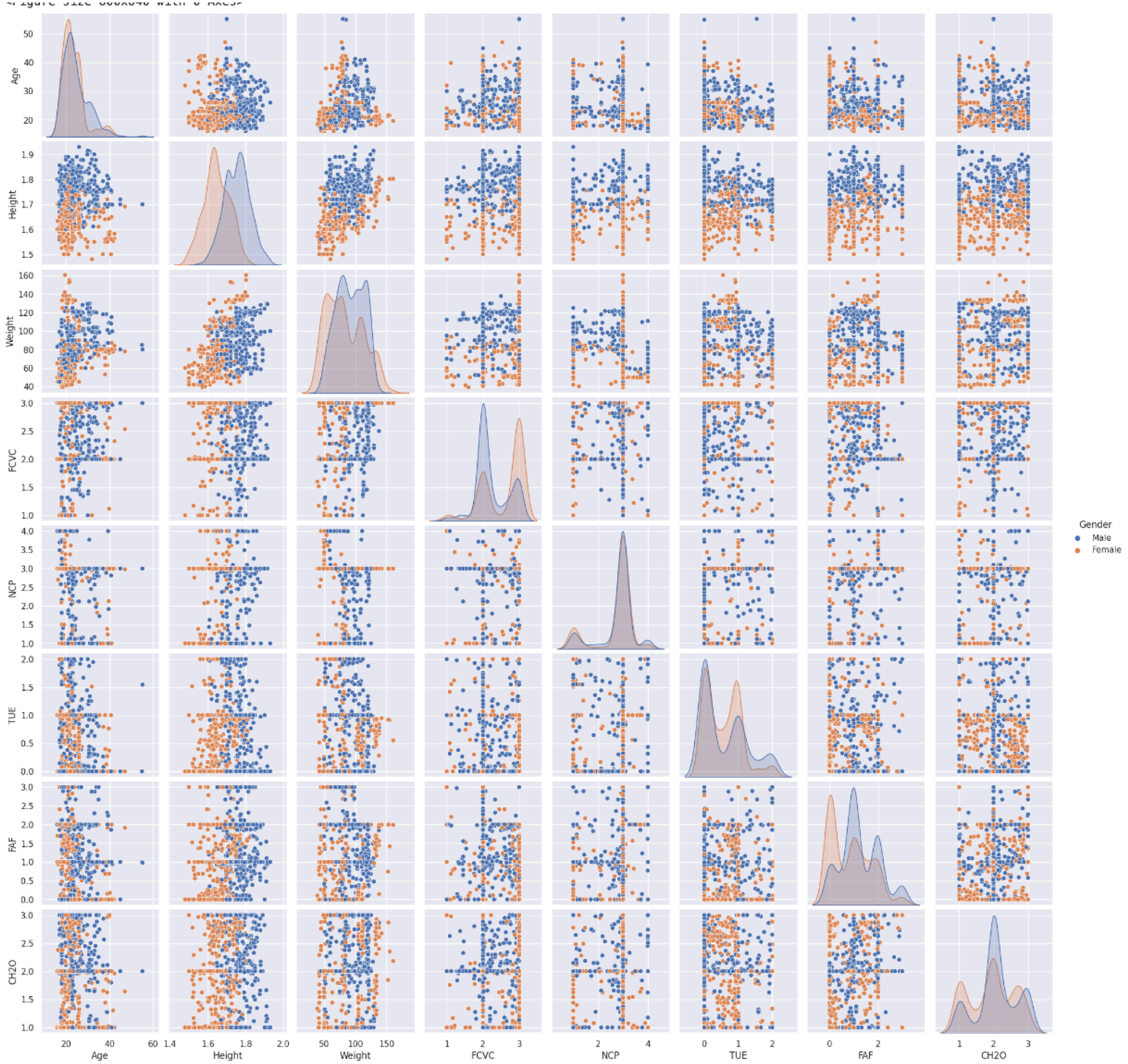
- Outliers
- Age/Weight
- Height/Weight

- 3 classes Obesity_Type_{i}: 0.975
- 2 classes Normal/ Insufficient_Weight: 0.94
- 2 classes Overweight_Level_{i}: 0.88



Distributions of num features according to gender

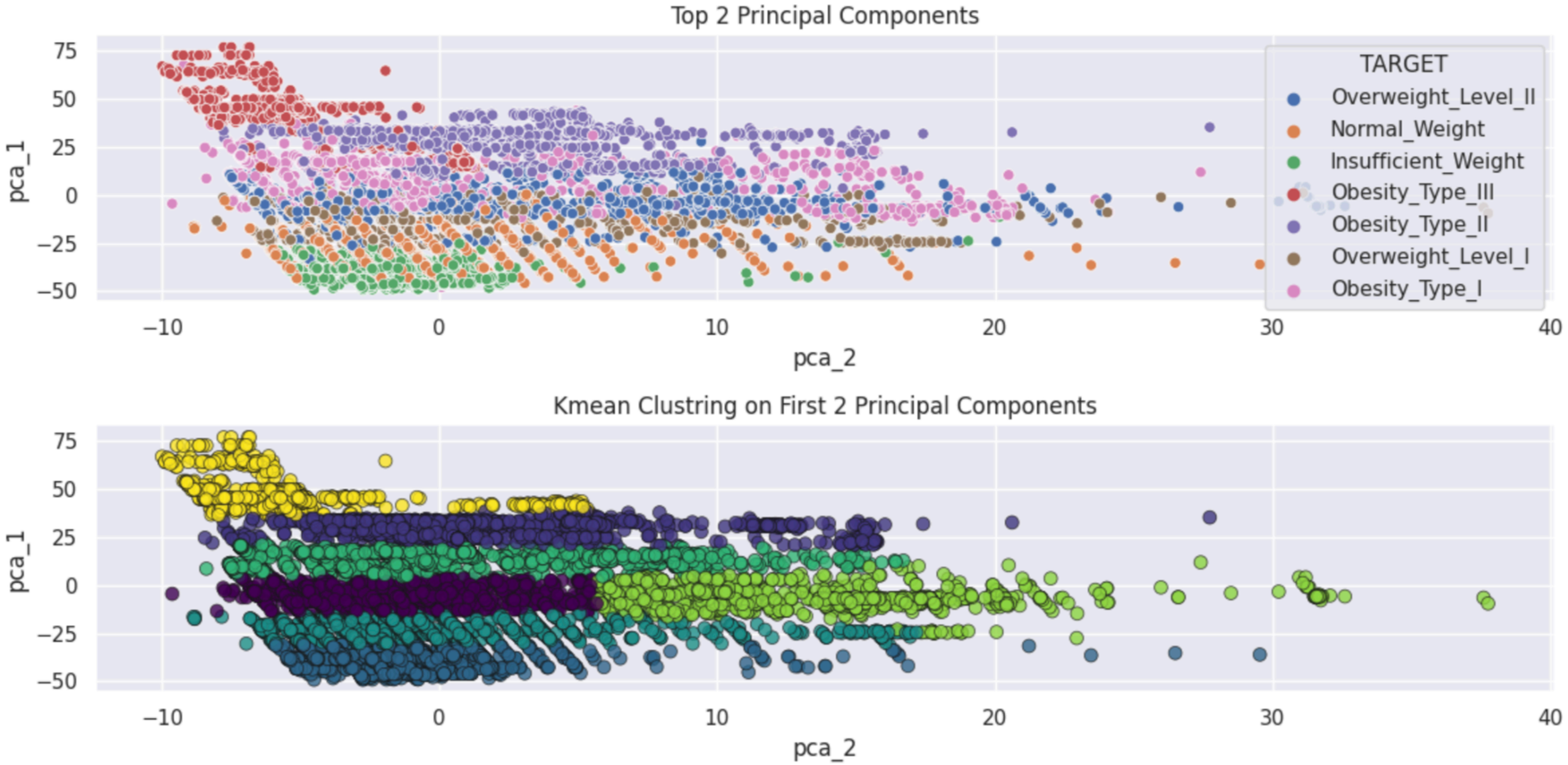
- 2 different models (cat boost)
- Female: 0.92 accuracy
- Male: 0.88 accuracy



Dimensionality reduction + clustering

- PCA
- TSNE
- UMAP

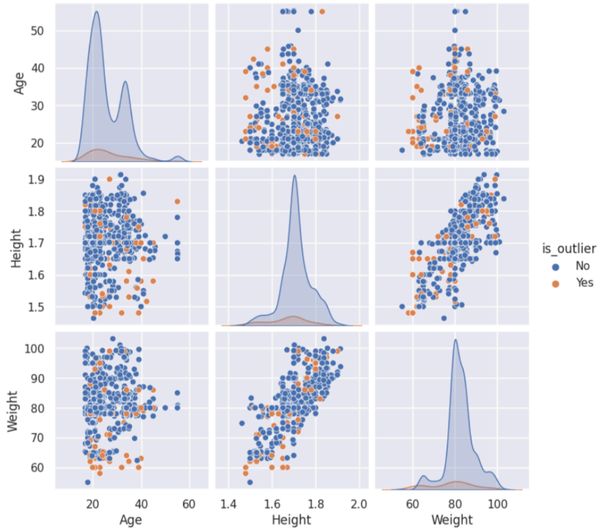
- K-means (n=7)

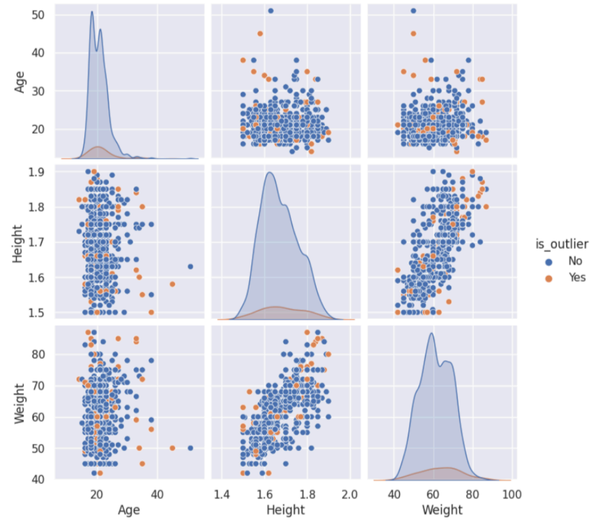


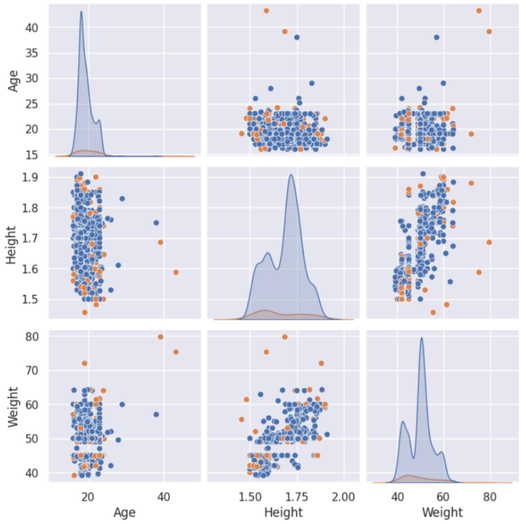

```
Variance Ratio for Each Principal Component:
Principal Component 1: 0.8891
Principal Component 2: 0.0637
```

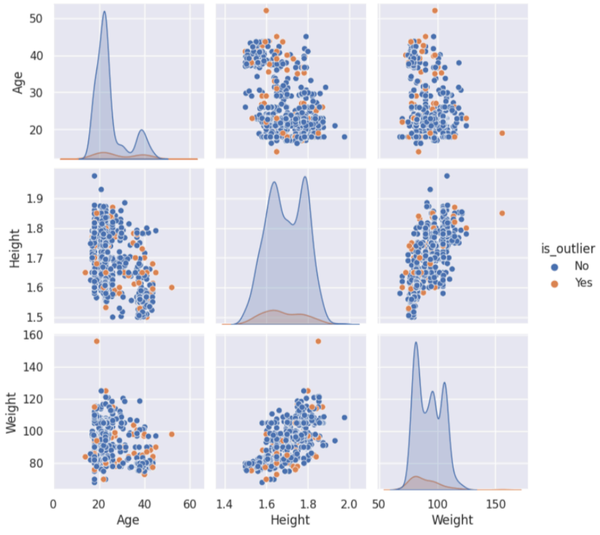
Outlier detection (+2% acc cv)

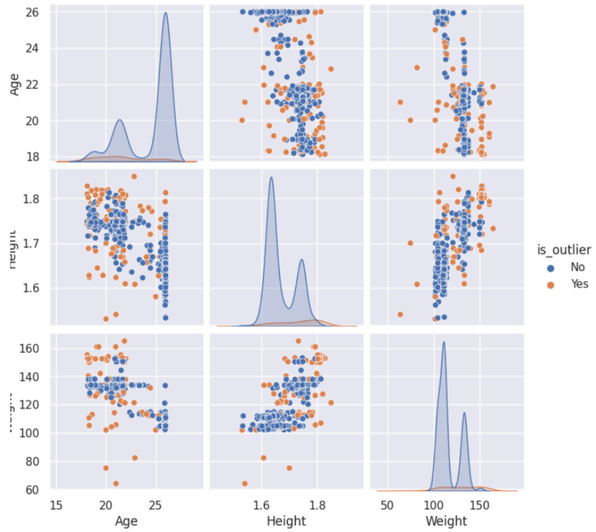
- IsolationForest
- 7 classes
- outlier_threshold = 0.1

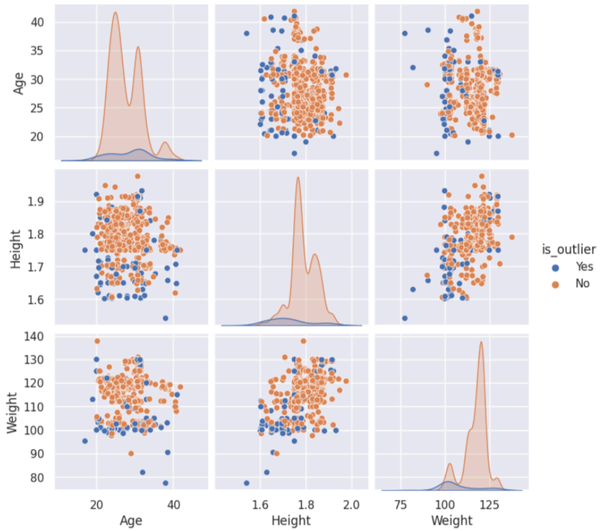


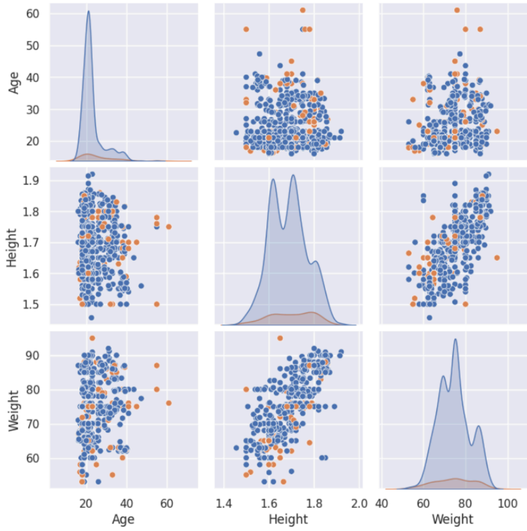






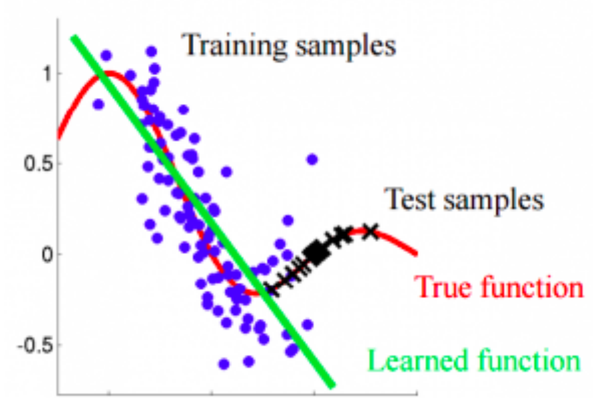


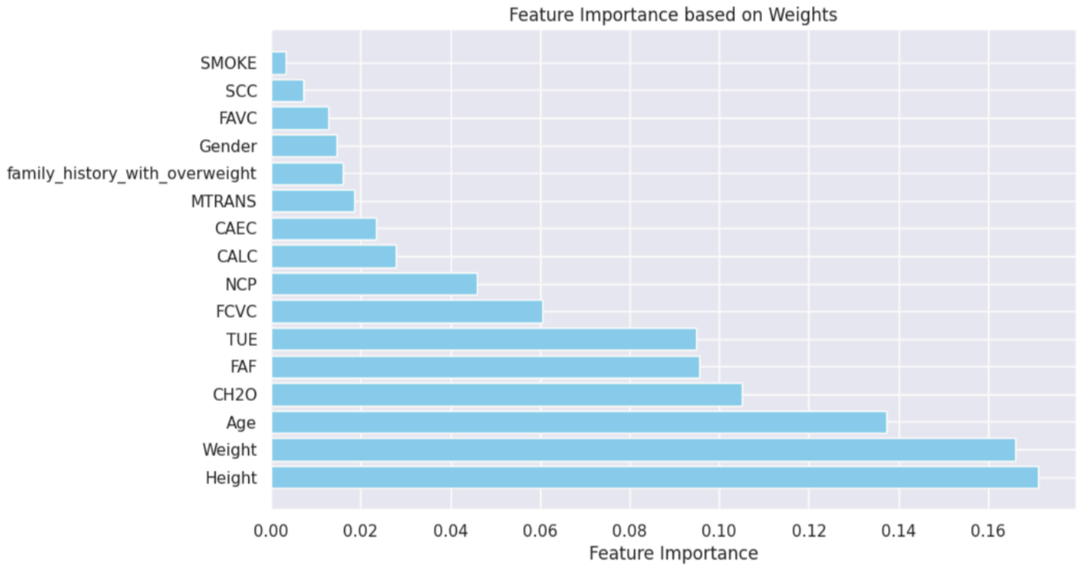




Covariate shift

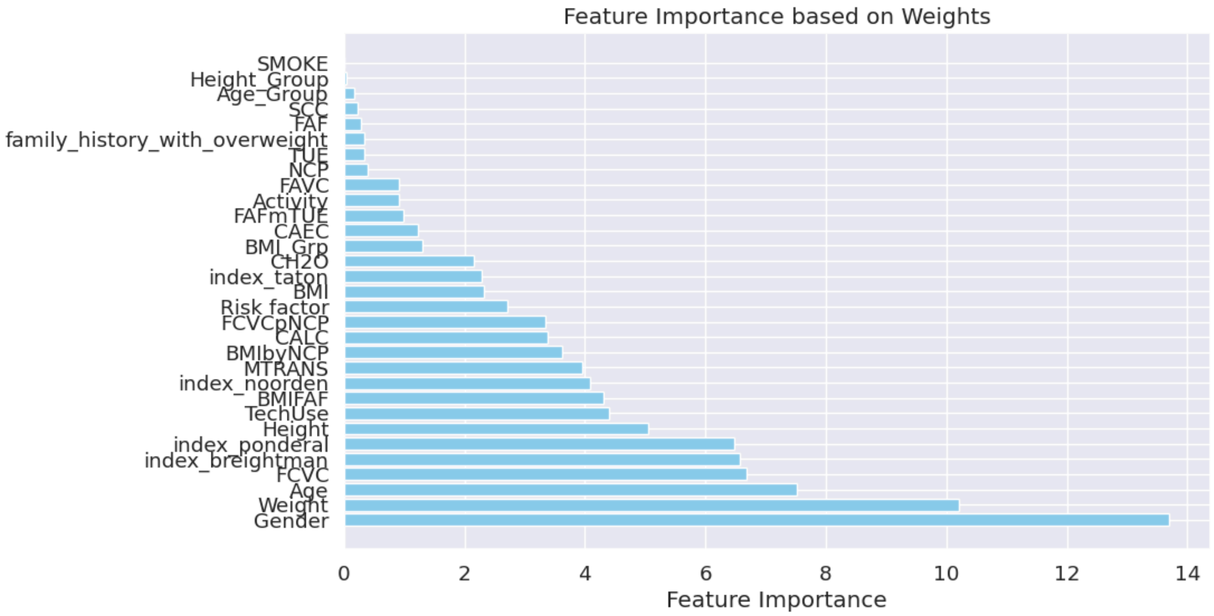
- Train_label = 0; Test_label = 1
- RandomForestClassifier
- Roc-Auc = 0.50
- Accuracy = 0.56





Baseline models

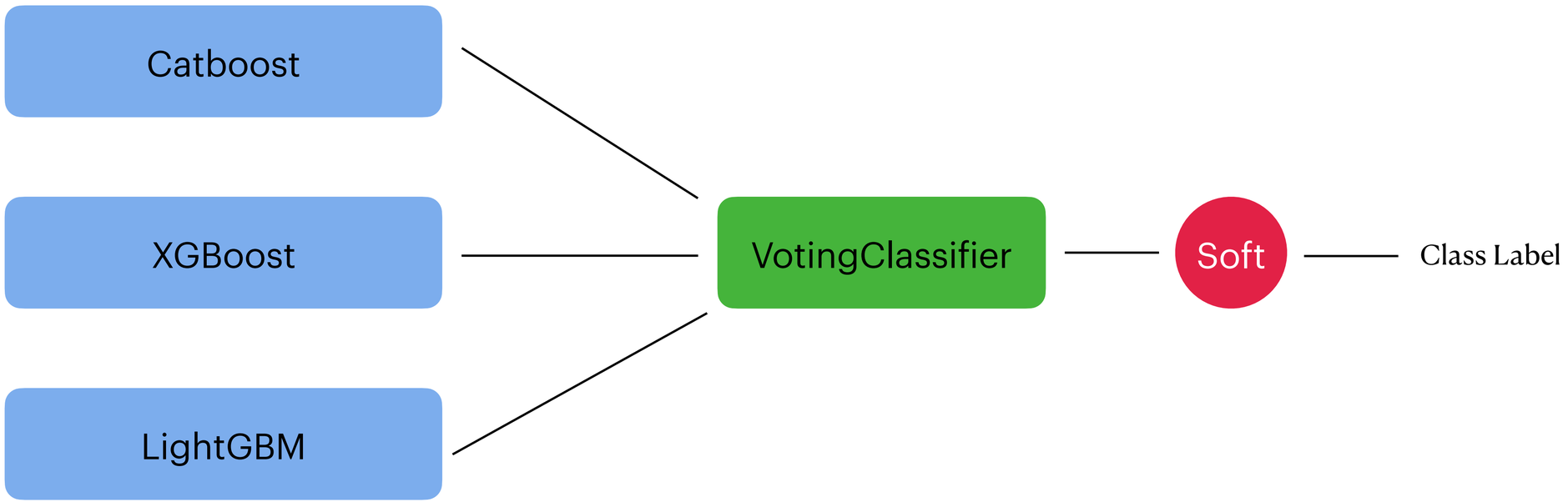
Catboost



XGBoost

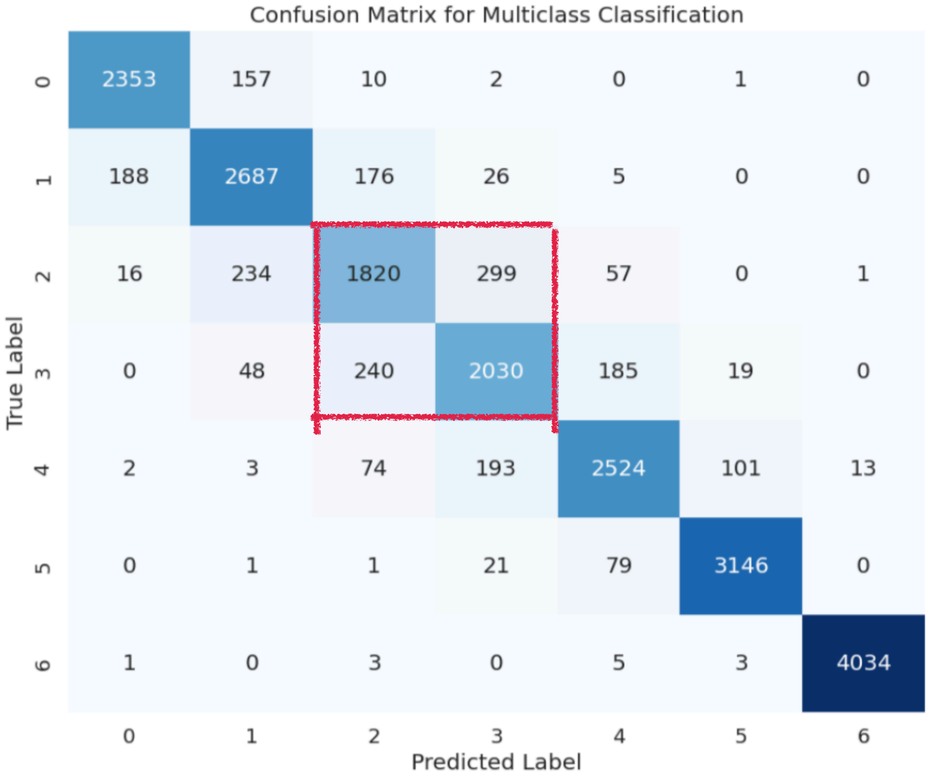
- Catboost accuracy ~ 0.903
- XGBoost accuracy ~ 0.899
- LightGBM ~ 0.902

Final model Voting (+1% acc cv)



Final results

- Cross validation accuracy = 0.917
- 5 folds



- most of the error is concentrated in 2 and 3 classes
- Overweight_Level_{I} classes

Unsuccessful Practices

- Pseudo Labeling
- Ordinal loss
- Stacking Ensemble
- Features Selection random Forest+RFE
- Random Search for baseline models

$$\mathcal{L}(\{y\},\{s\}) = -\sum_{i}\sum_{j=0}^{m-1} I_{y_i>j}\log(\text{sigmoid}(s_{i,j})) + I_{y_i\le j}\log(1-\text{sigmoid}(s_{i,j}))$$

Exploration Wishlist

- AutoML: autogluon framework
- Pyboost
- Optuna hyperparameter tuning
- Regression model
- Search for data leaks
- Binary classification (is_obesity) ->
  two separate classifiers

Thank you for your attention