Prediction of Obesity Levels using K Nearest Neighbor and Light Gradient Boosting Machine

Fathia Mohamed

Background

- Worldwide obesity tripled between 1975-2016 (less than 50 years)
- 1.9 billion (39%) adults overweight
- 650 million (13%) adults obese
- 340 million children and adolescents between ages 5-19 are overweight or obese

Obesity and overweight (2021, June 9). *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Obesity can lead to other health issues

- Cardiovascular diseases
- Diabetes
- Musculoskeletal disorders
- Cancer

Data

- From UC Irvine Machine Learning Repository

- 16 variables and the target variable (the obesity level category)

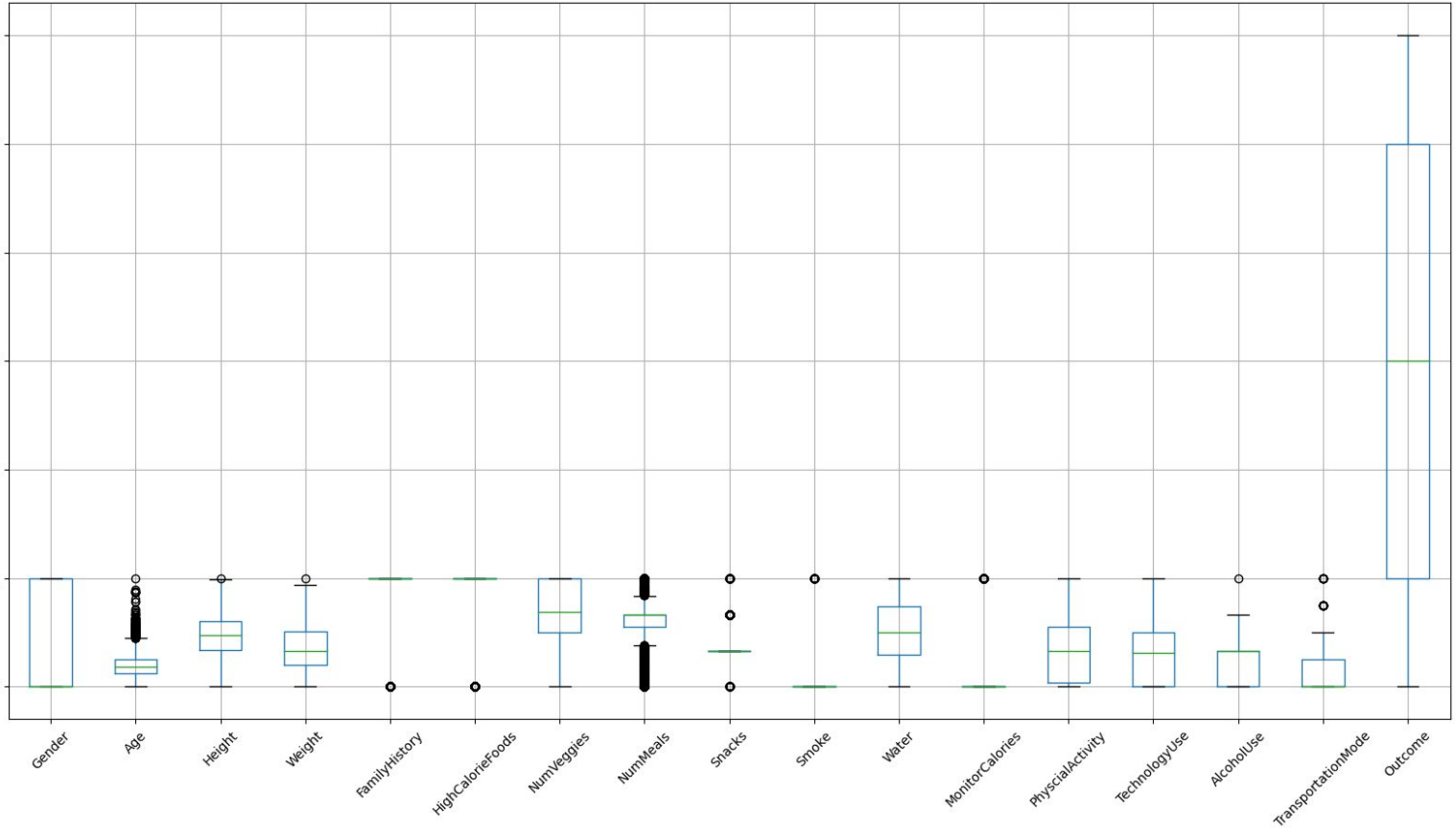
- Obesity level categories
  - Underweight
  - Normal
  - Overweight I
  - Overweight II
  - Obesity I
  - Obesity II
  - Obesity III

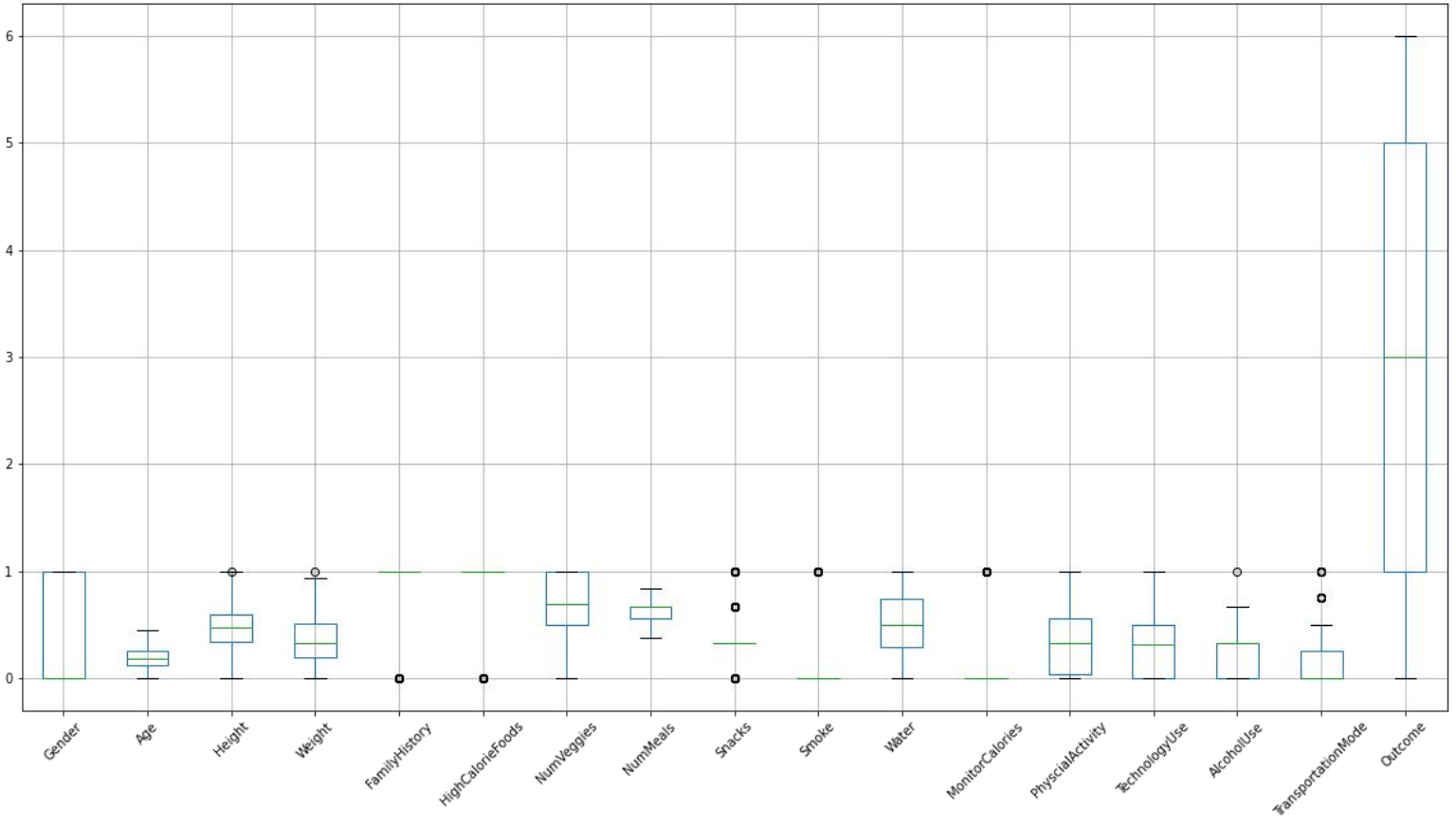
Exploration of Data

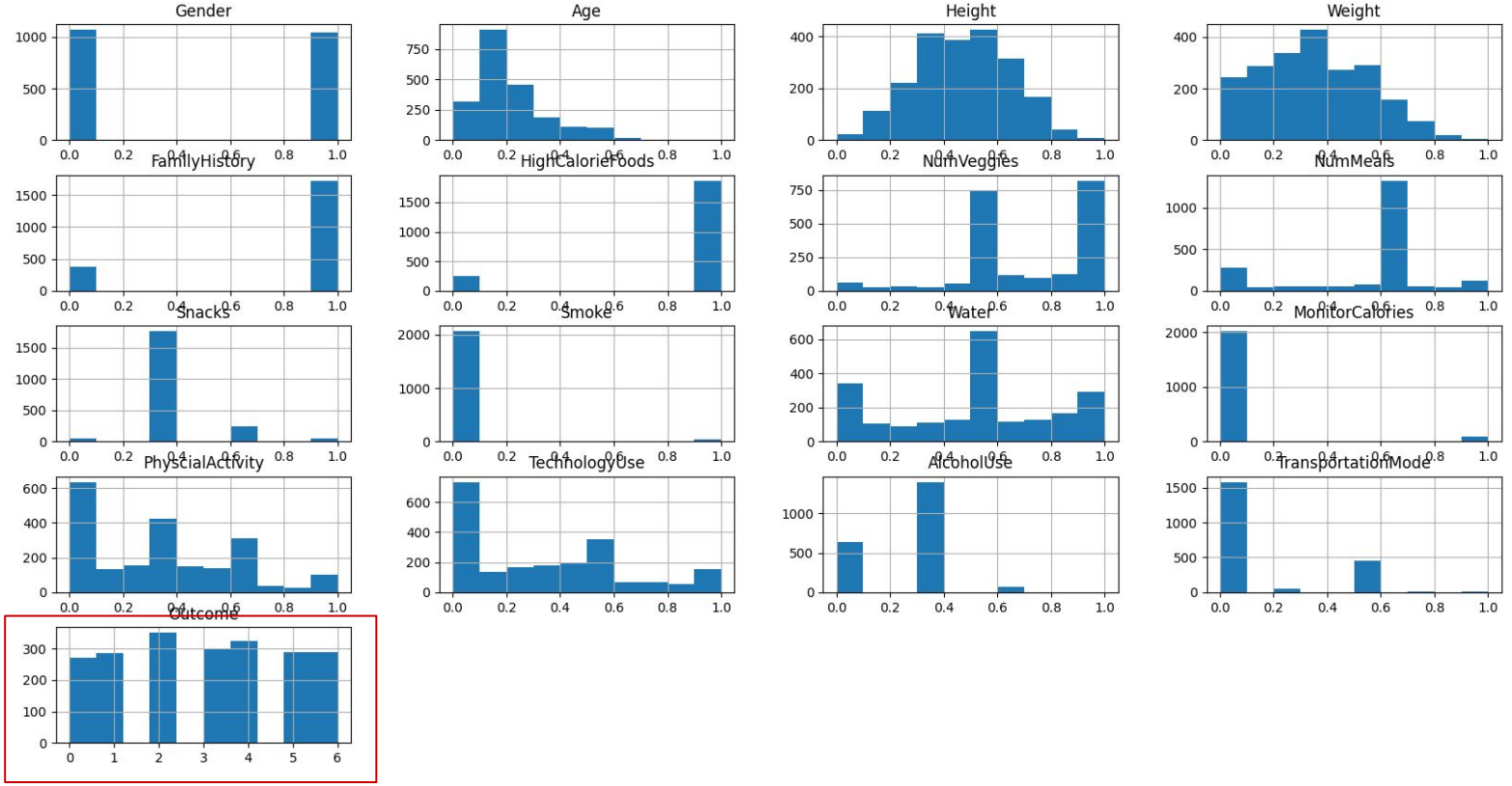
	Gender	Age	Height	Weight	FamilyHistory	HighCalorieFoods	NumVeggies	NumMeals	Snacks	Smoke	Water	MonitorCalories	PhyscialActivity	TechnologyUse	AlcoholUse	TransportationMode	Outcome
0	1	21.000000	1.620000	64.000000	1	0	2.0	3.0	1	0	2.000000	0	0.000000	1.000000	0	0	1
1	1	21.000000	1.520000	56.000000	1	0	3.0	3.0	1	1	3.000000	1	3.000000	0.000000	1	0	1
2	0	23.000000	1.800000	77.000000	1	0	2.0	3.0	1	0	2.000000	0	2.000000	1.000000	2	0	1
3	0	27.000000	1.800000	87.000000	0	0	3.0	3.0	1	0	2.000000	0	2.000000	0.000000	2	1	5
4	0	22.000000	1.780000	89.800000	0	0	2.0	1.0	1	0	2.000000	0	0.000000	0.000000	1	0	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	1	20.976842	1.710730	131.408528	1	1	3.0	3.0	1	0	1.728139	0	1.676269	0.906247	1	0	4
2107	1	21.982942	1.748584	133.742943	1	1	3.0	3.0	1	0	2.005130	0	1.341390	0.599270	1	0	4
2108	1	22.524036	1.752206	133.689352	1	1	3.0	3.0	1	0	2.054193	0	1.414209	0.646288	1	0	4
2109	1	24.361936	1.739450	133.346641	1	1	3.0	3.0	1	0	2.852339	0	1.139107	0.586035	1	0	4
2110	1	23.664709	1.738836	133.472641	1	1	3.0	3.0	1	0	2.863513	0	1.026452	0.714137	1	0	4

Normalized Data

	Gender	Age	Height	Weight	FamilyHistory	HighCalorieFoods	NumVeggies	NumMeals	Snacks	Smoke	Water	MonitorCalories	PhyscialActivity	TechnologyUse	AlcoholUse	TransportationMode	Outcome
0	1.0	0.148936	0.320755	0.186567	1.0	0.0	0.5	0.666667	0.333333	0.0	0.500000	0.0	0.000000	0.500000	0.000000	0.00	1
1	1.0	0.148936	0.132075	0.126866	1.0	0.0	1.0	0.666667	0.333333	1.0	1.000000	1.0	1.000000	0.000000	0.333333	0.00	1
2	0.0	0.191489	0.660377	0.283582	1.0	0.0	0.5	0.666667	0.333333	0.0	0.500000	0.0	0.666667	0.500000	0.666667	0.00	1
3	0.0	0.276596	0.660377	0.358209	0.0	0.0	1.0	0.666667	0.333333	0.0	0.500000	0.0	0.666667	0.000000	0.666667	0.25	5
4	0.0	0.170213	0.622642	0.379104	0.0	0.0	0.5	0.000000	0.333333	0.0	0.500000	0.0	0.000000	0.000000	0.333333	0.00	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2106	1.0	0.148443	0.491943	0.689616	1.0	1.0	1.0	0.666667	0.333333	0.0	0.364070	0.0	0.558756	0.453124	0.333333	0.00	4
2107	1.0	0.169850	0.563366	0.707037	1.0	1.0	1.0	0.666667	0.333333	0.0	0.502565	0.0	0.447130	0.299635	0.333333	0.00	4
2108	1.0	0.181362	0.570200	0.706637	1.0	1.0	1.0	0.666667	0.333333	0.0	0.527097	0.0	0.471403	0.323144	0.333333	0.00	4
2109	1.0	0.220467	0.546132	0.704079	1.0	1.0	1.0	0.666667	0.333333	0.0	0.926170	0.0	0.379702	0.293017	0.333333	0.00	4
2110	1.0	0.205632	0.544974	0.705020	1.0	1.0	1.0	0.666667	0.333333	0.0	0.931757	0.0	0.342151	0.357069	0.333333	0.00	4

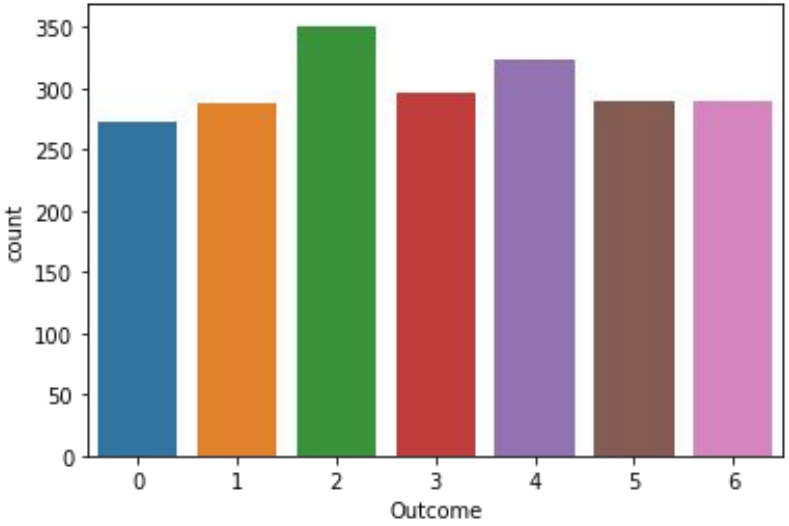

6
5
4
3
2
1
0
Gender
Age
Height
Weight
FamilyHistory
HighCalorieFoods
NumVeggies
NumMeals
Snacks
Smoke
Water
MonitorCalories
PhysicalActivity
TechnologyUse
AlcoholUse
TransportationMode
Outcome

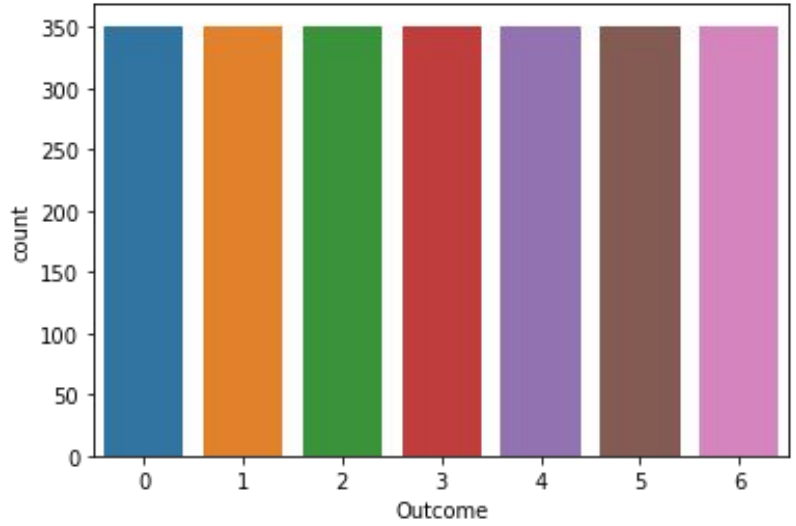

6
5
4
3
2
1
0
Gender
Age
Height
Weight
FamilyHistory
HighCalorieFoods
NumVeggies
NumMeals
Snacks
Smoke
Water
MonitorCalories
PhysicalActivity
TechnologyUse
AlcoholUse
TransportationMode
Outcome


Gender
Age
Height
Weight
FamilyHistory
HighCalorieFoods
NumVeggies
NumMeals
Snacks
Smoke
Water
MonitorCalories
PhysicalActivity
TechnologyUse
AlcoholUse
TransportationMode
Outcome

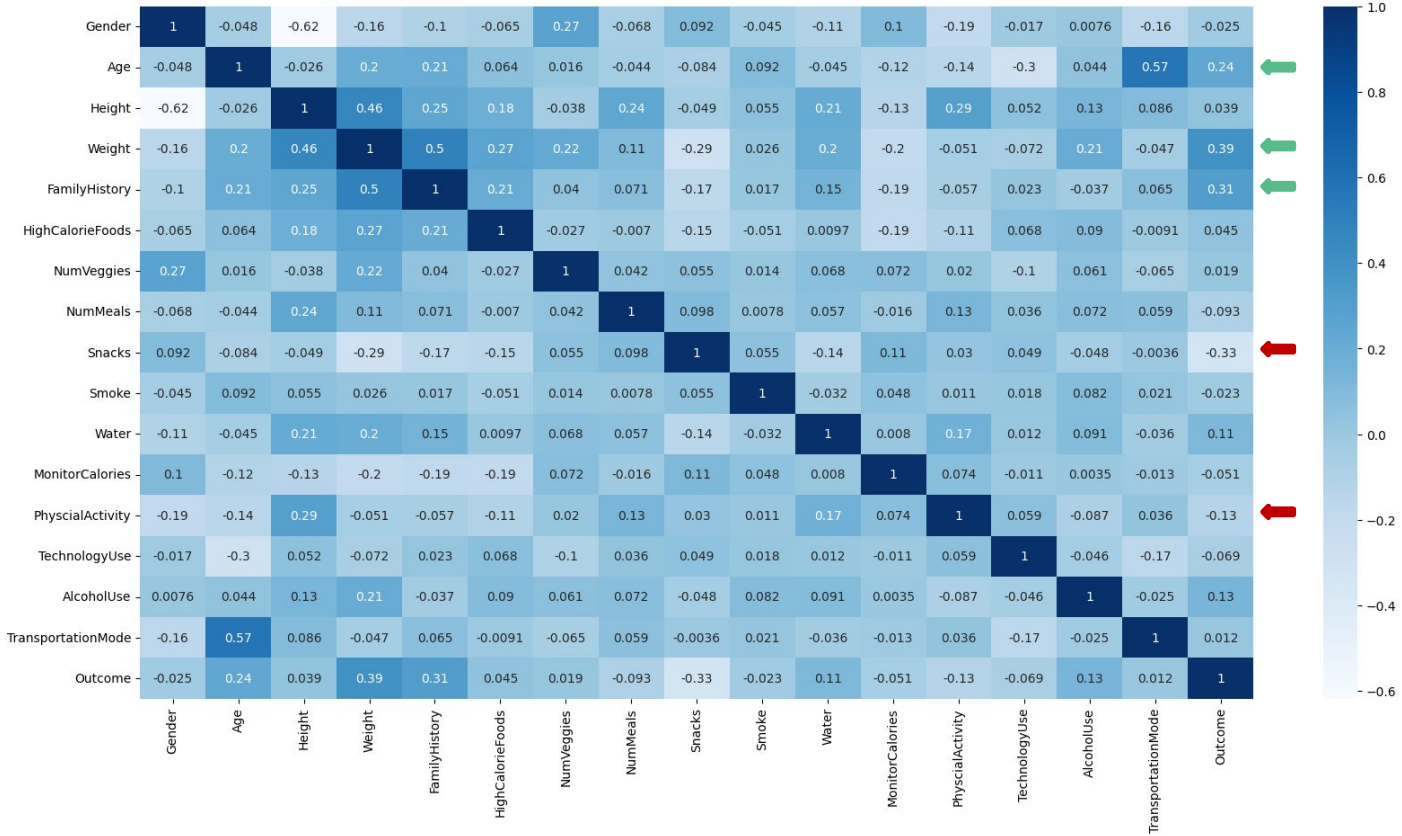
Synthetic Minority Oversampling Technique (SMOTE)

- Undersampling vs. Oversampling
- Generates new instances for the minority class


350
300
250
200
150
100
50
0
count
0
1
2
3
4
5
6
Outcome

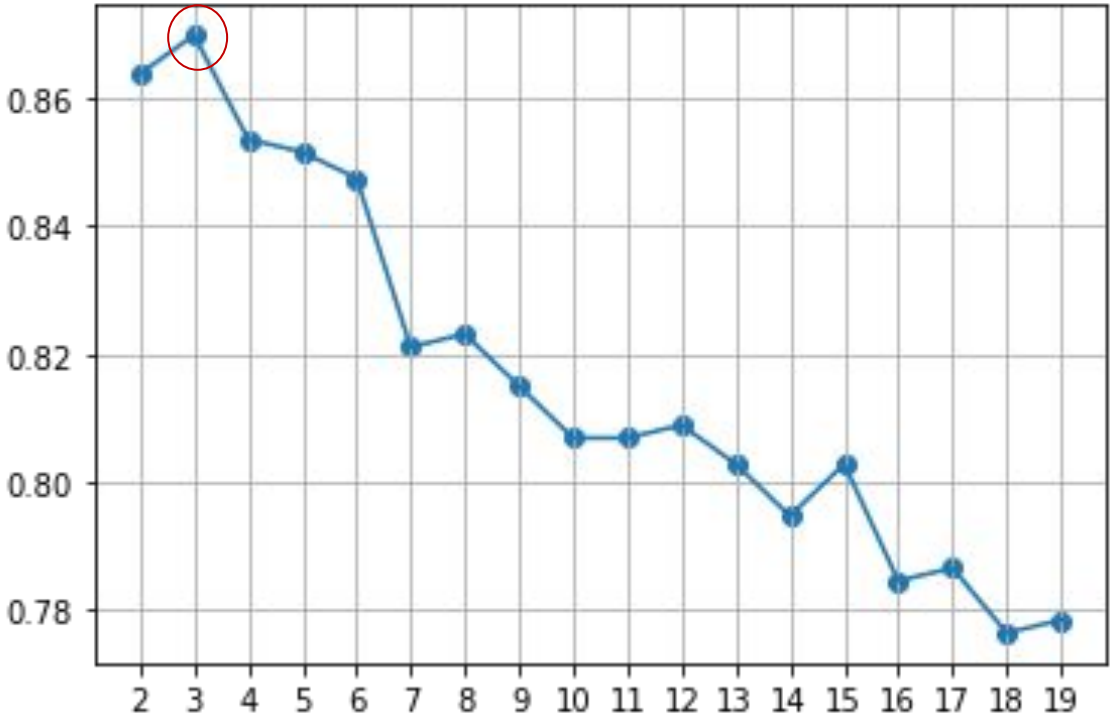

350
300
250
200
150
100
50
0
count
0
1
2
3
4
5
6
Outcome

	Gender	Age	Height	Weight	FamilyHistory	HighCalorieFoods	NumVeggies	NumMeals	Snacks	Smoke	Water	MonitorCalories	PhyscialActivity	TechnologyUse	AlcoholUse	TransportationMode	Outcome
Gender	1.000000	-0.048394	-0.618466	-0.161668	-0.102512	-0.064934	0.274505	-0.067600	0.091543	-0.044698	-0.107930	0.102633	-0.189607	-0.017269	0.007616	-0.164116	-0.024908
Age	-0.048394	1.000000	-0.025958	0.202560	0.205725	0.063902	0.016291	-0.043944	-0.083739	0.091987	-0.045304	-0.116283	-0.144938	-0.296931	0.044487	0.567983	0.236170
Height	-0.618466	-0.025958	1.000000	0.463136	0.247684	0.178364	-0.038121	0.243672	-0.048818	0.055499	0.213376	-0.133753	0.294709	0.051912	0.129732	0.085768	0.038986
Weight	-0.161668	0.202560	0.463136	1.000000	0.496820	0.272300	0.216125	0.107469	-0.287493	0.025746	0.200575	-0.201906	-0.051436	-0.071561	0.206677	-0.046615	0.387643
FamilyHistory	-0.102512	0.205725	0.247684	0.496820	1.000000	0.208036	0.040372	0.071370	-0.169787	0.017385	0.147437	-0.185422	-0.056673	0.022943	-0.036676	0.065036	0.313667
HighCalorieFoods	-0.064934	0.063902	0.178364	0.272300	0.208036	1.000000	-0.027283	-0.007000	-0.150068	-0.050660	0.009719	-0.190658	-0.107995	0.068417	0.089520	-0.009102	0.044582
NumVeggies	0.274505	0.016291	-0.038121	0.216125	0.040372	-0.027283	1.000000	0.042216	0.054670	0.014320	0.068461	0.071852	0.019939	-0.101135	0.060781	-0.065098	0.018522
NumMeals	-0.067600	-0.043944	0.243672	0.107469	0.071370	-0.007000	0.042216	1.000000	0.097801	0.007811	0.057088	-0.015624	0.129504	0.036326	0.071747	0.059022	-0.092616
Snacks	0.091543	-0.083739	-0.048818	-0.287493	-0.169787	-0.150068	0.054670	0.097801	1.000000	0.055282	-0.144995	0.109179	0.030110	0.048567	-0.047540	-0.003556	-0.327295
Smoke	-0.044698	0.091987	0.055499	0.025746	0.017385	-0.050660	0.014320	0.007811	0.055282	1.000000	-0.031995	0.047731	0.011216	0.017613	0.082471	0.021045	-0.023256
Water	-0.107930	-0.045304	0.213376	0.200575	0.147437	0.009719	0.068461	0.057088	-0.144995	-0.031995	1.000000	0.008036	0.167236	0.011965	0.091386	-0.035638	0.108868
MonitorCalories	0.102633	-0.116283	-0.133753	-0.201906	-0.185422	-0.190658	0.071852	-0.015624	0.109179	0.047731	0.008036	1.000000	0.074221	-0.010928	0.003463	-0.012794	-0.050679
PhyscialActivity	-0.189607	-0.144938	0.294709	-0.051436	-0.056673	-0.107995	0.019939	0.129504	0.030110	0.011216	0.167236	0.074221	1.000000	0.058562	-0.086799	0.036098	-0.129564
TechnologyUse	-0.017269	-0.296931	0.051912	-0.071561	0.022943	0.068417	-0.101135	0.036326	0.048567	0.017613	0.011965	-0.010928	0.058562	1.000000	-0.045864	-0.165571	-0.069448
AlcoholUse	0.007616	0.044487	0.129732	0.206677	-0.036676	0.089520	0.060781	0.071747	-0.047540	0.082471	0.091386	0.003463	-0.086799	-0.045864	1.000000	-0.025492	0.134632
TransportationMode	-0.164116	0.567983	0.085768	-0.046615	0.065036	-0.009102	-0.065098	0.059022	-0.003556	0.021045	-0.035638	-0.012794	0.036098	-0.165571	-0.025492	1.000000	0.012268
Outcome	-0.024908	0.236170	0.038986	0.387643	0.313667	0.044582	0.018522	-0.092616	-0.327295	-0.023256	0.108868	-0.050679	-0.129564	-0.069448	0.134632	0.012268	1.000000


Gender 1 -0.048 -0.62 -0.16 -0.1 -0.065 0.27 -0.068 0.092 -0.045 -0.11 0.1 -0.19 -0.017 0.0076 -0.16 -0.025
Age -0.048 1 -0.026 0.2 0.21 0.064 0.016 -0.044 -0.084 0.092 -0.045 -0.12 -0.14 -0.3 0.044 0.57 0.24
Height -0.62 -0.026 1 0.46 0.25 0.18 -0.038 0.24 -0.049 0.055 0.21 -0.13 0.29 0.052 0.13 0.086 0.039
Weight -0.16 0.2 0.46 1 0.5 0.27 0.22 0.11 -0.29 0.026 0.2 -0.2 -0.051 -0.072 0.21 -0.047 0.39
FamilyHistory -0.1 0.21 0.25 0.5 1 0.21 0.04 0.071 -0.17 0.017 0.15 -0.19 -0.057 0.023 -0.037 0.065 0.31
HighCalorieFoods -0.065 0.064 0.18 0.27 0.21 1 -0.027 -0.007 -0.15 -0.051 0.0097 -0.19 -0.11 0.068 0.09 -0.0091 0.045
NumVeggies 0.27 0.016 -0.038 0.22 0.04 -0.027 1 0.042 0.055 0.014 0.068 0.072 0.02 -0.1 0.061 -0.065 0.019
NumMeals -0.068 -0.044 0.24 0.11 0.071 -0.007 0.042 1 0.098 0.0078 0.057 -0.016 0.13 0.036 0.072 0.059 -0.093
Snacks 0.092 -0.084 -0.049 -0.29 -0.17 -0.15 0.055 0.098 1 0.055 -0.14 0.11 0.03 0.049 -0.048 -0.0036 -0.33
Smoke -0.045 0.092 0.055 0.026 0.017 -0.051 0.014 0.0078 0.055 1 -0.032 0.048 0.011 0.018 0.082 0.021 -0.023
Water -0.11 -0.045 0.21 0.2 0.15 0.0097 0.068 0.057 -0.14 -0.032 1 0.008 0.17 0.012 0.091 -0.036 0.11
MonitorCalories 0.1 -0.12 -0.13 -0.2 -0.19 -0.19 0.072 -0.016 0.11 0.048 0.008 1 0.074 -0.011 0.0035 -0.013 -0.051
PhyscialActivity -0.19 -0.14 0.29 -0.051 -0.057 -0.11 0.02 0.13 0.03 0.011 0.17 0.074 1 0.059 -0.087 0.036 -0.13
TechnologyUse -0.017 -0.3 0.052 -0.072 0.023 0.068 -0.1 0.036 0.049 0.018 0.012 -0.011 0.059 1 -0.046 -0.17 -0.069
AlcoholUse 0.0076 0.044 0.13 0.21 -0.037 0.09 0.061 0.072 -0.048 0.082 0.091 0.0035 -0.087 -0.046 1 -0.025 0.13
TransportationMode -0.16 0.57 0.086 -0.047 0.065 -0.0091 -0.065 0.059 -0.0036 0.021 -0.036 -0.013 0.036 -0.17 -0.025 1 0.012
Outcome -0.025 0.24 0.039 0.39 0.31 0.045 0.019 -0.093 -0.33 -0.023 0.11 -0.051 -0.13 -0.069 0.13 0.012 1
Gender Age Height Weight FamilyHistory HighCalorieFoods NumVeggies NumMeals Snacks Smoke Water MonitorCalories PhyscialActivity TechnologyUse AlcoholUse TransportationMode Outcome
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

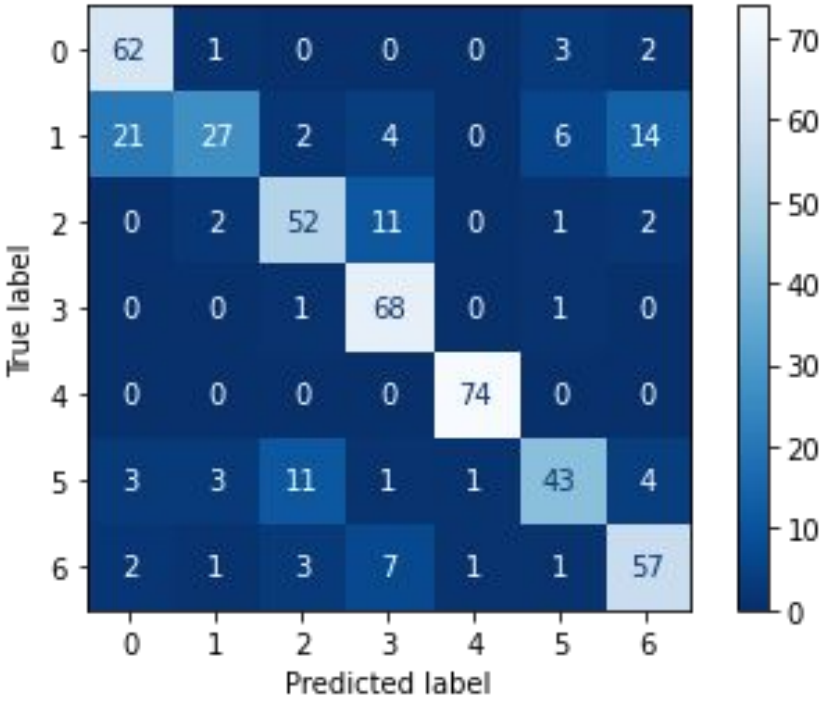
K Nearest Neighbor (KNN)

- Supervised machine learning
- Used for classification and regression
- Similar data points have similar labels/values
- Splits data into training and test sets

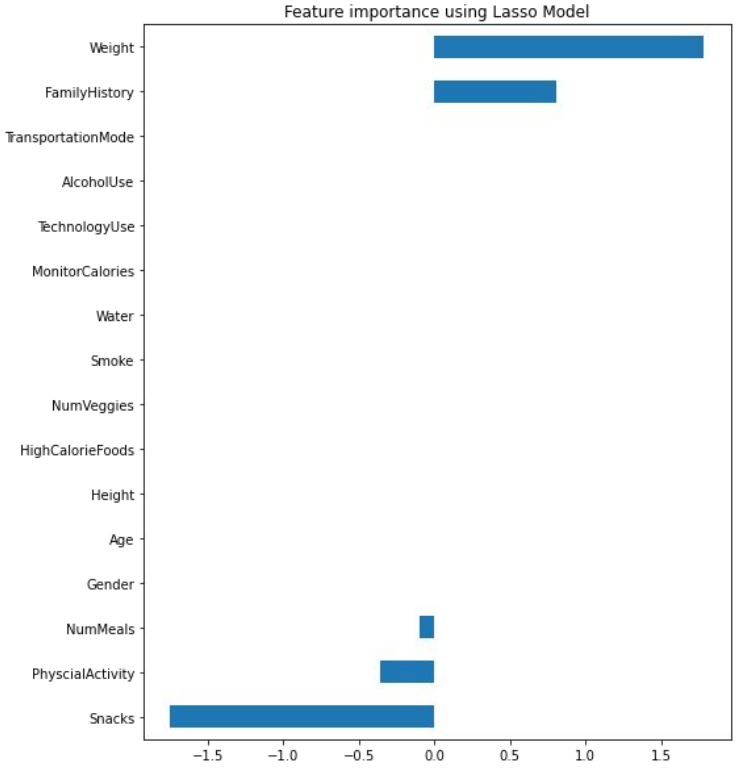

0.86
0.84
0.82
0.80
0.78
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

```
Accuracy: 0.8699186991869918
```

Evaluation of Clustering Model



Feature Importance - KNN



Accuracy: 0.7658536585365854

Evaluation of the most optimal model using Pycaret

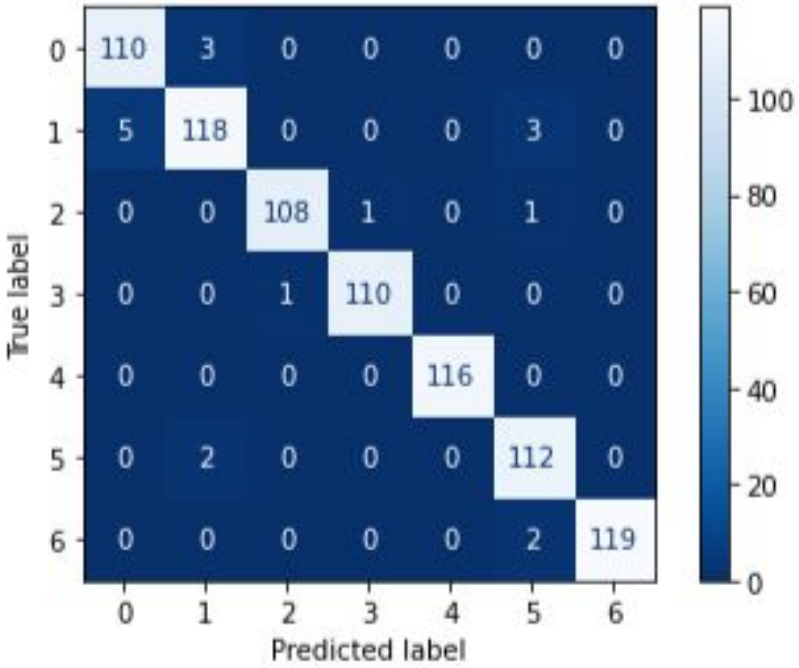
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9756	0.9988	0.9756	0.9766	0.9756	0.9715	0.9717	9.9220
gbc	Gradient Boosting Classifier	0.9651	0.9979	0.9651	0.9664	0.9651	0.9593	0.9595	0.8470
rf	Random Forest Classifier	0.9581	0.9977	0.9581	0.9620	0.9584	0.9511	0.9517	0.0940
et	Extra Trees Classifier	0.9552	0.9966	0.9552	0.9589	0.9554	0.9477	0.9483	0.0740
dt	Decision Tree Classifier	0.9360	0.9627	0.9360	0.9391	0.9361	0.9253	0.9258	0.0100
lda	Linear Discriminant Analysis	0.8964	0.9896	0.8964	0.9016	0.8965	0.8792	0.8800	0.0100
knn	K Neighbors Classifier	0.7836	0.9438	0.7836	0.7845	0.7764	0.7475	0.7498	0.0650
lr	Logistic Regression	0.7406	0.9422	0.7406	0.7376	0.7308	0.6973	0.6998	0.8920
svm	SVM - Linear Kernel	0.7120	0.0000	0.7120	0.7510	0.6987	0.6640	0.6744	0.0400
ridge	Ridge Classifier	0.6335	0.0000	0.6335	0.6321	0.6115	0.5724	0.5796	0.0070
nb	Naive Bayes	0.5457	0.9123	0.5457	0.5754	0.4777	0.4699	0.5023	0.0100
ada	Ada Boost Classifier	0.2943	0.7778	0.2943	0.2684	0.1966	0.1767	0.2004	0.0520
qda	Quadratic Discriminant Analysis	0.1431	0.0000	0.1431	0.0205	0.0358	0.0000	0.0000	0.0220
dummy	Dummy Classifier	0.1396	0.5000	0.1396	0.0195	0.0342	0.0000	0.0000	0.0080

Light Gradient Boosting Machine (LGBM)

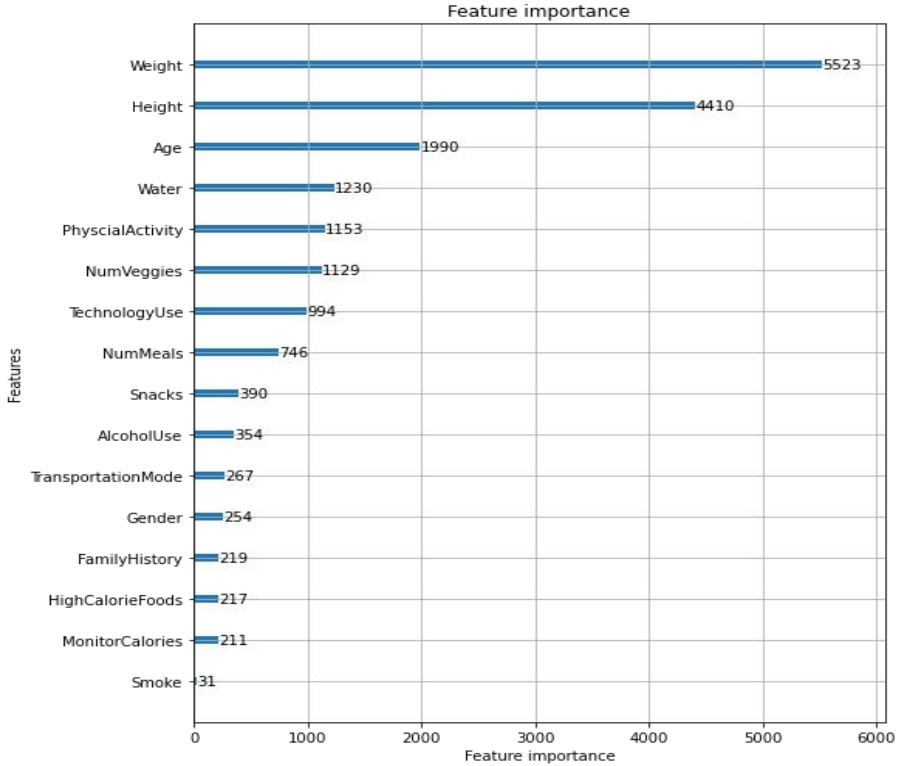
- Combines weak learners to create a strong predictive model
- Increases efficiency of the model

```
Training accuracy 1.0000
Testing accuracy 0.9778
```

Confusion Matrix - LGBM



Feature Importance - LGBM



```
Training accuracy 1.0000
Testing accuracy 0.9679
```

Conclusions

- KNN model - 87% accuracy
- LGBM model - 98% accuracy
- Great potential use in healthcare
- Future considerations

References

- Ali, M. (2021, November). PyCaret Tutorial: A beginner's guide for automating ML workflows using PyCaret. *Datacamp*.
  https://www.datacamp.com/tutorial/guide-for-automating-ml-workflows-using-pycaret?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720830&utm_adgroupid=157156377071&utm_device=c&utm_keywor
- Brownlee, J. (2021, January 5). Multi-Class Imbalanced Classification. *Machine Learning Mastery*. https://machinelearningmastery.com/multi-class-imbalanced-classification/
- Data Normalization with Pandas. (n.d.). *GeeksforGeeks*. https://www.geeksforgeeks.org/data-normalization-with-pandas/
- Mondal, A. (2023, August 17). Complete guide on how to Use LightGBM in Python. https://www.analyticsvidhya.com/blog/2021/08/complete-guide-on-how-to-use-lightgbm-in-python/
- Obesity and overweight (2021, June 9). *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
- Sathpathy, S. (2023, November 17). SMOTE for Imbalanced Classification with Python. *Analytics Vidhya*.
  https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/#h-smote-synthetic-minority-oversampling-technique
- Shafi, A. (2023, February). K-Nearest Neighbors (KNN) Classification with scikit-learn. *Datacamp*.
  https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720830&utm_adgroupid=157156377311&utm_device=c&utm_keyword
- Shetye, A. (2019, February 10). Feature Selection with sklearn and Pandas. *Medium*. https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b
- Zafar, A. (2022, February 15). Handling Outliers in
  Pandas. *Medium*. https://medium.com/@arsalan_zafar/handling-outliers-in-pandas-5cd872eef508#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6ImY4MzNlOGE3ZmUzZmU0Yjg3ODk0ODIxOWExNjg0YWZhMzczY2E4NmYiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJod