Obesity Prediction Using Machine Learning Methods

Nidhi Mankala|Tris Marie Joe|Naveen Kumar N| SCOPE

Introduction

In the current times, obesity is usually predicted using body mass index, but this diagnosis might not be always accurate. There are various other factors that can cause obesity. In our project we take into consideration all these external correlations to make predictions.

Motivation

It is important to try various methods in public health sector and so we believe that incorporating Machine Learning will help to improve predictions.

SCOPE of the Project

Analysis of different machine learning methods: Support Vector Machine, XGBoost, Adaboost, Gradient Boosting, Random forest classifier, category boosting and light gradient boosting. The models are trained on the dataset and is then evaluated using performance metrics such as accuracy, recall, precision score, ROC curve.

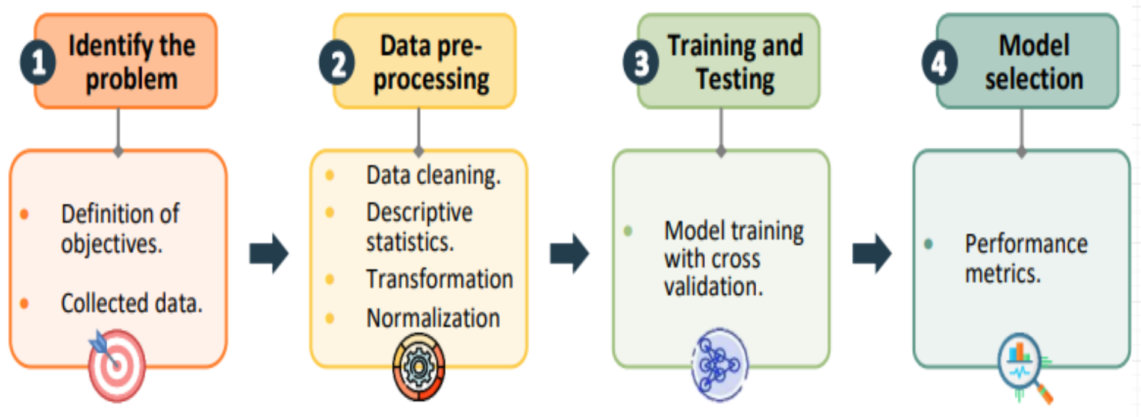
Methodology

This work seeks to predict obesity using data from the 'Estimation of Obesity Levels based on Eating Habits and Physical Condition dataset' from the UCI Machine Learning Repository. There are a total of 2111 instances of data collected, with 17 properties.

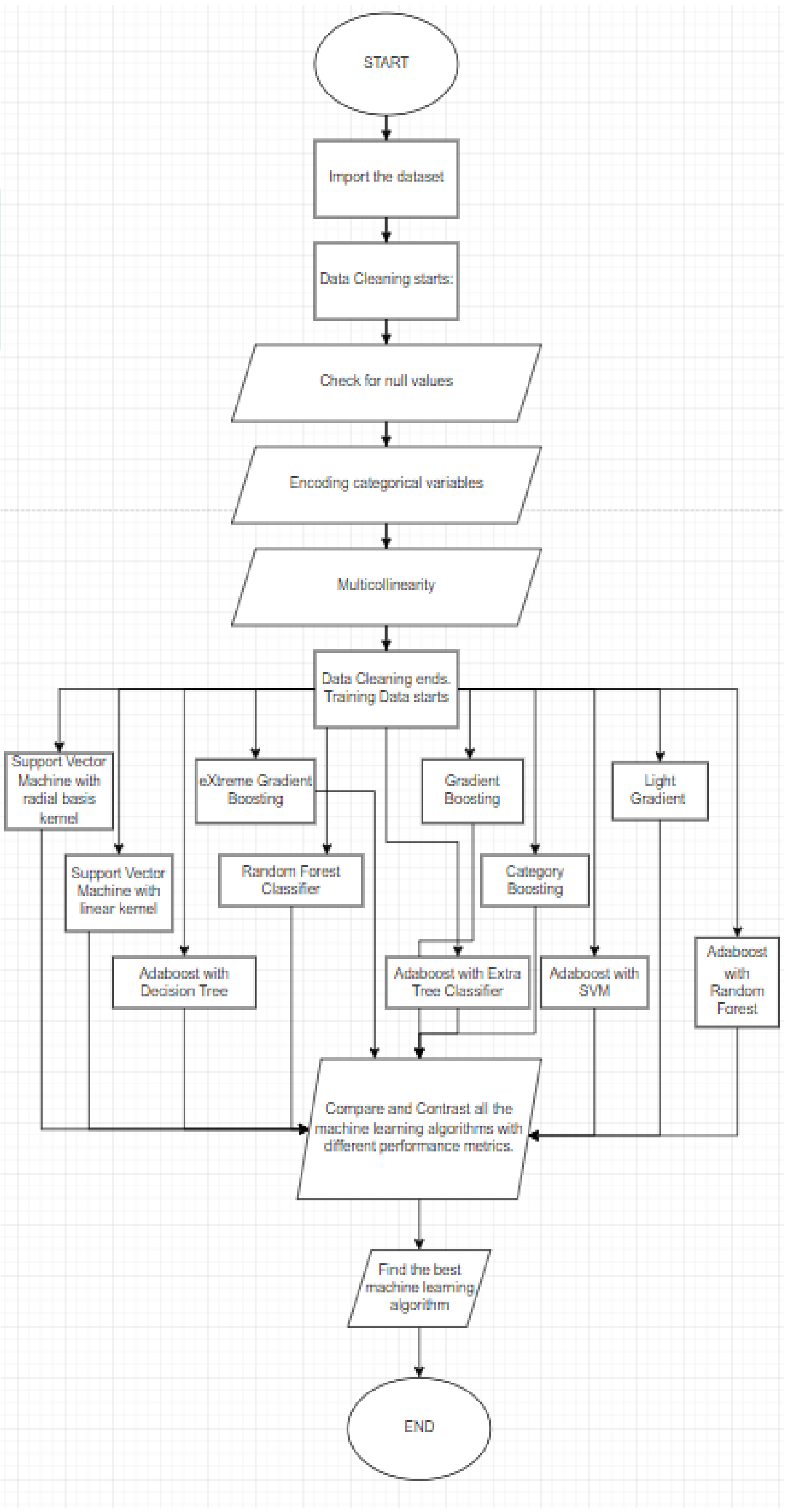
Features of dietary habits		
6	c_ECFF	Do you eat high caloric food frequently?
7	c_EVM	Do you usually eat vegetables in your meals?
8	c_MMHD	How many main meals do you have daily?
9	c_EFBM	Do you eat any food between meals?
10	c_SMOKE	Do you smoke?
11	c_WDRD	How much water do you drink daily?
12	c_DRAL	How often do you drink alcohol?
Physical condition features		
13	c_MCED	Do you monitor the calories you eat daily?
14	c_HPHA	How often do you have physical activity?
15	c_TTEC	How long do you use technological devices?
16	c_TRANSP	What type of transportation do you usually use?

List of esoteric features of the dataset that give a more accurate prediction

We start the project by importing our dataset, performing some methods for data cleaning to make our dataset ready. We then tried various different machine learning algorithms on this dataset and then compare the results of these algorithms with different performance metrics.

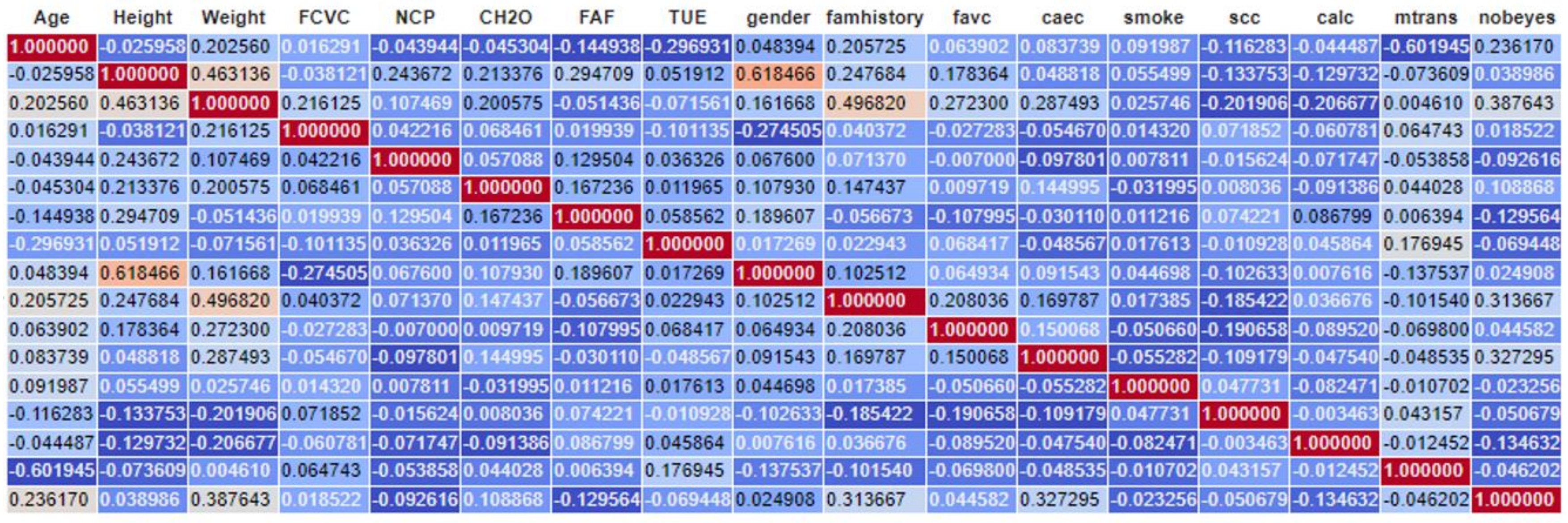


Pipeline diagram of the prediction model for the identification of obese or overweight people


START
Import the dataset
Data Cleaning starts:
Check for null values
Encoding categorical variables
Multicollinearity
Data Cleaning ends. Training Data starts
Support Vector Machine with radial basis kernel
eXtreme Gradient Boosting
Gradient Boosting
Light Gradient
Support Vector Machine with linear kernel
Random Forest Classifier
Category Boosting
Adaboost with Decision Tree
Adaboost with Extra Tree Classifier
Adaboost with SVM
Adaboost with Random Forest
Compare and Contrast all the machine learning algorithms with different performance metrics.
Find the best machine learning algorithm
END

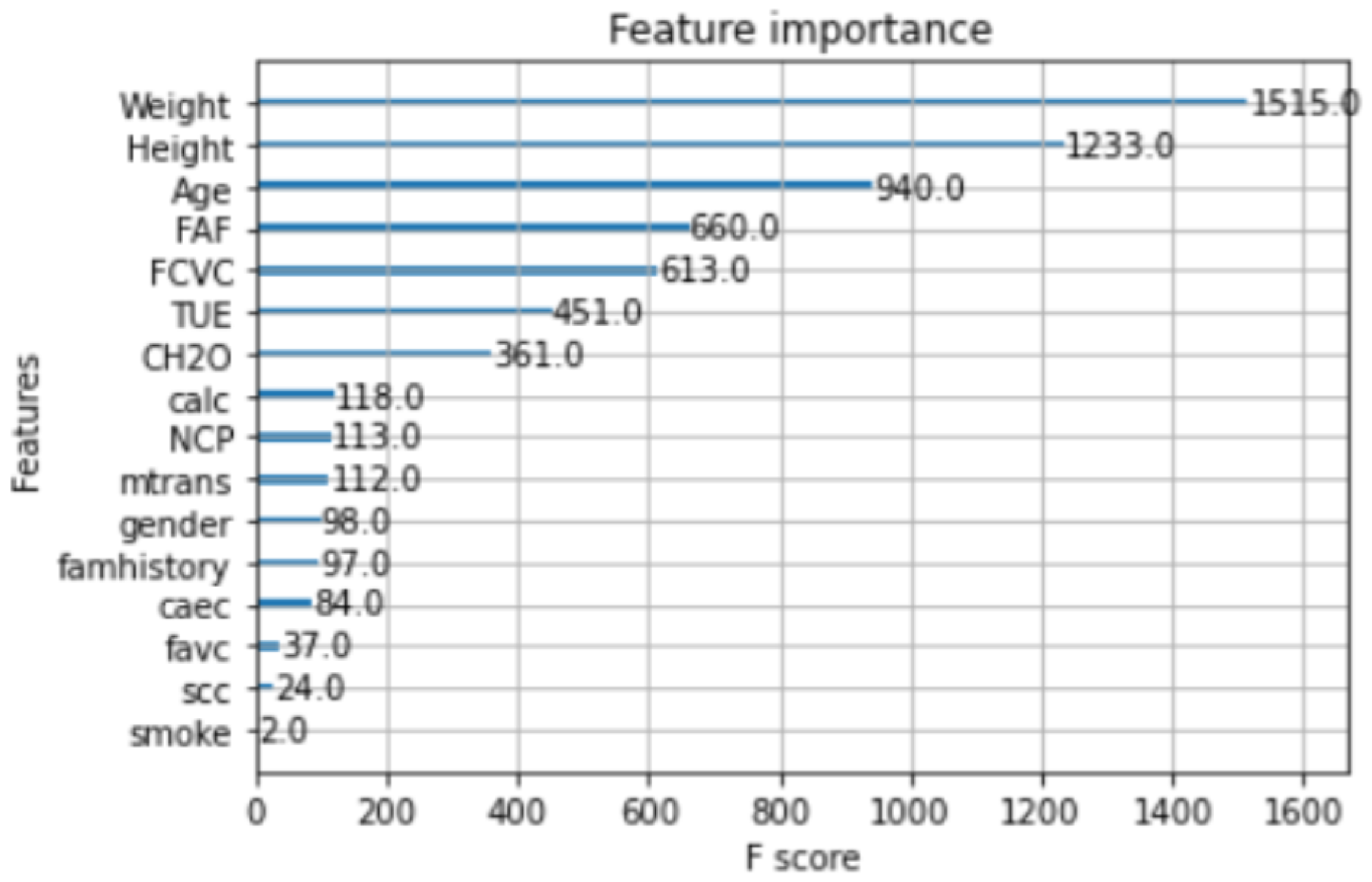
Architecture of Project

Results



Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE	gender	famhistory	favc	caec	smoke	scc	calc	mtrans	nobeyes
1.000000	-0.025958	0.202560	0.016291	-0.043944	-0.045304	-0.144938	-0.296931	0.048394	0.205725	0.063902	0.083739	0.091987	-0.116283	-0.044487	-0.601945	0.236170
-0.025958	1.000000	0.463136	-0.038121	0.243672	0.213376	0.294709	0.051912	0.618466	0.247684	0.178364	0.048818	0.055499	-0.133753	-0.129732	-0.073609	0.038986
0.202560	0.463136	1.000000	0.216125	0.107469	0.200575	-0.051436	-0.071561	0.161668	0.496820	0.272300	0.287493	0.025746	-0.201906	-0.206677	0.004610	0.387643
0.016291	-0.038121	0.216125	1.000000	0.042216	0.068461	0.019939	-0.101135	-0.274505	0.040372	-0.027283	-0.054670	0.014320	0.071852	-0.060781	0.064743	0.018522
-0.043944	0.243672	0.107469	0.042216	1.000000	0.057088	0.129504	0.036326	0.067600	0.071370	-0.007000	-0.097801	0.007811	-0.015624	-0.071747	-0.053858	-0.092616
-0.045304	0.213376	0.200575	0.068461	0.057088	1.000000	0.167236	0.011965	0.107930	0.147437	0.009719	0.144995	-0.031995	0.008036	-0.091386	0.044028	0.108868
-0.144938	0.294709	-0.051436	0.019939	0.129504	0.167236	1.000000	0.058562	0.189607	-0.056673	-0.107995	-0.030110	0.011216	0.074221	0.086799	0.006394	-0.129564
-0.296931	0.051912	-0.071561	-0.101135	0.036326	0.011965	0.058562	1.000000	0.017269	0.022943	0.068417	-0.048567	0.017613	-0.010928	0.045864	0.176945	-0.069448
0.048394	0.618466	0.161668	-0.274505	0.067600	0.107930	0.189607	0.017269	1.000000	0.102512	0.064934	0.091543	0.044698	-0.102633	0.007616	-0.137537	0.024908
0.205725	0.247684	0.496820	0.040372	0.071370	0.147437	-0.056673	0.022943	0.102512	1.000000	0.208036	0.169787	0.017385	-0.185422	0.036676	-0.101540	0.313667
0.063902	0.178364	0.272300	-0.027283	-0.007000	0.009719	-0.107995	0.068417	0.064934	0.208036	1.000000	0.150068	-0.050660	-0.190658	-0.089520	-0.069800	0.044582
0.083739	0.048818	0.287493	-0.054670	-0.097801	0.144995	-0.030110	-0.048567	0.091543	0.169787	0.150068	1.000000	-0.055282	-0.109179	-0.047540	-0.048535	0.327295
0.091987	0.055499	0.025746	0.014320	0.007811	-0.031995	0.011216	0.017613	0.044698	0.017385	-0.050660	-0.055282	1.000000	0.047731	-0.082471	-0.010702	-0.023256
-0.116283	-0.133753	-0.201906	0.071852	-0.015624	0.008036	0.074221	-0.010928	-0.102633	-0.185422	-0.190658	-0.109179	0.047731	1.000000	-0.003463	0.043157	-0.050679
-0.044487	-0.129732	-0.206677	-0.060781	-0.071747	-0.091386	0.086799	0.045864	0.007616	0.036676	-0.089520	-0.047540	-0.082471	-0.003463	1.000000	-0.012452	-0.134632
-0.601945	-0.073609	0.004610	0.064743	-0.053858	0.044028	0.006394	0.176945	-0.137537	-0.101540	-0.069800	-0.048535	-0.010702	0.043157	-0.012452	1.000000	-0.046202
0.236170	0.038986	0.387643	0.018522	-0.092616	0.108868	-0.129564	-0.069448	0.024908	0.313667	0.044582	0.327295	-0.023256	-0.050679	-0.134632	-0.046202	1.000000

Correlation matrix of all the features in the dataset



Feature importance graph shows the feature which has the highest correlation to obesity.

.

Comparative analysis of different performance metrics on all the algorithms

ML MODELS	accuracy	precision	recall	f1
SVM with radial basis kernel	77.3	76.7	77.3	76.8
SVM with linear kernel	96.9	96.9	96.7	96.8
eXtreme Gradient Boosting	97.6	97.6	97.55	97.57
Adaboost with Decision Tree	93.1	92.9	92.8	92.8
Adaboost with Random Forest	96.2	96.4	96.1	96.1
Random Forest Classifier	95.5	95.6	95.3	95.3
Gradient Boosting	95.9	95.8	95.83	95.83
Adaboost with Extra Tree classifier	93.6	93.6	93.3	93.4
Adaboost with SVM	82.2	82.5	82.09	82.1
Category Boosting (CatBoost)	97.8	97.9	97.8	97.8
Light Gradient Boosting Method	99.05	99.04	99.09	99.06

Conclusion

From our comparative analysis we see that the Light Gradient Boosting Method which gives an accuracy of 99.09% is the best ML algorithm for this prediction based model.

We have also optimized the results of other machine learning models to give a accurate prediction.

This product will help future developers use this prediction model to create product that can help people check their health digitally, hence contributing to smoother economic transactions.

References

1. & Sánchez Hernández, A. B. (2019). Obesity level estimation software based on decision trees.
2. Cervantes, R. C., & Palacio, U. M. (2020). Estimation of obesity levels based on computational intelligence. *Informatics in Medicine Unlocked*, *21*, 100472..