A Hybrid Machine Learning Model for Estimation of Obesity Levels

Akash Choudhuri
Roll: E1773U194002
M.Sc Capstone Project
Supervisor: Prof. Sudarsan Padhy

2nd September, 2021

Abstract

Obesity has always been a problem which has plagued humans for many generations, which, since the 1975, almost doubled to turn into a global epidemic. The current human dependence on technology has contributed to the problem even more, with the effects visibly pronounced in late teenagers and early adults. Researchers till date, have tried numerous ways to determine the factors that cause obesity in early adults.

On that frontier, our hybrid machine-learning model uses the help of some supervised and unsupervised data mining methods like Extremely Randomized Trees, Multilayer Perceptron and XGBoost using Python to detect and predict obesity levels and help healthcare professionals to combat this phenomenon. Our dataset is a publicly available dataset in the UCI Machine Learning Repository, containing the data for the estimation of obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.

Keywords: Obesity, Extremely Randomized Trees, Multilayer Perceptron, XGBoost.

Overview

- Introduction
  - Motivation
  - Related Works
  - The Problem Statement
  - Research Gap
- Hybrid Model
  - Hybrid Model Architecture
  - Results
  - Comparison of Results by the Hybrid Model with Previous Research Works
  - Why Choose a Hybrid Model?
- Performance of Model in Covid-19
  - The Modified Dataset
  - Results
- Conclusions
  - Future Direction of Research
  - Key features of this work
- References

Introduction

Motivation

- Obesity has become a global epidemic in this decade, that has doubled since 1980, with serious consequences for health in children, teenagers and adults.
- More than 190 million adults, 18 years old or older, are suffering from alteration of their weight. This has been the result of unhealthy lifestyle choices which included increased consumption of fast food, increased smoking frequency, and decrease in physical activity.
- Obesity has severe effects, ranging from cardiovascular issues to depression.
- Although much research has been performed to determine the effect of obesity on various age groups very little prior research has been done to quantify and analyze the effect of obesity in early adults.
- The lifestyle of early adults has been severely affected by the Covid-19 pandemic with them being highly susceptible to obesity due to lack of avenues of physical activity due to Covid-19 restrictions in public places.
- This study, thus aims at creating awareness about obesity in early adults.

Important Related Works

- In 2012, Adnan and Husain developed a model based on the Naïve Bayes' Model [1] with a hybrid approach using genetic algorithms to optimize factors used in the prediction of juvenile obesity, with a low percentage of negative samples vs. positive samples. They obtained 19 parameters to be implemented in prediction with a precision of 75%.
- In 2019, Eduardo De-La-Hoz-Correa et al. [2] presented data for the estimation of obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. They generated 77% of the data synthetically using the Weka tool and the SMOTE filter, and they collected 23% of the data directly from users through a web portal. They used the popular J48 Algorithm in decision trees using Weka and achieved precision of 97.4% and recall of 97.8%.
- In 2020, Cervantes et al. [3] proposed and compared a SVM Model with a Decision Tree model using the obesity dataset generated from the University students of Latin American countries and achieved high precision and recall values using the decision tree. Further, they proposed a Decision Tree + Simple K-Means model, where they achieved 98.5% precision and 98.5% recall.

The Problem Statement

- Given a dataset containing the lifestyle habits of young undergraduate students residing in Mexico, Peru, and Colombia, our job is to construct a mathematical/machine learning model to estimate and predict the level of obesity in them.
- We will also explore the relationship between the various lifestyle habits to obesity.
- In addition to that, we will perform scenario testing on the dataset with respect to Covid-19 and then try to observe the performance of our given model in this scenario.

Research Gap

- All the previous related works used tools like Weka. Although these tools do not need any sort of coding, the models created by these tools are not flexible and we cannot try to tune the various model parameters to improve performance. This can only be achieved by coding.
- The previous works of research mostly used common machine learning algorithms. However, current advancements in this domain have introduced various machine learning models which offer higher performance if their model parameters are tuned properly.
- It is also to be noted that the previous works did not focus on data preprocessing methods like feature selection and feature scaling methods. This also serves as an avenue to improve performance.

Hybrid Model

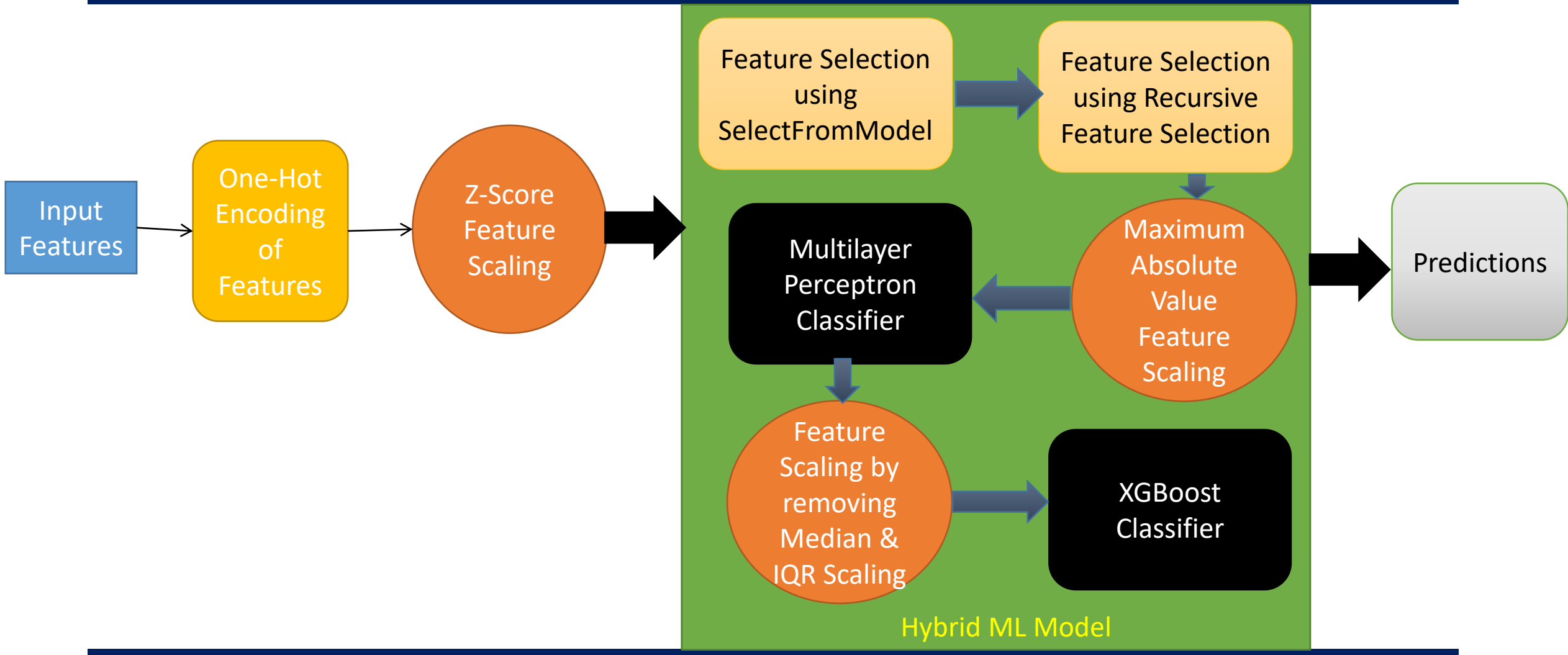
Some Concepts of Machine Learning Used

- **Classification** is the subset of supervised learning problems where the target variable is a qualitative variable (that is, the variables take on values in one of K different classes). Our problem is a **6-class classification problem**.
- To put all features into the same standing in many machine learning methods, we must scale so that one big number does not damage the model simply because of its high magnitude. This is **data scaling**.
- The dimensionality of a dataset is defined as the number of input variables or characteristics. **Dimensionality reduction** techniques are those that reduce the number of variables in a dataset.
- **Multilayer Perceptron (MLP)** [5] is a deep learning model consisting of fundamental parts called neurons, which arrange themselves to form layers network and the activation function determines the output of each neuron given a set of inputs. A typical MLP architecture consists of an input layer, few hidden layers and an output layer. Our MLP **has 1 hidden layer with 100 neurons in it**.
- **XGBoost** [6] is a scalable Tree Boosting System. It is actually the implementation of gradient boosting trees designed for speed and performance.

Hybrid Model Architecture

- The string-type categorical features as well as the target variable (type of obesity) in the data is One-Hot Encoded (ie, creating a binary expression of all the string-type categorical features). We then get a data whose all features are either Integer- type or Float-type.
- We apply z-score feature transformation to all the float type feature columns except the columns which have been One-Hot Encoded.
- We perform initial feature selection using the SelectFromModel method. The tree-based algorithm that was used was Extremely Randomized Trees Algorithm.
- We perform another set of feature selection process using the Recursive Feature Elimination method. The tree-based algorithm that was used was Extremely Randomized Trees Algorithm.
- We then again scale the features of the data by scaling it down to the maximum absolute value scaling.
- The scaled features are then passed to the Multilayer-Perceptron Classifier. The Multilayer Perceptron Classifier does an initial classification of 6 classes with intermediate float values.
- The predictions of the Multilayer-Perceptron are again scaled by removing the median and scales the data according to the quantile range (defaults to IQR: Inter-quartile Range). The IQR is the range between the 1st quartile (25^{th} quantile) and the 3rd quartile (75^{th} quantile).
- The scaled outputs of the Multilayer Perceptron are then fed into the XGBoost Model and the predictions of the overall model are then generated.

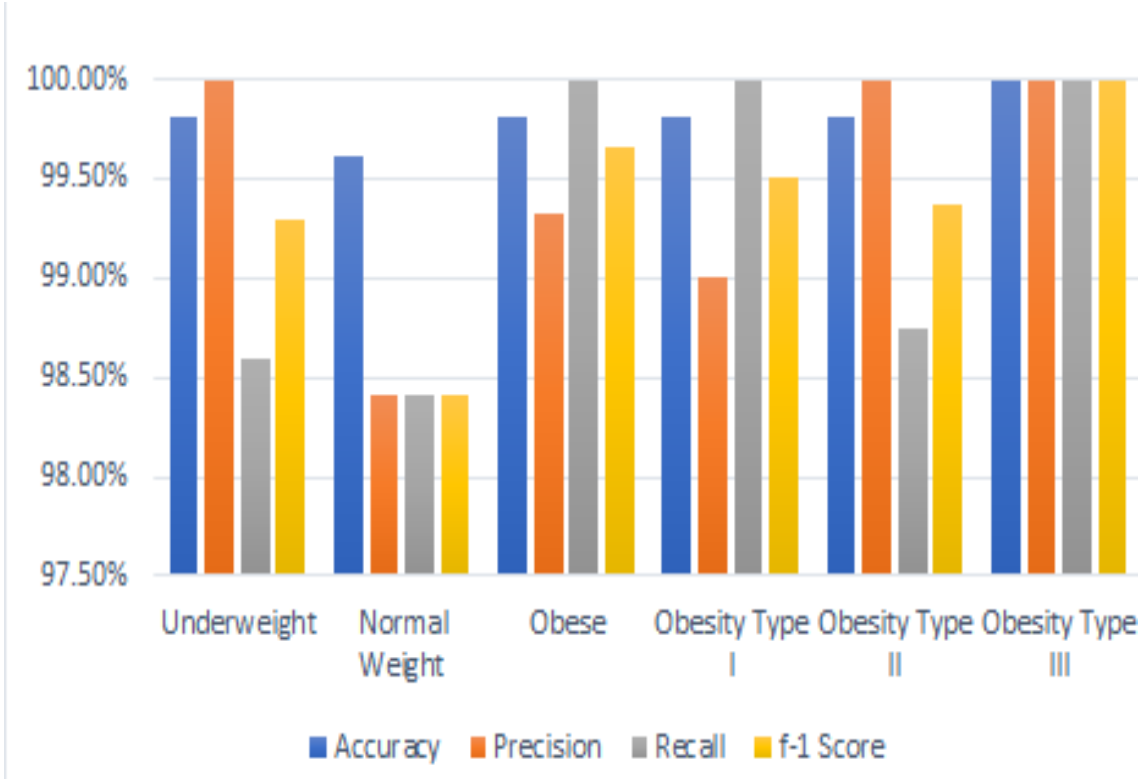
Hybrid Model Architecture cont.



Model Parameters Used

Model Part	Parameter Values
Feature Selection (Select From Model)	Extremely Randomized Trees Classifier: • Number of trees in forest = 100. • Criterion = Gini. • Maximum Number of Features to consider best split = 0.1
Feature Selection (Recursive Feature Elimination)	Extremely Randomized Trees Classifier: • Number of trees in forest = 100. • Criterion = Gini. • Maximum Number of Features to consider best split = 0.4
Multilayer Perceptron Classifier	•Number of hidden layers = 1. • Number of neurons in hidden layer = 100. • Optimizer = Adam (Learning rate = 0.5). • L2 Regularization Parameter value = 0.1.
XGBoost Classifier	• Learning Rate = 0.5. • Maximum Depth of a Tree = 7. • Minimum sum of instance weight (hessian) needed in a child = 1. • Subsample ratio of the training instances = 0.8500000000000001

Results

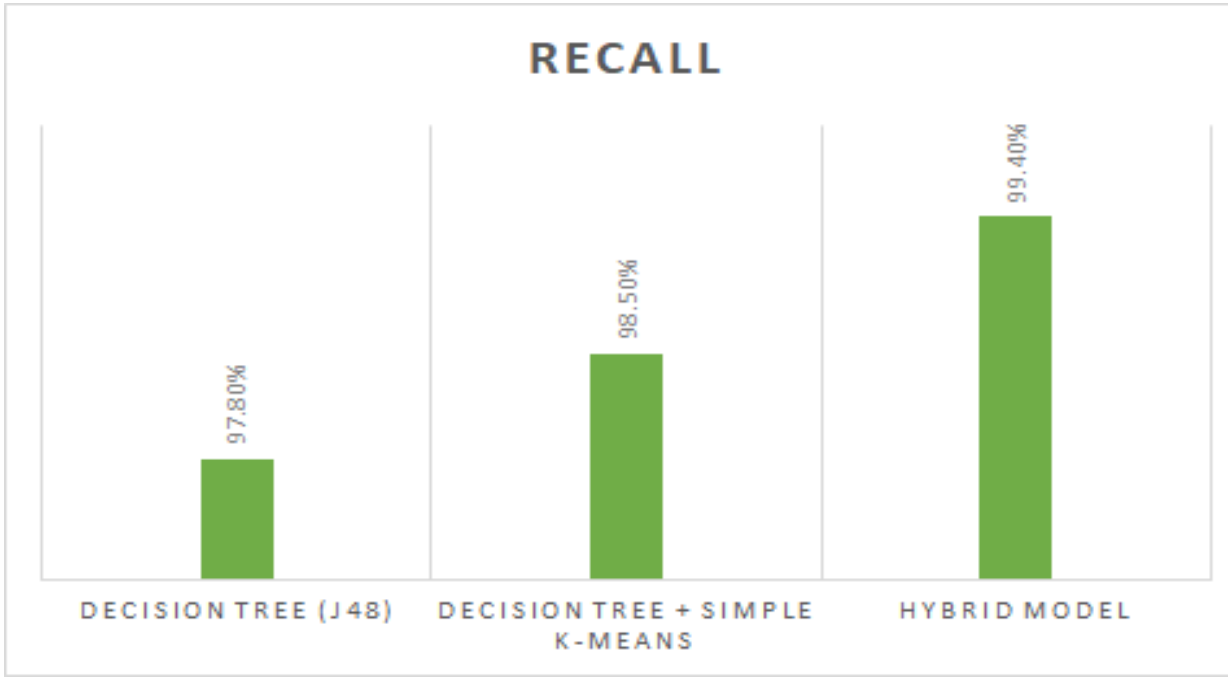


Class wise prediction performance measures of the model

		Predicted Class					
		Underweight	Normal Weight	Obese	Obesity Type I	Obesity Type II	Obesity Type III
Actual Class	Underweight	70	1	0	0	0	0
	Normal Weight	0	62	1	0	0	0
	Obese	0	0	149	0	0	0
	Obesity Type I	0	0	0	101	0	0
	Obesity Type II	0	0	0	1	79	0
	Obesity Type III	0	0	0	0	0	64

Confusion Matrix of the predictions made by the model

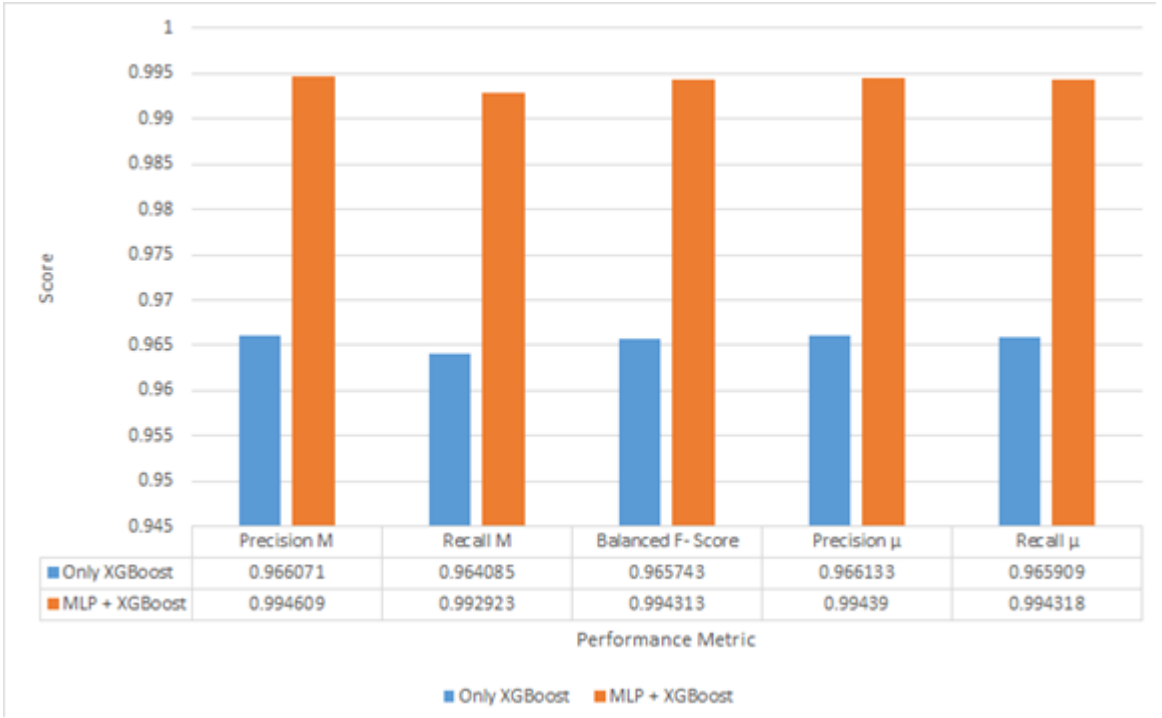
Comparison of Results to Previous Works



Comparison of Prediction Performance Metrics between the Champion (proposed model)-Challenger (models proposed in [1] and [2]) Models.

Why Choose a Hybrid Model?

The idea of choosing the XGBoost model germinated by its frequent usage in Kaggle competitions to increase performance. However, Multilayer Perceptron added to the architecture seems to offer much better performance than that given by the XGBoost Model. This is shown below:



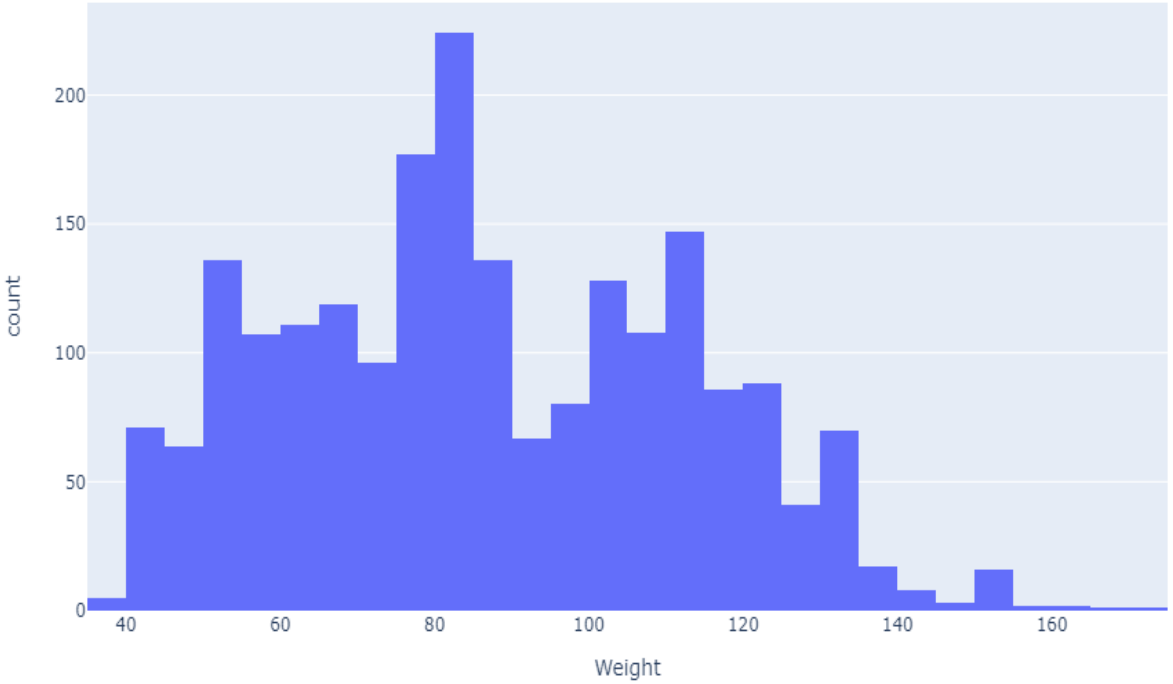
	Precision M	Recall M	Balanced F- Score	Precision μ	Recall μ
Only XGBoost	0.966071	0.964085	0.965743	0.966133	0.965909
MLP + XGBoost	0.994609	0.992923	0.994313	0.99439	0.994318

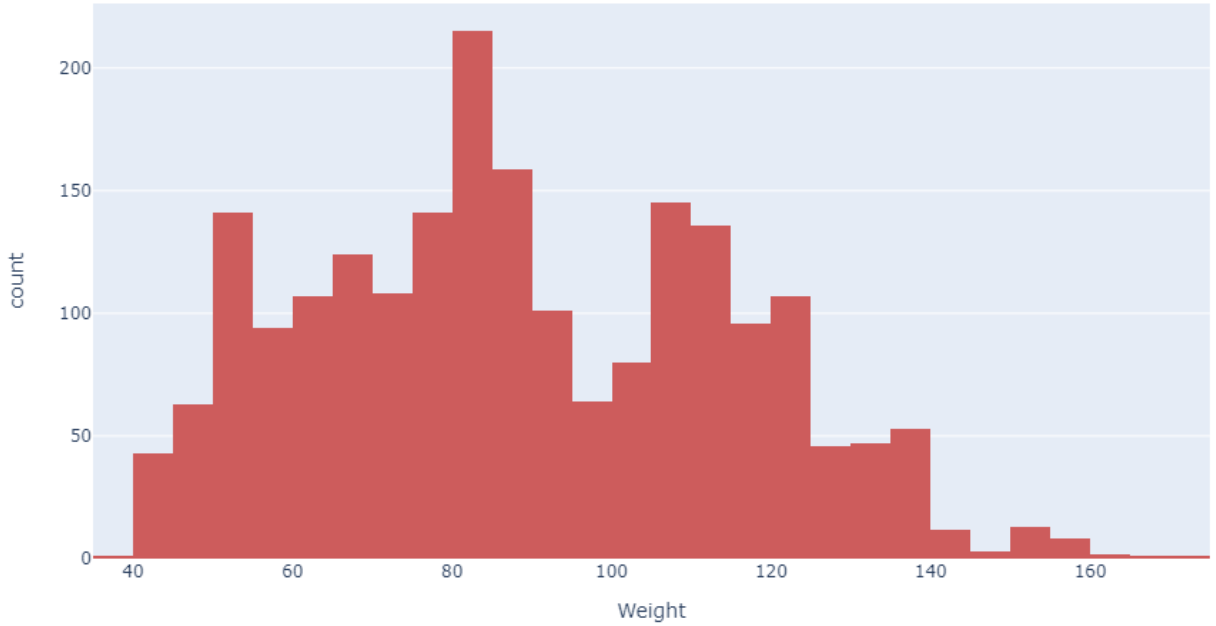
The Performance of the Model in Covid-19

The Modified Dataset

- According to Lin et al. study [4] on a total of 7444 weight measurements from 269 unique study participants (residing in 37 states and Washington, District of Columbia, USA), post shelter in place participants experienced steady weight gain of 0.27Kg every 10 days irrespective of geographic location or co morbidities.
- So, we decided to use the findings of [4] and assumed the period of Covid-19 to be on peak and lockdowns in various countries to be actively during the time Covid-19 peaked in Latin America.
- Thus, we are considering the period of lockdown due to Covid-19 to be held from 20th March, 2020 up till 1st September, 2020 and using the value of average weight gain mentioned in [4] by initializing a random multiplier between 0-0.27Kg of every 10 days for the given period.

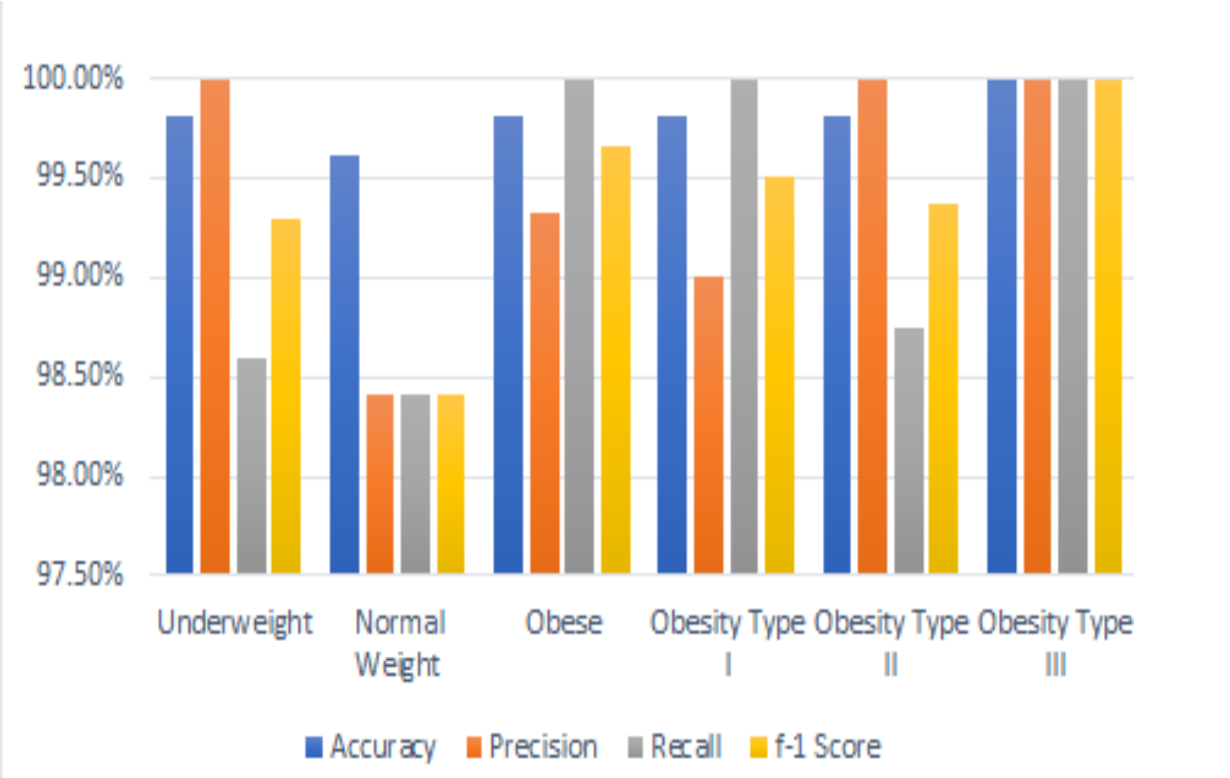
The Modified Dataset cont.



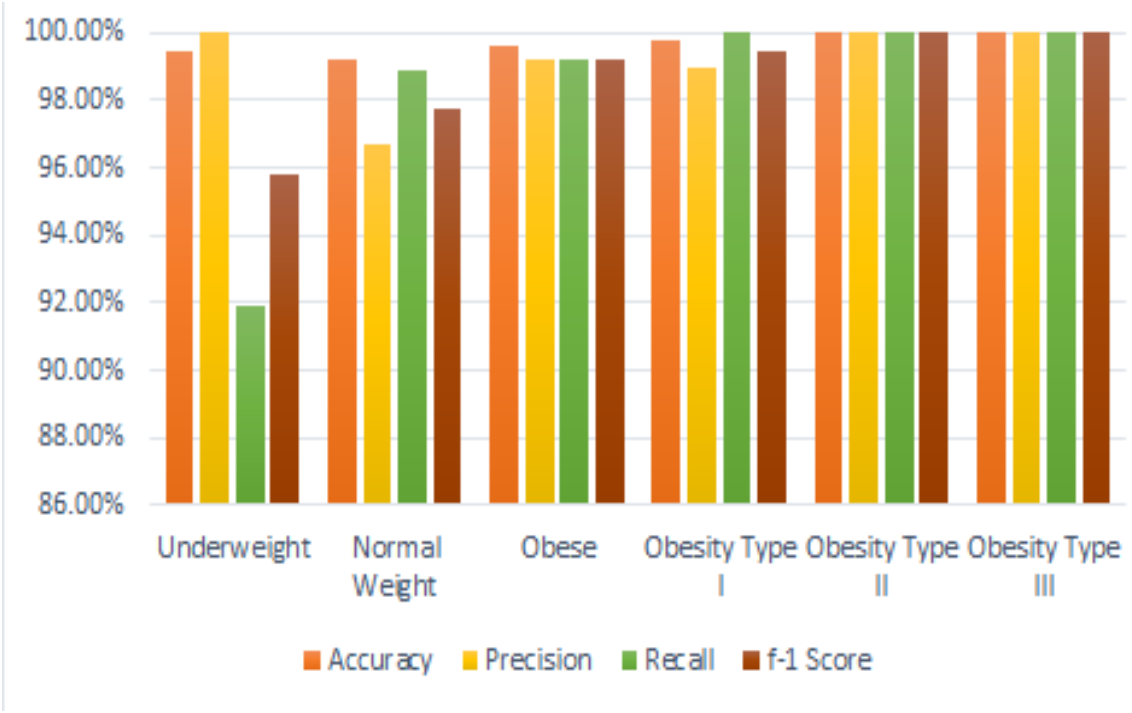


Histogram of Weights before Covid-19 (shown in blue) and after Covid-19 (shown in red)

Results



Class wise prediction performance measures of the model in original case



Class wise prediction performance measures of the model in Covid-19

Conclusions

Future Direction of Research

- An attempt to further improve classification quality on this dataset could be made by aggregating and fine tuning more Machine Learning models into the model pipeline.
- Efforts can be made to present more publicly available datasets on obesity, which would be homogenous in terms of the features and descriptions, which can inspire further research in this area.
- Some of the used machine learning approaches could not be optimized as extensively as it would be desirable due to computation resource limitations. Therefore, especially concerning Multilayer Perceptron (Artificial Neural Network), a more in-depth optimization process would be desirable to come closer to a global optimum in terms of classification results.

Key Features

- In our model pipeline, we utilized a plethora of dimensionality reduction and feature scaling methods. We also used two widely popular prediction supervised learning methods, namely the Multilayer Perceptron Model (also known as Artificial Neural Network) and the tree-based ensemble technique XGBoost.
- Our constructed model gave a prediction accuracy of 99.43%, a precision value of 0.994 and a recall value of 0.994, which, by all means, were sizeable improvements from the performance metrics generated from previous works of research working with the same dataset.
- By stress testing on the context of Covid-19, we also proved that our model is valid and very flexible to changes and this work has been able to create a newer Machine Learning Method to predict obesity in human beings.

References

1. Adnan MHBM, Husain W. A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction". In: Computer & information science (ICCIS), 2012 international conference on, vol. 1. *IEEE; 2012. p. 281–5.*
2. De la Hoz Manotas, Alexis & De la Hoz Correa, Eduardo & Mendoza, Fabio & Morales, Roberto & Sanchez, Beatriz. (2019). Obesity Level Estimation Software based on Decision Trees. *Journal of Computer Science. 15. 10. 10.3844/jcssp.2019.67.77*.
3. Cervantes, Rodolfo & Palacio, Ubaldo. (2020). Estimation of obesity levels based on computational intelligence. Informatics in Medicine Unlocked. 21. 100472. 10.1016/j.imu.2020.100472.
4. Lin AL, Vittinghoff E, Olgin JE, Pletcher MJ, Marcus GM. Body Weight Changes During Pandemic-Related Shelter-in-Place in a Longitudinal Cohort Study. *JAMA Netw Open. 2021;* 4(3):e212536. doi:10.1001/jamanetworkopen.2021.2536.
5. Haykin, Simon S. *Neural Networks: A Comprehensive Foundation.* Upper Saddle River, N.J.: Prentice Hall, 1999.
6. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining (pp. 785–794).* New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Thank You

Questions/Queries?
Feel Free to ask