Al-Imam University
Computer Science Department
College of Computer and Information Sciences
CS364: Machine Learning
2nd Semester 2022 / 1443 H

Project Report:

Obesity Level Estimator

Group members:

Student Name	Student ID	Email
Nouf Mohammed Alajmi	441022333	nomoalajmi@sm.imamu.edu.sa
Abeer Mohammed Aldosari	440023513	amsaldosari13@sm.imamu.edu.sa
Razan Saad Alabulkarim	440021710	rsaalabdulkrim@sm.imamu.edu.sa

Due date: 14th May 2022
Section code: 371

Supervisor:

Dr. Wojdan BinSaeedan

Table of contents:

1. Introduction:

Obesity, one of the dangerous diseases in the world. It's caused by eating too much and moving a little, which is something that could anybody do. Some other cases are caused by other reasons like having another disease. In this project, we implemented a Machine Learning (ML) model to diagnose the level of obesity of people. The dataset is taken from Kaggle [1], which is dataset that is collected from individuals from the countries of Mexico, Peru, and Colombia.

2. Problem formulation:

In this section, we will discuss the details of problem formulation and the chosen dataset.

2.1 Problem definition:

The problem is to estimate the level of obesity of an individual, based on their eating habits, physical condition, and many health problems such as diabetes, heart disease, and some cancers.

2.2 Initial State:

Initial step is to prepare the dataset. The "Level of Obesity" dataset from Kaggle consists of **2113** records and **16** features. The features are:

no.	Features
1	Gender
2	Age
3	Height
4	Weight
5	Smoke
6	Family history with overweight
7	Calories consumption monitoring (SCC)
8	Physical activity frequency (FAF)
9	Time using technology devices (TUE)
10	Transportation used (MTRANS)
11	Frequent consumption of high caloric food (FAVC)
12	Frequency of consumption of vegetables (FCVC)
13	Number of main meals (NCP)
14	Consumption of food between meals (CAEC)
15	Consumption of water daily (CH20)
16	Consumption of alcohol (CALC)

Table 1: Features Table

2.3 Action:

Our main goal is to build a ML model, train it and make it predict or estimate the level of individuals. We chose to use Random Forest model; the implementation details are in section (4.2).

2.4 Goal state:

Our aim is to estimate the level of obesity of individuals, the levels are:

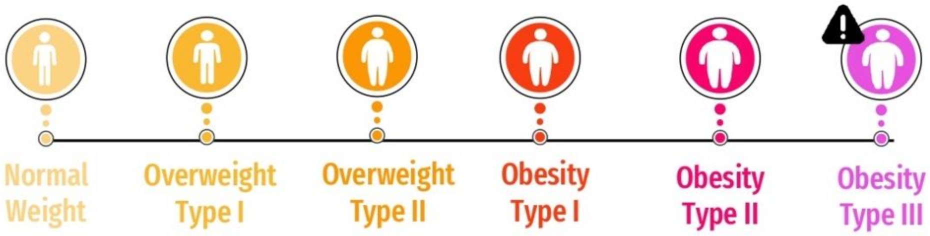

Figure 1: Obesity Estimations

In addition, there is an "insufficient weight", which indicates that the model does not have information to determine the obesity level.

3. ML model: Random Forest

In this project, we chosen Random Forest classification model which is one of the supervised machine learning algorithms that are widely used in classification and regression. Instead of a single decision tree, it is implemented into several decision trees, and based on each tree of its expected result, it takes the most votes. One of the most important features of a random forest is that it can dealing with a data set that contains continuous variables in regression and categorical variables in classification. We chosen Random Forest because the flexibility of implementation and our past knowledge. Also, it results in good performance in our dataset.

4. Implementation details:

In this section, we will discuss the implementation steps of the Random Forest model along with the Level of Obesity dataset.

4.1 Programming language and environment:

We built our model using Python programming language, due to its simplicity and flexibility. Also, it includes several helpful libraries to implement ML models effectively with simple commands.

For programming environment, we used Jupyter Notebook to implement our model along with detailed notes, then upload it in Google Colab. You can see Instructions on compiling/running the programs in the attached README file.

4.2 Implementation steps:

You can see the implementation steps of out model in the attached Google Colab link, which contains the source code of the built model along with detailed comments and steps.

5. Results and discussion:

In our project, we built a Random Forest model to estimate the obesity level of individuals. We faced some challenges and problems, which will be discussed in this section.

5.1 Feature selection:

In the process of preparing the dataset, we used the feature selection technique, which is a common technique used in ML models to reduce the number of features and only use the important features. This approach is recommended to reduce the size of the model which affects the time complexity or running the model.

After performing feature selection, the figure bellow shows the feature's level of importance. Some features are strongly affecting the prediction value while others barely relied on while training the model.

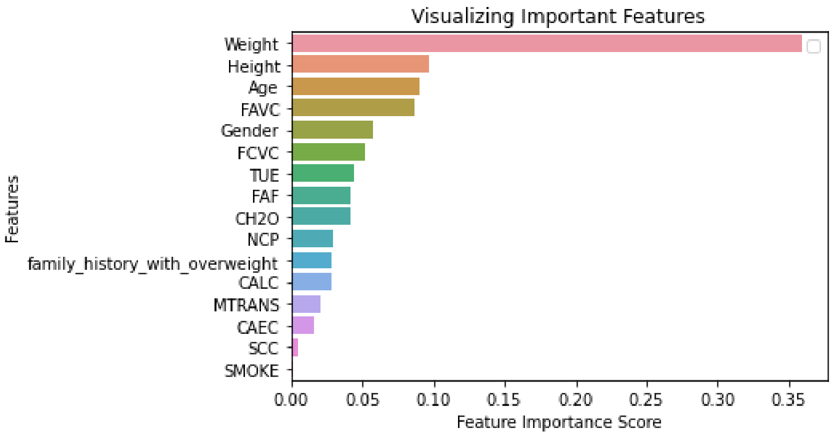

Figure 2: Important Features

We built our model **with** and **without** feature selection to set the difference of performance between them, and that using feature selection doesn’t affect the model's performance negatively. We tested their performance for 10 times and take the average of accuracy.

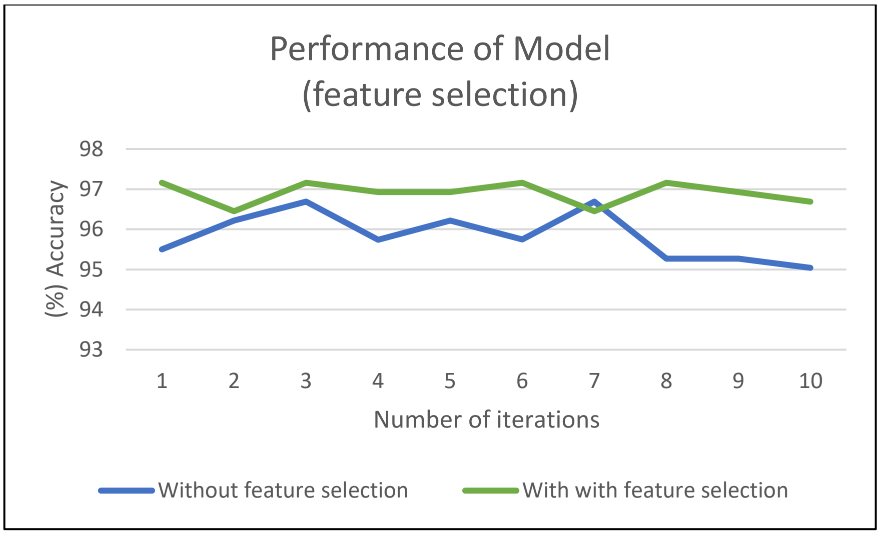

Figure 3: Performance of model (feature selection)

Avg accuracy without feature selection: 95.84 %
Avg accuracy with feature selection: 96.90 %

The results show that using feature selection improves the performance of the model.

5.2 Hyperparameter tunning (GridSearchCV):

After applying feature selection, we want to perform a hyperparameter tunning step which is to chose the best parameters to be used in building the model in order to improve the performance and avoid overfitting. We used GridSearchCV famous technique in this step.

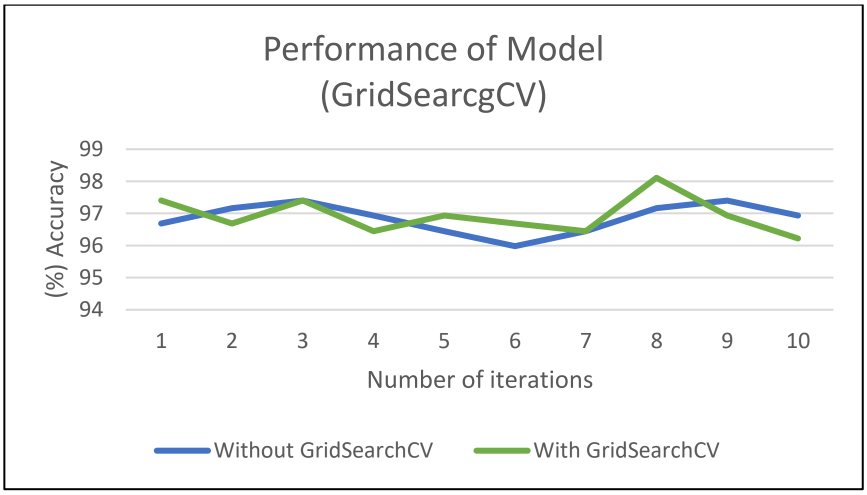

Figure 4: Performance of model (GridSearchCV)

Avg accuracy without GridSearchCV: 96.85 %
Avg accuracy with GridSearchCV: 96.93 %

The results show that using GridSearchCV improves the performance of the model with little differences.

5.3 Overall performance:

We used the GridSearchCV technique in order to improve the performance as well as avoiding overfitting. However, after running the model for several times it always results in 100% accuracy of training. Only one time we have captured the training reduced to 99%. It seems that the dataset itself is clean and does not include outliers and noisy data. It shows good training performance on average of 96.93% which indicates that the model is trained well and can predict values effectively.

6. Conclusion:

In conclusion, we presented in this project an Obesity Level Estimator application using Random Forest mode. We faced several challenges since it is our first time to deal with Random Forest model, also the time taken to find a suitable dataset to implement our project. Overall results shows that the model shows good prediction results with average of 96% accuracy.

7. References:

[1] "Obesity Levels & Life Style | Kaggle." https://www.kaggle.com/code/mpwolke/obesity-levels-life-style/notebook (accessed April 29, 2022).

8. Appendix:

Along with this pdf folder, there are multiple files, which are the Google Colab link, which includes the source code and the GitHub link which contains all the documents along with the README file.