Estimation Of Obesity Levels Based On Eating Habits And Physical Condition

Gamze Aksu
171180005
Computer Engineering
Gazi University

Abstract — **Estimation of obesity levels based on eating habits and physical condition dataset is examined. With four different classification algorithms, it is estimated whether people are obese or not with the data in the dataset. K-Nearest Neighbors algorithm, Support Vector Machines, Decision Tree algorithm and Random Forest algorithm were used. Two different encoding methods were applied to the data. These are Label Encoding and One Hot Encoding. Among these algorithms, the algorithm that gives the best score value is the Decision Tree with Label Encoding applied to the data. The algorithm that gives the worst score value is the Decision Tree algorithm with One Hot Encoding applied to the data.**

Keywords—Python, Data, Mining, Obesity

I. Introduction

In this project, Estimation of obesity levels based on eating habits and physical condition dataset [1] are examined. There are 2111 in the dataset. There are seven different obesity levels to estimate. These are as follows:

- Insufficient Weight,
- Normal Weight,
- Overweight Level I,
- Overweight Level II,
- Obesity Type I,
- Obesity Type II,
- Obesity Type III.

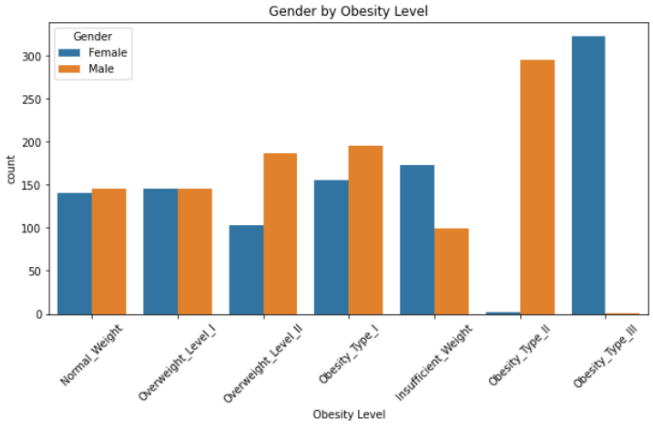
This data set was created by collecting the answers to 16 different questions asked to people on a website. In the data set, the BMI values of the people were calculated by taking the age, height and weight values of the people. According to the BMI values, the data is labelled whether it is obese or not. Age, height and weight values in the data set are deleted. Is it possible to predict whether people are obese by only asking questions about their health? The answer to this question is explored in this paper. [2]

II. Purpose

The genetic characteristics of people, their diet, and their daily movements also affect their health. In today's world, there is a problem of obesity due to social life. This project is carried out in order to make an analysis of this problem.

III. Research Questions

1. What is the distribution of the number of men and women at each obesity level?

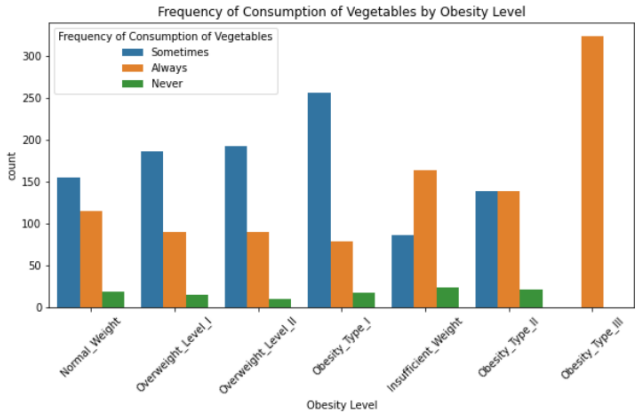


Although the number of male and female at other obesity levels is approximately close, there is a large difference in the number of males and females at Obesity Type 2 and Obesity Type 3 levels. When Obesity Type 3 is filtered only for women and the number of women's obesity levels is examined, it is seen that Obesity Type 3 has the highest number.

Similarly, when only men are filtered and the obesity level numbers of men are examined, it is seen that the biggest part is Obesity Type 2.

2. What is the distribution of the number of frequency of consumption of vegetables at each obesity level?

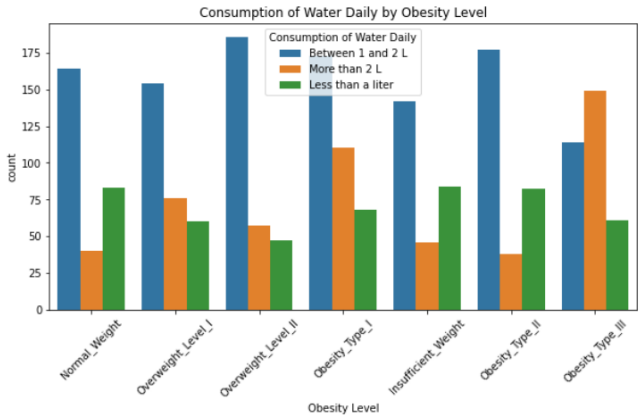
As can be seen in the graph, Never is the lowest value in every Obesity level, which is normal, but it is very strange that there is only Always in Obesity Type 3 level.



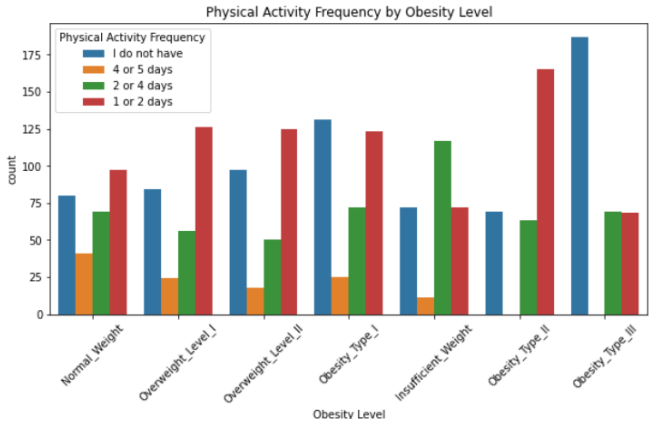
3. What is the distribution of the number of consumption of water daily at each obesity level?

Although the amount of water consumed daily is between 1 and 2 liters, we observe that the amount of water consumed daily at high levels such as obesity type 2 is low.

But strangely enough, the water level consumed daily is quite high at the level of obesity type 3.

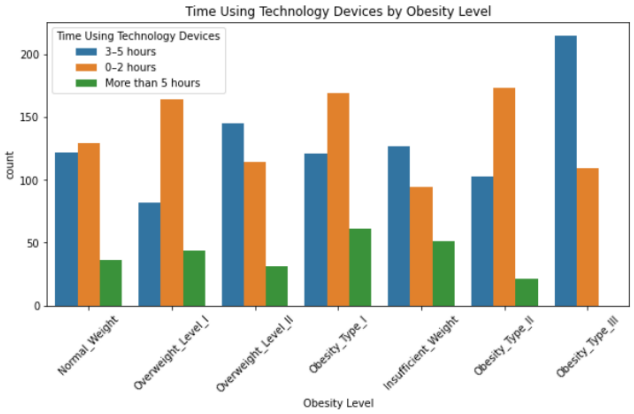


4. What is the distribution of the number of physical activity frequency at each obesity level?



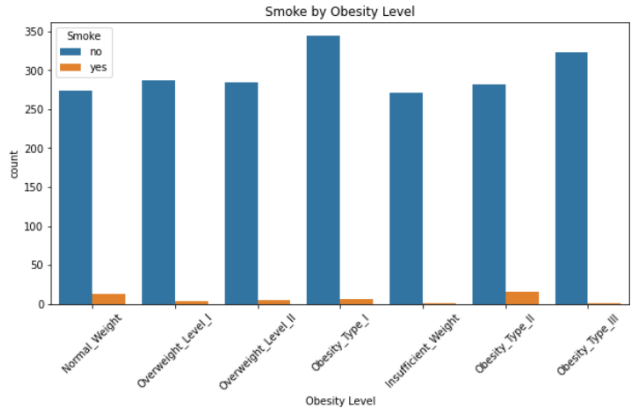
In obesity type 2 and obesity type 3 levels, the number of physical activity for 4 or 5 days is zero, but as the frequency of physical activity decreases, the level of obesity increases. It is observed that the number of physical activities for 1 or 2 days is high, but this was not found in the levels of insufficient weight and obesity type 3.

5. What is the distribution of the number of time using technology at each obesity level?

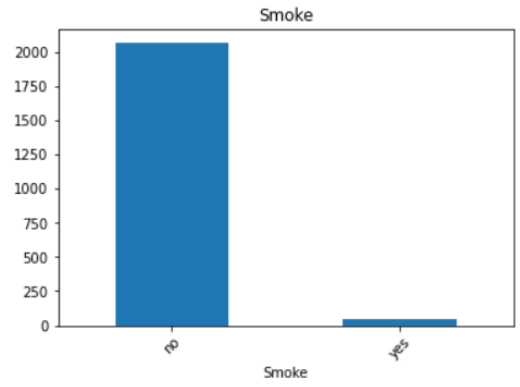


Using a technological device between 0 and 2 hours can be seen as normal for a person. While the number of people between 0-2 hours at each level is the highest, it is 3-5 hours more in overweight level 2 and obesity type 3 groups. There is no person who uses technology for more than 5 hours at the level of obesity type 3 in a way. However, it can be said that the use of technological devices does not have an effect on obesity.

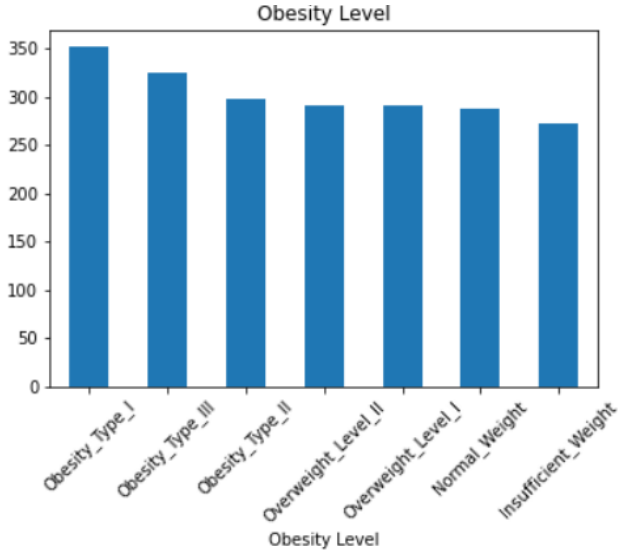
6. What is the distribution of the number of smoke at each obesity level?



There are many no answers at each level and number of yes answers is almost zero. It cannot be said that this result has much benefit in the training of the model. Looking at the smoking graph across all data shows the same result.



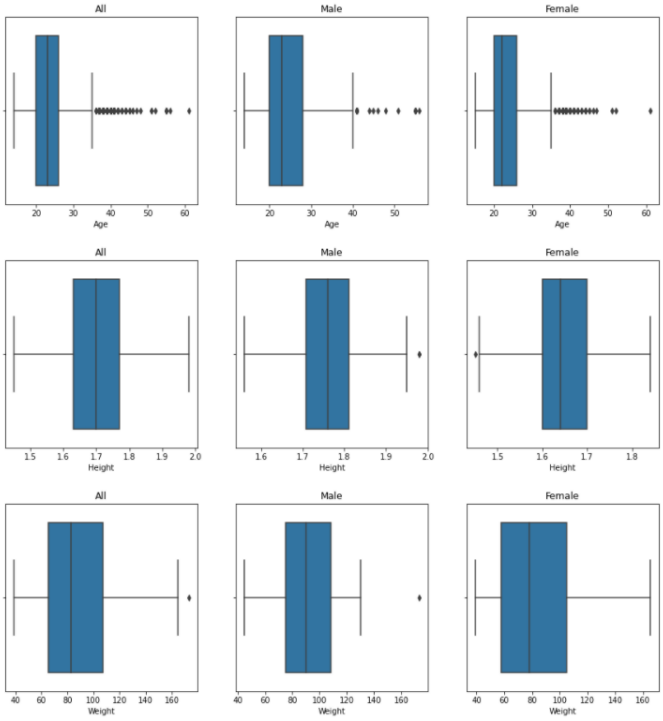
7. What is the distribution of obesity levels?



It seems that there is approximately the same number of data from each level, which indicates that the data is distributed properly. It is useful for training the model.

IV. Outlier

Outlier data in the dataset is shown with the help of boxplot. Outliers were visualized in three separate groupings: men, women, and all data. Here, the points outside the box plot represent the outlier points. Normally this outlier data needs to be cleaned. However, since the number of data is already very small, these data will not be removed within the project.



V. Methodology

After Dataset was classified by K-nearest neighbors, support vector machines, decision tree and random forest classification algorithms.

In order to find the best parameters of these algorithms, the Grid Search CV class was used. When a model is created, the model has many parameters. It is not known which of these different hyper parameters is the best. But experienced people can guess. Grid Search aims to find the best hyper parameter values by trying all the parameters given as a dictionary in model training. Cross Validation is also used. The CV value in the project is set to 10. The data is divided into 10 parts and one part at a time is separated as a test. In order for the model not to see all the data, the number of test data is determined as 10% of all data.

In addition, two different techniques were used to convert data from text to numbers. These are Label Encoding and One Hot Encoding. In Label Encoding, all values in a column are converted to numbers starting from 0 in alphabetical order. Since the "Number of Main Meals" attribute value in this data already consists of numbers and its order is not disturbed, it is not included in this Encoding. In One Hot Encoding, each unique value in a column is added to the data as another column. In short, in Label Encoding, values in a column are converted into numbers alphabetically, while in One Hot Encoding, each unique value is added as another attribute.

K-nearest neighbors algorithm is a supervised learning algorithm. The K value is the number of nearest neighbors. For example, if k = 5, classification is made according to its 5 nearest neighbors. Finds nearest neighbors based on different distance metrics such as Euclid, Manhattan and Minkowski.

The support vector machine algorithm is a supervised learning algorithm. Draws a line to separate points placed on a plane. This line needs to be drawn furthest from each point class.

Decision tree algorithm is a supervised learning algorithm. As like its name, it is a tree-based algorithm. It works by dividing the data into small clusters. It works for both numeric and categorical data. Decision trees use different algorithms such as gini and entropy to divide data into small clusters. One of the biggest problems of decision trees is overfitting. It memorizes the data when the data is divided too much.

Random forest algorithm is a supervised learning algorithm. It consists of randomly generated decision trees. It is an Ensemble algorithm. It finds a solution to the overfitting problem of the decision tree algorithm. Because the prediction with the highest vote is selected as the final prediction from the random forest algorithm.

The metrics used are:

- Accuracy is the percentage of samples correctly classified.
- Recall is a metric that shows how many of the transactions that should have been positively predicted were positively predicted. It is an important parameter to consider when the costs of False Negative are high.
- Precision shows how many of the positively predicted values are actually positive. It is an important parameter to be considered when the costs of False Positive are high.
- F1score is the harmonic average of precision and recall values. It prevents incorrect model selection in unevenly distributed data sets.

VI. Findings

After all the trainings were done, Decision Tree got the highest score value with the train data with Label Encoding applied. Accuracy score value is 0.69. Then comes the Random Forest algorithm in the second place. Accuracy values are 0.69 with Label Encoding and 0.68 with One Hot Encoding. The algorithm with the worst accuracy score is Decision Tree, which has 0.64 accuracy.

Encoding	Model	Score
Label	Decision Tree	0.693396
Label	Random Forest	0.688679
OHE	Random Forest	0.679245
OHE	SVC	0.674528
Label	SVC	0.674528
OHE	KNN	0.669811
Label	KNN	0.655660
OHE	Decision Tree	0.641509

Let's examine all the algorithms that have been applied, from the one with the highest score value to the one with the lowest score value.

A. *Decision Tree*

The Decision Tree algorithm has the highest accuracy score. First of all, after the data were digitized with Label Encoding, the Decision Tree algorithm was applied with Grid Search CV. The best score value in the Grid Search CV algorithm was 0.66. As a result, the best parameters were determined as

- criterion : gini,
- max_leaf_nodes: 92,
- min_samples_split: 4,
- splitter: random.

Considering the Precision and Recall values, a good classification was made for Obesity Type 3 and the FN value was zero. For others, classification is not as good as Obesity Type 3. Looking at the macroaverage value, precision and recall averages are 0.69.

Decision Tree Classification Report

	precision	recall	f1-score	support
Insufficient_Weight	0.75	0.69	0.72	26
Normal_Weight	0.53	0.50	0.51	20
Obesity_Type_I	0.59	0.69	0.64	42
Obesity_Type_II	0.76	0.93	0.84	30
Obesity_Type_III	0.94	1.00	0.97	30
Overweight_Level_I	0.67	0.56	0.61	36
Overweight_Level_II	0.57	0.43	0.49	28
accuracy			0.69	212
macro avg	0.69	0.69	0.68	212
weighted avg	0.69	0.69	0.69	212

B. *Random Forest*

The Random Forest algorithm takes the second place for the data with Label Encoding applied. The best score value in the Grid Search CV algorithm was 0.71. The best parameters obtained by the Random Forest algorithm as a result of the Grid Search CV algorithm were determined as follows:

- criterion: entropy,
- max_depth: 8,
- max_features: auto,
- n_estimators: 200.

Considering the Precision and Recall values, a good classification was made for Obesity type 3 again. When looking at other Obesity Levels, it is seen that some of them increase while others decrease. As a result, precision and recall macroaverage values are 0.69.

Random Forest Classification Report

	precision	recall	f1-score	support
Insufficient_Weight	0.79	0.73	0.76	26
Normal_Weight	0.58	0.55	0.56	20
Obesity_Type_I	0.64	0.60	0.62	42
Obesity_Type_II	0.60	0.93	0.73	30
Obesity_Type_III	0.91	1.00	0.95	30
Overweight_Level_I	0.69	0.50	0.58	36
Overweight_Level_II	0.62	0.54	0.58	28
accuracy			0.69	212
macro avg	0.69	0.69	0.68	212
weighted avg	0.69	0.69	0.68	212

C. *SVC*

Since the algorithm with the third highest accuracy score is the Random Forest algorithm again, this algorithm has been skipped. The algorithm with the fourth highest accuracy score is the SVC algorithm. This algorithm is trained for data with One Hot Encoding applied to this value. The best score value in the Grid Search CV algorithm was 0.74. The best parameter values from the Grid Search CV algorithm are as follows:

- C: 1,
- degree: 4,
- gamma: scale,
- kernel: poly.

Looking at the Precision and Recall values, it is seen that the highest values are again at the Obesty Type 3 level. Then the highest value is Obesity Type 2 level.

Looking at the macroaverage values for Precision and Recall values, they were 0.67 and 0.68, respectively.

SVC Classification Report

	precision	recall	f1-score	support
Insufficient_Weight	0.80	0.77	0.78	26
Normal_Weight	0.40	0.50	0.44	20
Obesity_Type_I	0.62	0.50	0.55	42
Obesity_Type_II	0.74	0.87	0.80	30
Obesity_Type_III	0.97	1.00	0.98	30
Overweight_Level_I	0.56	0.50	0.53	36
Overweight_Level_II	0.60	0.64	0.62	28
accuracy			0.67	212
macro avg	0.67	0.68	0.67	212
weighted avg	0.68	0.67	0.67	212

D. K-Nearest Neighbors (KNN)

In the last place is the KNN algorithm. This KNN algorithm is trained with One Hot Encoding applied data. The best score value in the Grid Search CV algorithm was 0.70. The best parameters in the Grid Search CV algorithm are:

- algorithm: auto,
- metric: manhattan,
- n_neighbors: 7,
- weights: distance.

The macroaverage values for the Precision and Recall values were 0.63 and 0.66, respectively.

KNN Classification Report

	precision	recall	f1-score	support
Insufficient_Weight	0.70	0.81	0.75	26
Normal_Weight	0.31	0.20	0.24	20
Obesity_Type_I	0.65	0.57	0.61	42
Obesity_Type_II	0.68	0.87	0.76	30
Obesity_Type_III	0.81	1.00	0.90	30
Overweight_Level_I	0.67	0.61	0.64	36
Overweight_Level_II	0.62	0.54	0.58	28
accuracy			0.67	212
macro avg	0.63	0.66	0.64	212
weighted avg	0.65	0.67	0.65	212

VII. Conclusion

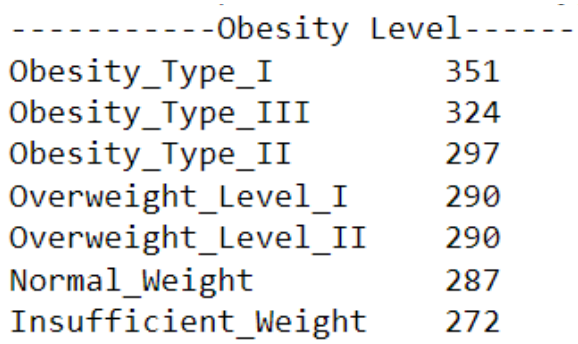
Estimation of obesity levels based on eating habits and physical condition dataset was analyzed in this project. By removing the age, height and weight values in the data, it was predict the obesity levels of the people with the health data. Within the scope of this project, outlier data was not cleaned due to the small number of data.

Label Encoding and One Hot Encoding were applied to the data. Obesity Levels were estimated by K-Nearest Neighbors, SVC, Decision Tree and Random Forest algorithms.

According to the accuracy values, they can be sorted from the highest accuracy score value to the lowest accuracy score value as follows:

1. Decision Tree
2. Random Forest
3. SVC
4. K-Nearest Neighbors

There may be several reasons for the low accuracy values. First of all, the low number of data can be given as one of the reasons for the low accuracy values. There are only 2111 records in the dataset. And there are seven different obesity levels. When the records that are already few in number are also divided into 7 different levels, even fewer records remain for each level prediction. This may be a reason for the low accuracy values.



Another reason for the low accuracy values may be outlier data. Not removing outlier data may reduce accuracy.

Another reason for the low accuracy values is that the required values for the Grid Search CV algorithm could not be given. Better parameters may have been overlooked.

References

[1] https://archive-beta.ics.uci.edu/ml/datasets/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

[2] Palechor, F. M., & Manotas, A. de la H. (2019). *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 25, 104344.*

[3] https://www.w3schools.com/python/matplotlib_pie_charts.asp

[4] https://seaborn.pydata.org/generated/seaborn.countplot.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[7] https://scikit-learn.org/stable/modules/tree.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[9] Ay, Ş.(2020). Model Performansını Değerlendirmek — Metrikler https://medium.com/deep-learning-turkiye/model-performans%C4%B1n%C4%B1-de%C4%9Ferlendirmek-metrikler-cb6568705b1