Semester Project

Ali Aqeel Zafar[0000000162380879]

[1] Eotvos Lorand University,Budapest, Egyetem tér 1-3, 1053, Hungary
[2] aliaqeelzafar34@gmail.com

Abstract. This report contains an overall understanding of how to analyze the Obesity data set by using different methods like Clustering, Regression, and Classification. In addition, presenting some ideas on how can frequent pattern mining can be utilized to uncover some patterns in the data and to enhance the prediction.

Keywords: Semester Project · Data Science · Clustering · Regression · Classification.

1 Data Preprocessing

1.1 Processing the Data Containing Null Values and Assigning Labels

[3] It had been seen that the data set contained null values in Height, Weight and Age features. In order to remove, the missing values they were replaced by the whole feature's mean with the help of Simple Imputer library.Then by using label encoder library I labeled the Family history overweight,Fav,Smoke,Scc and Nobesity. I also labeled Calc,Caec,Mtrans and Gender with label encoder . The look of the data set is as follows:

Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	...	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObesity
0	21.000000	1.620000	64.000000	1	0	2.0	3.0	...	0	2.000000	0	0.000000	1.000000	3	3	1
0	21.000000	1.520000	56.000000	1	0	3.0	3.0	...	1	3.000000	1	3.000000	0.000000	2	3	1
1	23.000000	1.800000	77.000000	1	0	2.0	3.0	...	0	2.000000	0	2.000000	1.000000	1	3	1
1	27.000000	1.800000	87.000000	0	0	3.0	3.0	...	0	2.000000	0	2.000000	0.000000	1	4	5
1	22.000000	1.780000	89.800000	0	0	2.0	1.0	...	0	2.000000	0	0.000000	0.000000	2	3	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	20.976842	1.710730	131.408528	1	1	3.0	3.0	...	0	1.728139	0	1.676269	0.906247	2	3	4
0	21.982942	1.748584	133.742943	1	1	3.0	3.0	...	0	2.005130	0	1.341390	0.599270	2	3	4
0	22.524036	1.752206	133.689352	1	1	3.0	3.0	...	0	2.054193	0	1.414209	0.646288	2	3	4
0	24.361936	1.739450	133.346641	1	1	3.0	3.0	...	0	2.852339	0	1.139107	0.586035	2	3	4
0	23.664709	1.738836	133.472641	1	1	3.0	3.0	...	0	2.863513	0	1.026452	0.714137	2	3	4

Fig. 1. Look of the dataset after data pre processing.

2 Clustering

2.1 K-Means

[1][2] The K-MEANS clustering was done with features Height and Weight. I chose these because to see how Weight is distributed according to Height. It shows that most of the points are closer to cluster center. The clusters are well defined. It shows that how the people according to their height are distributed

with their weight. In order to find the cluster centers it was done through the elbow method in order to find out the optimum number of cluster centers. The reason I used the elbow method because it is hard to find the number of cluster points which is hyper parameter by hit and trial method. Also if we run with out the elbow method K-means clustering it will take time by increasing the number of clusters one by one. It is also computationally efficient in terms as it quickly defines the desired clusters. Also defines clusters in a well defined way meaning showing how they look.

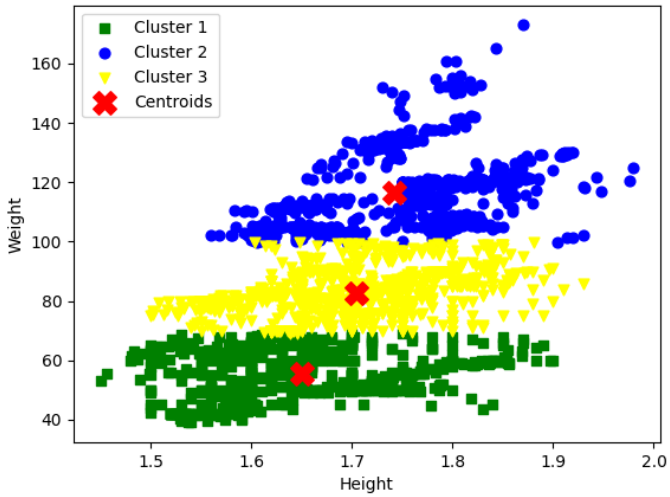

Fig. 2. Clusters with n = 3 between Height and Weight feature of the from data set.

According to the above figure people in cluster 1 have between 40kg and 70kg with height range from 1.5m to 1.9m this means that these are normal weight people. Then according to cluster 2 have height of 1.5m to 1.7m with weight range from 70kg to 100kg this means that these are medium weight people. Then in the third cluster people have height 1.55m to 1.9m and weight 100kg to 160kg, this shows that they are obese people.

2.2 DB Scan

[4][2] Then DB-Scan clustering was done between the Weight and Ncp,Weight and Caec,Weight and Calc.Also the clusters are well defined. The below image is of Weight and Caec and it shows that most of the people consumes food 2 times between meals per day have weight between 40kg to 155kg.Followed by people consuming food 1 time between meals per day have weight between 40kg

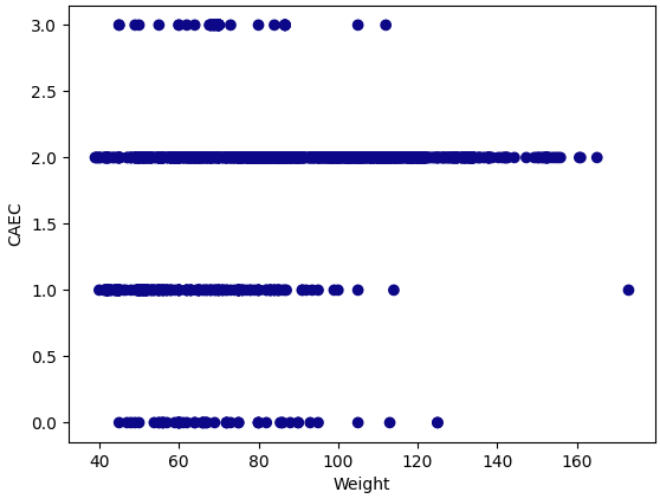

Fig. 3. Cluster generated between Weight and CAEC feature of the from data set.

to 120kg. Then people consuming food 0 times between meals per day have weight between 50kg and 125kg. Then people consuming food 3 times between meals per day have weight between 40kg and 120kg and these people are less in quantity.

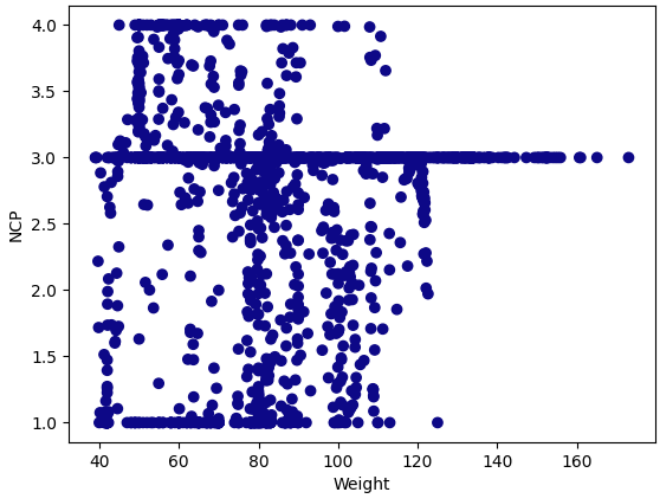

Fig. 4. Cluster generated between Weight and NCP feature of the from data set.

The above figure shows that overall people range from weight from 40kg to 160kg with taking meals per day 1 to taking 4 meals per day. The most densely populated area are of those people that have weight between 40kg to 160kgs that consumes meals per day from 1 to 3 meals per day.

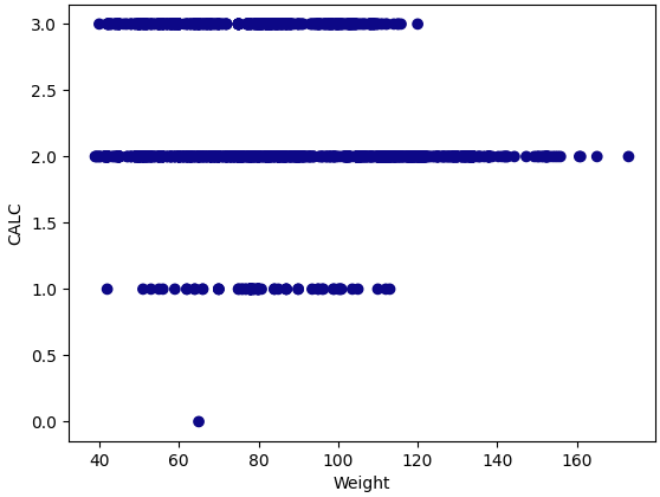

Fig. 5. Cluster generated between Weight and CALC feature of the from data set.

The above image shows that most of the poeple are part of that area who consumes alcohol 2 times of on daily basis and have weight ranging from 40kg to 130 kgs. Followed by those people who consume alcohol 3 times have weight ranging from 40kg to 110kg. Then a few people consume alcohol 1 time per day have weight between 40kg to 117kg.Last there is only 1 person who does not consume alcohol and have weight a little bit over 60kg. After increasing the radius (eps) to 25 and number of points in neighbourhood to 100 the outliers were more visible from the cluster during DBSCAN as can be shown in above two images. The radius and number of point in neighbour hood are the hyper parameters of the clustering algorithm. The reason I used DBScan because it identifies the points that are outliers robustly and gives their visibility in a more dominant way. Then the next reason is that it gives a visibility of the clusters based on how densely are the points in the clusters are populated based on the radius of defining the which tells neighbour hood and the number inside the neighbourhood which helps the upcoming points position to be established and in which cluster they belongs.

2.3 Agglomerate Hierarchical Clustering

[4][2] Then Agglomerate Hierarchical Clustering was done by Weight and Faf. The reason I consider Weight and Faf is to show how many people according to their Weight do physical activities.

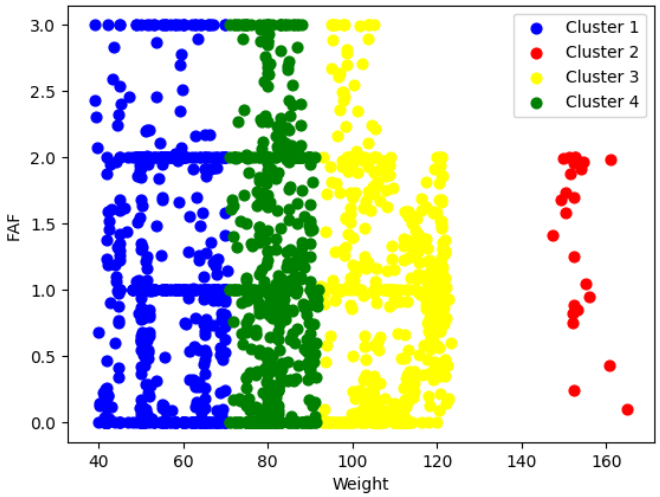

Fig. 6. Cluster generated between Weight and FAF feature of the from data set.

It showed in the above image that in cluster 1 people that weigh between 40kg to 75kg and do physically activity between 0 to 3 times. Then people in cluster 4 weigh between 75kg to 95 kg who physical activity between 0 to 3 times. Followed by cluster 3 which shows people weighing between 95 to 125 do physical activity between 0 to 2 times a day. Lastly cluster 2 has least number of people who are overweight and do physical activity 0 to 3 times a day. The reason I used this type of clustering is to understand that what is the similarity between FAF and Weight meaning how they are related to each other. The next thing was the hyper parameters of this type of clustering is not too hard to measure which is distance the between the clusters.

3 Regression

3.1 Understanding the Co-relation between the Features

[5] The first thing that was to find out the correlation between the features that had positive relationship with each other these were Age,Height,Gender,Weight, Family history overweight, Caec and Faf. The feature that were positively co

related with Nobesity were Age, Height, Weight and Overweight. The positive relationship means that the features are directly proportional with each other. Below we can see the relationship between features through heat map.

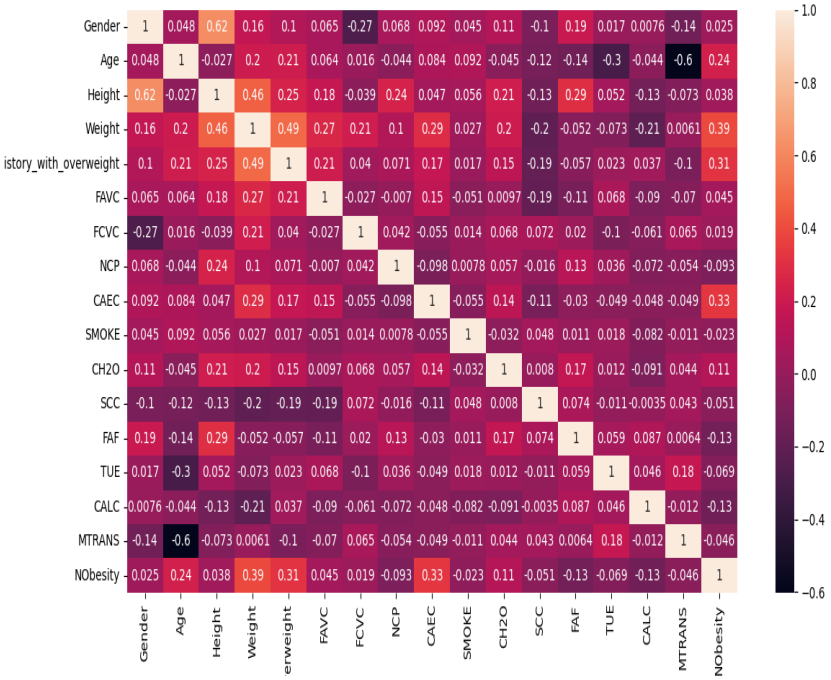

Fig. 7. Heat Map telling about positive and negative correlation between the features

3.2 Execution of Multiple Linear Regression Algorithm

[5][6] I divided my dataset into seventy five percent of train set and twenty five percent of test set. After running first linear regression algorithm on the features that were Age,Height,Gender,Weight,Family history overweight, Caec and Faf with the final predicted value to be Nobesity, the root mean squared error and mean squared error values are as follow: Root Mean Square Error is 1.6803561547748596 and Mean Square Error is 1.427363464522431 It shows that the predicted values(Nobesity values) of the model and actual value(Nobesity value) the difference between them were minimal. Then after this I used another regression algorithm but this time using Age, Height, Weight and Overweight because the above mentioned features had a positive relationship with Nobesity feature and the final predicted value to be Nobesity. After running the second regression algorithm the root mean square error is 1.703936572943416 and the mean absolute is 1.460638798474195.It shows that the RMSE and MAE of first

linear regression model is less than the second linear regression model so it means that Age, Height, Weight , Gender,Family history with overweight, Caec and Faf.This means that if a person has high weight, has family history of overweight, high consumption of food between meals,more age and did little or none physical activity will be more likely to be overweight. The reason I used multiple linear regression is because to find out how relationship between the dependent variable (Nobesity) and the independent variable Age,Height,Gender,Weight,Family history overweight, Caec and Faf is and this can be shown in the above description that is showing calculation of root mean square error and mean absolute error.

4 Classification

4.1 Finding best hyper parameters for KNN Algorithm

[7] I divided my dataset into seventy five percent of train set and twenty five percent of test set. The classification algorithm that was used was K Nearest Neighbour Classifier so in order to use it, its hyper parameters needs to be find that are best to on overall data set. So by using Grid Search CV it is showed that the best hyper parameter values are n neighbours = 1 and p = 1. p = 1 is the distance meaning Minkowsi distance is being used for determining the distance between new features point and class feature point.The hyper parameter were selected based on these features: (Age , Height, Gender,Weight,Family history with overweight,Caec and Faf) that had a positive relationship with each other through heat map. The reason I used KNN algorithm is to find out when new data entry of the person comes based on the given parameters then to which obesity level will it be classified then.

4.2 Executing the Nearest Neighbour Classifier

Classification Report

	precision	recall	f1-score	support
0	0.85	0.92	0.88	62
1	0.82	0.65	0.72	71
2	0.87	0.89	0.88	91
3	0.98	0.98	0.98	87
4	0.99	1.00	0.99	68
5	0.79	0.88	0.83	72
6	0.83	0.82	0.82	77
accuracy			0.88	528
macro avg	0.87	0.88	0.87	528
weighted avg	0.88	0.88	0.87	528

Fig. 8. Classification report of the classification model according to data inputs

[6][7] According to classification report as shown above, after executing the classification model,the model was able to achieve the 0.88 accuracy which is a good accuracy pointing out that the model classified correctly the people's obesity level. Also it gave good precision,recall,f1-score and support which were also good as well. The reason I used classification algorithm as to identify how people fall under the Obesity levels given in Nobesity feature based on the on the features like Age, Gender, Weight, Family history with overweight, Caec and Faf.

5 Frequent Pattern Mining

5.1 Ideas to do Frequent Patterning Mining On Obesity Data Set

[9][10]First thing that can be done is to first group the Nobesity feature with Age, Weight,Gender,Family history overweight,Faf,Ncp,Calc,Caec and Ch20. Then we can apply Apriori algorithm in order to find out which frequent itemsets are made. The generation of frequent item sets will be based on the minimum

support that is the threshold. Another reason for using Apriori algorithm is that association rule mining requires frequent item sets in order to make association rules. Next we can apply association rule mining to further clarify so that it can be shown that through which features can a person be considered obese. This clarification can be provided by applying metric which is lift. Lift is basically it tells how correlated is the antecedent and consequent in the association rule.If the lift is high it means that the features which are related to eating habits have strong correlation with Nobesity which is the corresponds to obesity level and the features related to eating habits that have week correlation with the Nobesity containing obesity levels will have a low lift. The decision which association rule is useful or not is based on if that association rule's lift is greater than 1 then it has strong co-relation, if it is less than 1 then it has weak co-relation and if it is near to 1 then there is no co-relation. The reason is that we can see what patterns are made based on the eating habits. Then next thing that can be done for further better analysis is to group the Nobesity feature with Age,Gender,Height,Weight and related fields based on physical condition like Scc and Faf. Identify the frequent itemsets again through Apriori algorithm and then find association rules that have strong corelation. The reason is that we can see what patterns are made based on the physical conditions.Lastly after getting the association rules from eating habits and from physical conditions we can identify those association rules which can be considered by combining both the eating habits features and the physical condition features to find out which association rules have strong correlation. Through frequent pattern mining we can also find new co relations between the features in the data and a result we can increasing the precision and accuracy in clustering, classification and regression.[8]

References

1. Dr. Horváth Tamás: ITDS Lecture 2 , slide 6–11 (2021)
2. Professor Kamuzora Adolf Rutebeka:Clustering Practical 6, http://localhost:8888/notebooks/Downloads/Pr 6//20Clustering.ipynb 2021
3. Professor Kamuzora Adolf Rutebeka: Pr 5 Data Processing , page 1–31 (2021)
4. Dr. Horváth Tamás: ITDS Lecture 2 , slide 20 (2021)
5. Professor Kamuzora Adolf Rutebeka:Regression Practical 7, http://localhost:8888/notebooks/Downloads/Pr_7/20$Regression.ipynb$2021
6. Dr. Horváth Tamás: ITDS Lecture 4 , pages 1–41 (2021)
7. Professor Kamuzora Adolf Rutebeka:Classification Practical 8 Part 1, http://localhost:8888/notebooks/Downloads/Pr_8/$ClassificationPart1.ipynb$2021
8. Thashmee Karunaratne:Is Frequent Pattern Mining useful in building predictive models? , pages 1–12
9. Dr. Horváth Tamás:ITDS Lecture 03.pdf , slide 1–34(2021)
10. Professor Kamuzora Adolf Rutebeka:FPM Practical 11, http://localhost:8888/notebooks/Downloads/Pr 11 20FPM.ipynb 2021