Examining Factors Influencing Obesity

1st Ömer Aras Kaplan
Computer Engineering (CE)
Yıldız Technical University (YTU)
İstanbul, Turkey
aras.kaplan@std.yildiz.edu.tr

2nd Ömer Buğrahan Çalışkan
Computer Engineering (CE)
Yıldız Technical University (YTU)
İstanbul, Turkey
bugrahan.caliskan@std.yildiz.edu.tr

***Abstract*—For over two decades, data science and data analytics had greater successes in their fields. Main impulse over this phenomenon is generally perceived as computer aided computing. With such computing speeds, more and more data can be processed into knowledge and information. Composed of 2117 entries and 17 columns of information, the data used for this project is an estimation of obesity levels from people of Northern Latin American countries. This paper contains information about how everyday choices people make impact their BMI, obesity and weight status. This paper also includes steps to determine effects of each element, steps to clear and rationalize data and steps to visualize data in a graphical fashion. The paper also contains classification and clustering results after their algorithmic steps. Each section contains said visual components to aid reading and understanding. The conclusion part presents information about different algorithms and their result involving project's model and dataset. Best resulting classification methods for this project were Random Forest and Decision Tree Classifiers. Best Resulting Clustering method was K-means. For better understanding, all methods in this document have a table regarding extra operations, such as gradual accuracy, Important Feature Accuracy and PCA. The program used to obtain the information in this article can be found here.**

***Index Terms*—obesity, BMI, classification, clustering, preprocessing, data analytics, data visualization, KNN, naive-bayes, random forest, SVM, decision tree, k-means, hierarchical clustering, agglomerative clustering, PCA, important feature accuracy**

I. Introduction

Obesity is generally defined as weighing above certain levels of individuals' ideal weight. [1] Altough it causes the lowest rate among individuals, obesity is an underlining cause of diseases like hypertension, diabetes, coronary diseases, certain types of cancer. [2] Obesity levels in the world are rapidly increasing [3]. While obesity is gaining more and more attention around the world, more data are collected and more are shared with researchers and society.

This document contains information about how lifestyle choices affect individuals' obesity level. Keeping in mind that obesity is determined by solely on height and weight. which are not lifestyle choices but merely results of an array of choices. In order to focus on lifestyle choices, some criteria have to be kept aside.

The project's dataset in use is generated as a hybrid. A quarter of entries are organic and collected from a web platform. Rest is gathered and generated by Weka tool [4]. The dataset has 2111 entries and 17 columns of information. These columns are mainly categorical except for height, weight and age.

The program used to obtain the information in this article can be found **here**.

A. Literature Review

In order to get brighter insights on the matter at hand, we examined previous involving the same dataset. To review said articles, we searched the site kaggle, where the dataset is found. While reviewing, classification and clustering methods, accuracy values, preprocessing steps were taken a look at.

1) ObesityDataSet: EDA, Data Prep, ML and HyperTuning: The first article removed height and weight from columns because height and weight are directly linked to bmi which is the default way to calculate labels. That way it deemed other features unnecessary for analysis. Upon data examination section, certain graphs were used to visualize data. The classifying methods used in the articles, Knn, decision tree, svm, are fairly common algorithms. So they were added to the project.

Having finished our project, after a second review on the article, we found out that accuracy values were higher. We conceived that it might be due to the fact that in our project, the BMI column in our project was reconfigured into a different label array. Additionally, while converting categorical values into numerical values, instead of adding new columns, logic values like 0, 1 were used, except for MTRANS column. In the article it was found that the most important feature was "Age", since "Age" is removed in our project for focus on life choices, the most important feature was different. [5]

2) Exploring and modelling Obesity Dataset: distributions kmeans boxplot visualization The second article was more focused on data visualization and clustering fields. Bivariate analysis, distribution charts, histograms were used. Also a correlation matrix was used to find influencing factors. Height and Weight features was still in the dataset used for training and testing.

The second article involved using Decision Tree Classifier and Receiver Operating Characteristic Curve for classifications. For clustering purposes, kmeans was used in the article.

As a result of not removing height and weight, the models found out that height and weight was the most important factors on labeling for obesity. While this is not wrong by any means, it does not reveal any information, system or insight

about other data in the set. There is no need for a machine learning model in a situation where the analyst has height, weight and the equation chart for obesity. [6]

B. Abbreviations

Abbreviations for column names are as follows:

- Frequent consumption of high caloric food(FAVC)
- Frequency of consumption of vegetables (FCVC)
- Number of main meals (NCP)
- Consumption of food between meals (CAEC)
- Consumption of water daily (CH20)
- Consumption of alcohol (CALC)
- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)
- Gender
- Age
- Height
- Weight

Age, height and weight columns are ordinal float values. Rest is categorical values.

C. Obesity Levels

For labeling purposes, formula for obesity and BMI levels of obesity are as follows:

$BMI = weight(kg)/height(m)^2$

- Insufficient Weight

$$BMI < 18.5$$

- Normal Weight

$$BMI < 25$$

- Overweight

$$BMI < 30$$

- Obesity Type I

$$BMI < 35$$

- Obesity Type II

$$BMI < 40$$

- Obesity Type III

$$BMI > 40$$

D. Information

Later in this document, steps will be found regarding preprocessing, classification and clustering. At the last section there will be more information about comparison of model results.

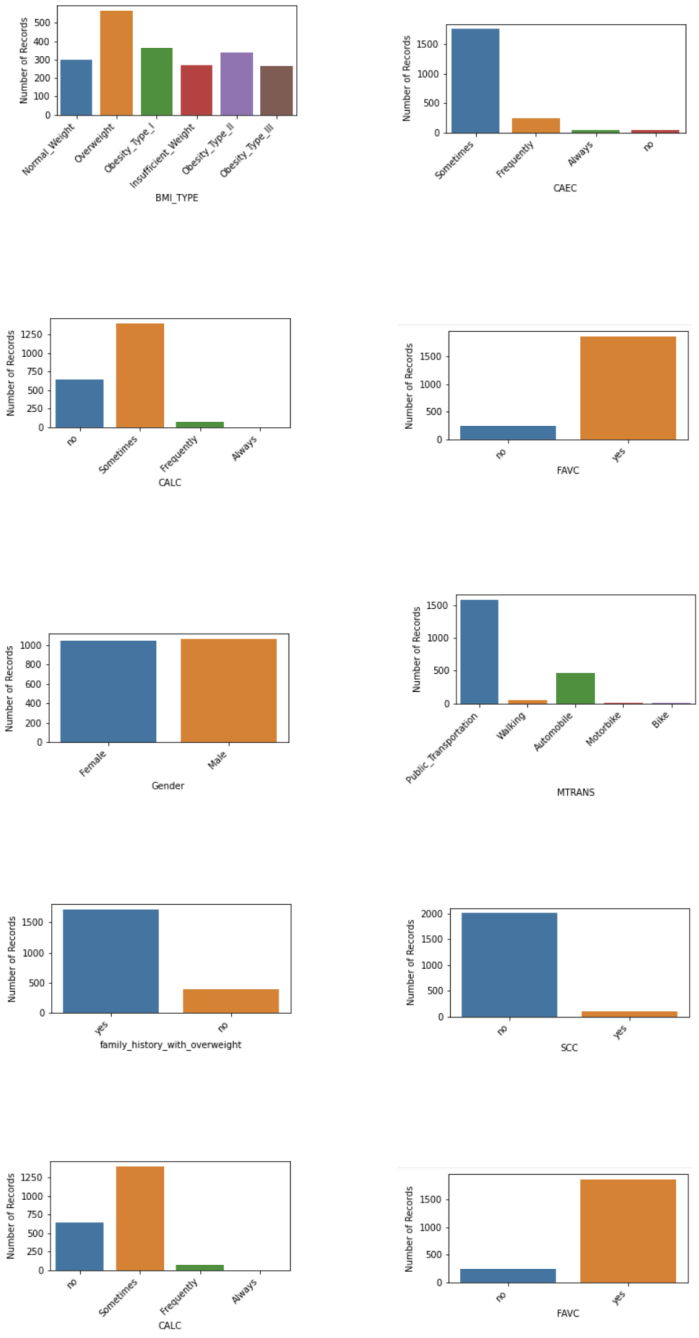
The gradual result sections on the models' result tables show the model's accuracy with 1-label radius of tolerance.

The project is developed on Jupyter Notebook. Pandas, Matplotlib, Seaborn, Numpy libraries were used in the process of analyzing and visualizing data.

II. Data Examination

A. Histogram Examination

To gain insight about the data at hand, variables are shown over a histogram plot.



Age, weight and height are the most affecting features on the model. Seeing the distribution of these features may present useful information.

Feature distributions are shown over histogram.

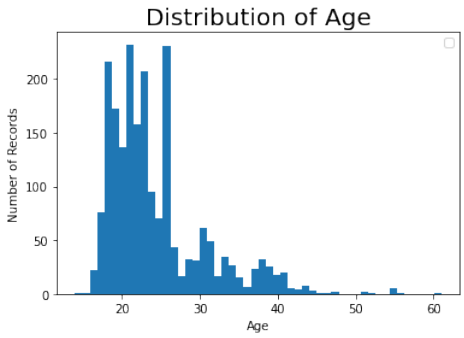


Fig. 1. Age Distribution

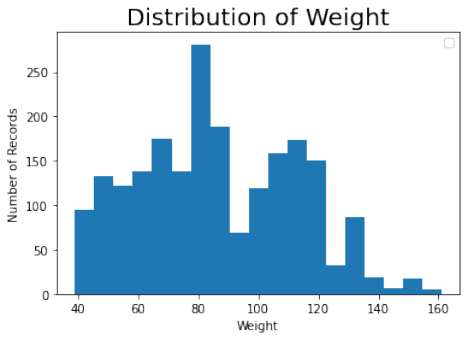


Fig. 2. Weight Distribution

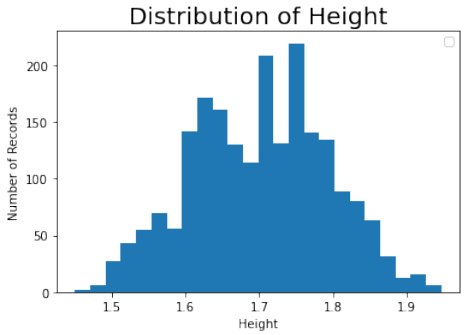


Fig. 3. Height Distribution

B. Bivariate Analysis

To show different feature distributions together while presenting mean, IQR and outliers; a boxplot with features spreading over BMI is generated.

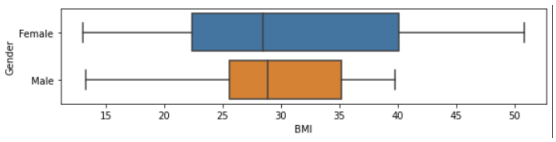


Fig. 4. Gender

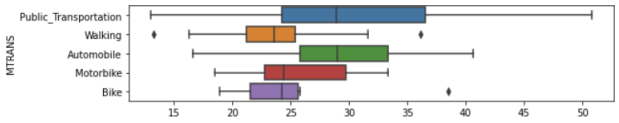


Fig. 5. MTRANS

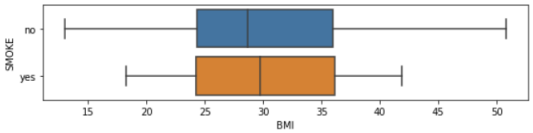


Fig. 6. Smoke

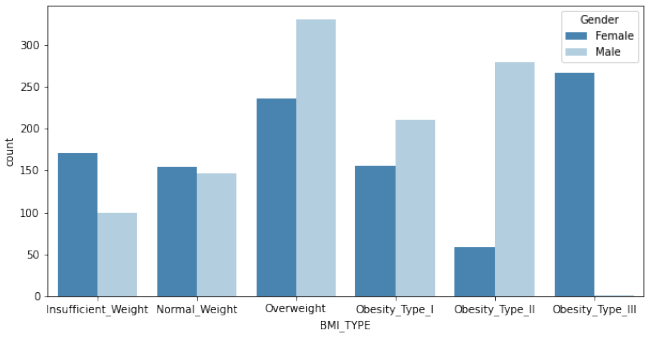


Fig. 7. BMI by Gender

C. Correlation Matrix

To gain the information about the most impactful features in dataset, a correlation matrix is in order.

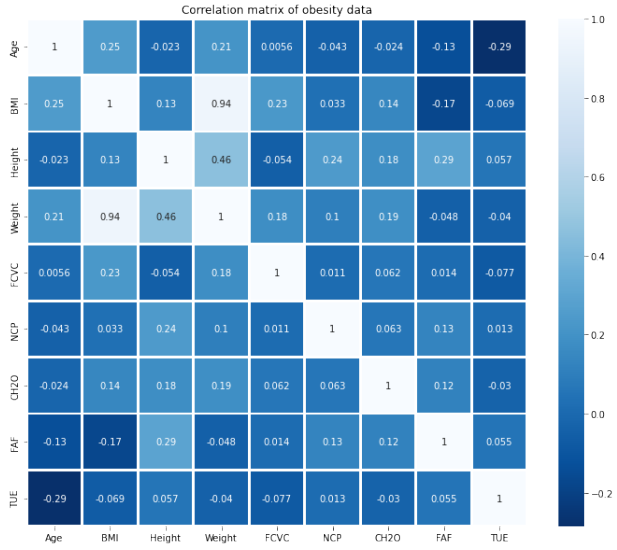


Fig. 8. Correlation Matrix

While checking the matrix, it can be seen that height, weight and BMI features have the most influence over the obesity levels.

III. Preprocessing

A. Removing outliers

Since it is very important to have logical and consistent data, outliers and entries with empty columns have to be removed. This part is also important for preventing the model from overfitting.

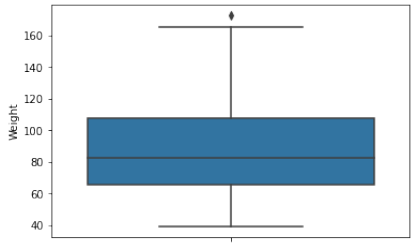


Fig. 9. Before Removing Outliers

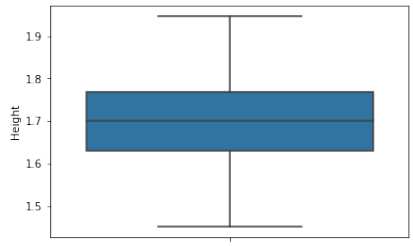


Fig. 10. After Removing Outliers

B. Feature Extraction

To deduce someone as overweight, obese or morbidly obese one has to check their BMI. Body Mass Index(BMI) is gathered by, dividing weight(in kilograms) by the square of height(in meters). The result is crosschecked by the table of intervals.

Although the column for obesity level exists in the original dataset, the categorization was different than the standard. So the old column is dropped, then new column is added by using mapping functions for given intervals.

C. Data Preparation

Due to the fact that this project focuses on lifestyle choices on obesity; height, age and gender should not be taken into consideration. As these fields can not be chosen or result of ongoing choices up until now.

While weight can be considered as a choice; it is more appropriate if it is thought as a result of an array of choices. Given the fact that weight feature is used first hand to compute the label, the feature itself should not be used while predicting labels, contrary to testing.

To comply with the project's needs, previously mentioned features are removed from dataset.

After dropping certain columns, ordinal values are converted to floats and categorical columns are mapped into different values according to their initial value.

IV. Feature Importance

Feature importance techniques are used in order to calculate a score for all features' importance. The higher the score, the more influence it has on the outcome of label.

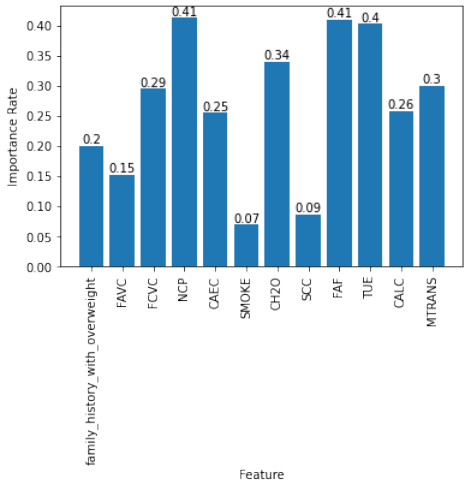


Fig. 11. Feature Importance Matrix

As it can be seen on the plot, there are many influential features on the dataset. Most important features are number of main meals(NCP), physical activity frequency(FAF) and time using technology devices(TUE).

V. Principal Component Analysis

To test results in a lower dimension a Principle Component Analysis was performed. Upon looking at principal component array, for the highest coverage with least amount of component vectors, 5 vectors could cover 80 percent of the dataset.

In the classification section, all PCA approaches were made with 5 PCA vectors.

VI. Find GINI

Gini coefficient represents the rate of labels after a decision node. A perfect model, meaning only one label after a decision, means a gini coefficient of 1. A random model means a gini coefficient of 0. To summarize, the higher gini coefficient is, the better.

Gini coefficient is a very useful way for node determination and knowledge discovery in data mining field [7].

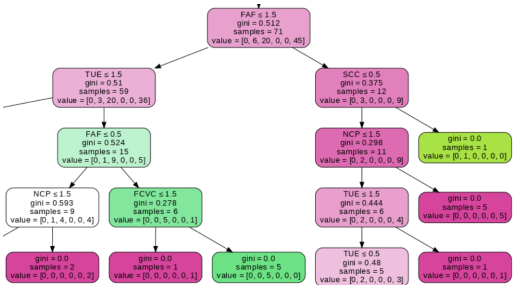


Fig. 12. Pruned Version of GINI Tree

Gini tree gives information about the path of queries involving program's way of analyzing objects for labeling. Gini tree created by the program is too big to include on this document.

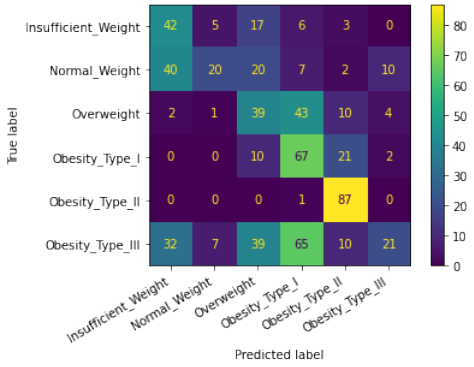
First levels of nodes are composed by queries about mostly FAF, TUE and NCP. Since the rest of the features influence the outcome less, it is harder to get a conclusive comparison. Therefore, there are more queries about the rest of the features in the deeper levels.

VII. Classification

A. Naive Bayes Classifier

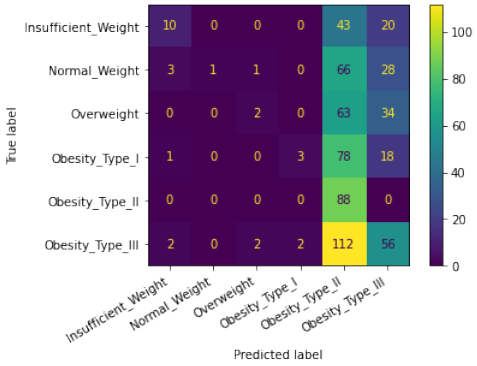
Naive Bayes classification methods are derived from Bayes' Theorem from probability theory. Put simply, Naive Bayes classifies objects using probability functions depending on their feature sets [8]. Naive Bayes approach is highly scalable. Meaning that it needs a certain amount of test values with independent features.

1) Raw testing: Raw Naive Bayes Test Matrix:

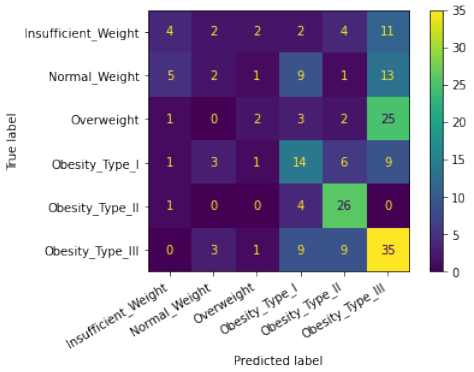


At the end of training and testing the model's accuracy was somewhat lower than expected. This might be due to dataset having many variables with similar importance rates and correlations.

2) Important Feature Based Testing: Naive Bayes with Important Feature Based Test Matrix



3) Principle Component Analysis Testing: Naive Bayes Test with Principle Component Analysis Matrix



4) Comparison:

Accuracy Metric	Accuracy
Raw Accuracy	0.36
Gradual Accuracy	0.49
Important Features Based	0.25
Principle Component Analysis	0.39

Apart from gradual accuracy metrics, which is not an accepted metric, training with PCA gave a slightly better result in comparison.

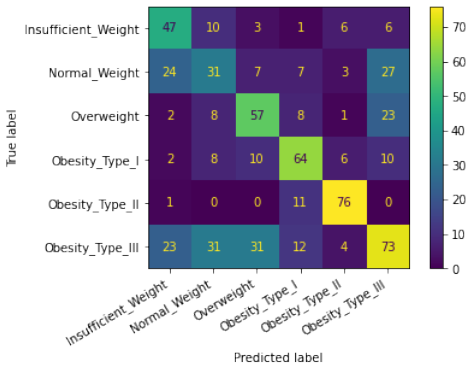
B. KNN - K Nearest Neighbor

K nearest neighbor algorithm uses different distance metrics to find the "k" nearest neighbors. The distance metric used in the project is euclidean distance. [9] [10]

$$\|a-b\| = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

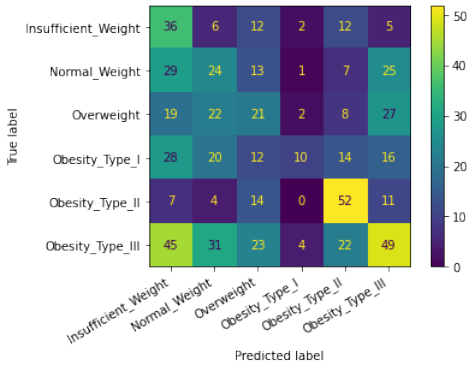
After finding the closest neighbors, model assigns the label of the most frequent label among neighbors [11].

1) Raw testing: Raw KNN Test Matrix:

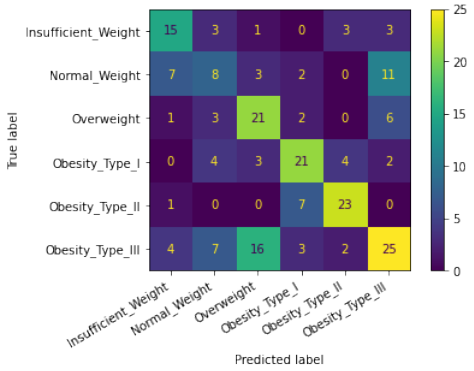


After the training, it is clear that KNN gave better results than Naive-Bayes. Unfortunately, the results are not as good as it should be.

2) Important Feature Based Testing: KNN with Important Feature Based Test Matrix



3) Principle Component Analysis Testing: KNN Test with Principle Component Analysis Matrix



4) Comparison:

Accuracy Metric	Accuracy
Raw Accuracy	0.56
Gradual Accuracy	0.71
Important Features Based	0.30
Principle Component Analysis	0.53

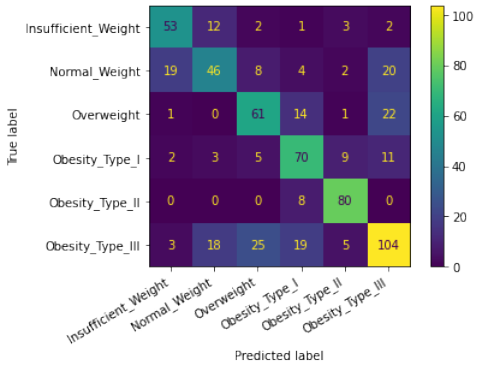
Important features based training reduced the accuracy drastically and contrary to expectations, principle component analysis reduced the accuracy only slightly.

It is common for PCA to reduce accuracy. Afterall, dimensions are decreased. Nevertheless, only 0.03 percent drop is very acceptable.

C. Random Forest Classification

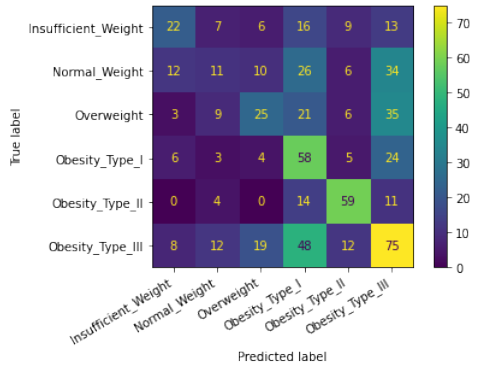
Random forest classifiers are made up of a large number of desicion trees. Each tree in the forest generates a class prediction. After prediction phase, the most frequent prediction is assigned as label. Random forest trees are more robust against overfitting but this comes as a cost to their accuracy.

1) Raw testing: Raw Random Forest Classifier Test Matrix

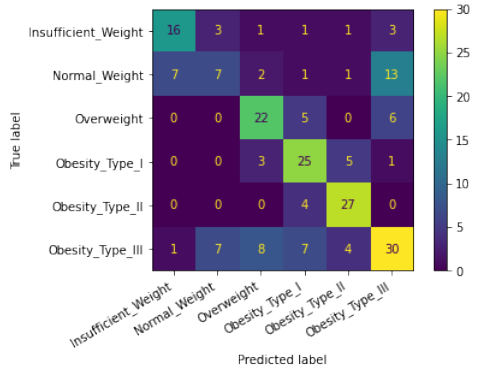


Random forest classification gave the best test results for the training. A probable reason would be the fact that most of the features where categorical.

2) Important Feature Based Testing: Random Forest Classifier with Important Feature Based Test Matrix



3) Principle Component Analysis Testing: Random Forest Classifier Test with Principle Component Analysis Matrix



4) Comparison:

Accuracy Metric	Accuracy
Raw Accuracy	0.66
Gradual Accuracy	0.78
Important Features Based	0.39
Principle Component Analysis	0.60

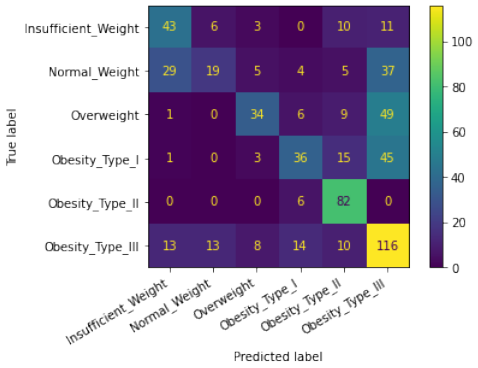
Although random forest classification was more successful, from the metric comparison perspective; it was not very

different than the previous classifiers. PCA and IFB turned out as expected. PCA slightly lowering accuracy for reduced dimensionality while IFB was drastically lower.

D. Support Vector Machines

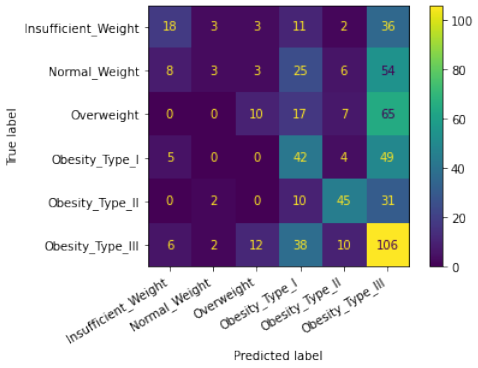
SVMs are supervised learning models that is used for classification and regression analysis. Support vector machine finds a hyperplane in an N-dimensional space that distinctly classifies the data points. [12]

1) Raw testing: Raw SVM Test Matrix

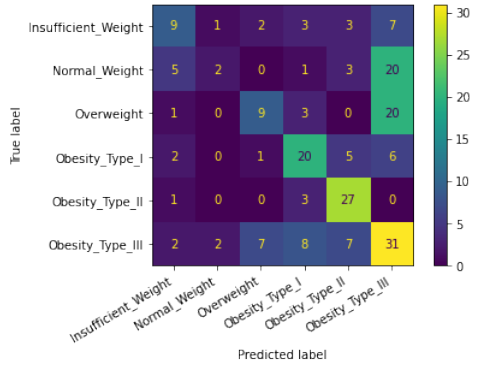


SVM training and testing resulted in a similar conclusion as KNN. The difference being, wrongfully labeled objects was scattered much more than KNN.

2) Important Feature Based Testing: SVM with Important Feature Based Test Matrix



3) Principle Component Analysis Testing: SVM Test with Principle Component Analysis Matrix



4) Comparison:

Accuracy Metric	Accuracy
Raw Accuracy	0.51
Gradual Accuracy	0.65
Important Features Based	0.35
Principle Component Analysis	0.46

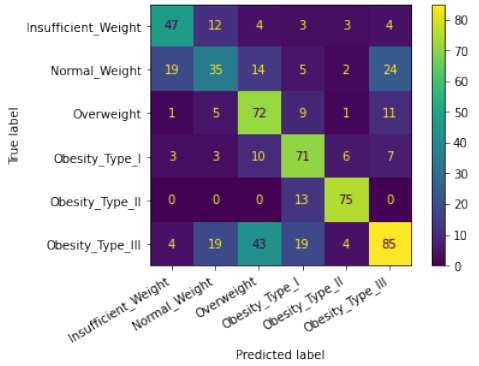
At the end of SVM analysis, it can be seen that SVM could not function as properly as the rest when regarding the edge labels such as ObesityTypeIII. Another point to focus on was, although SVM had worse results, Important Features Based Approach gave better results than KNN's Important Features Based Approach.

E. Decision Tree Classifier

Decision Tree classifiers uses an array of queries to make decisions similar to humans. Every time a decision is made the feature space is segmented into regions. A perfect node should separate one label from others. While going down the tree the amount of probable classes decreases. If the last leaf of the tree is impure by classes, the most common class is assigned to the object. [13]

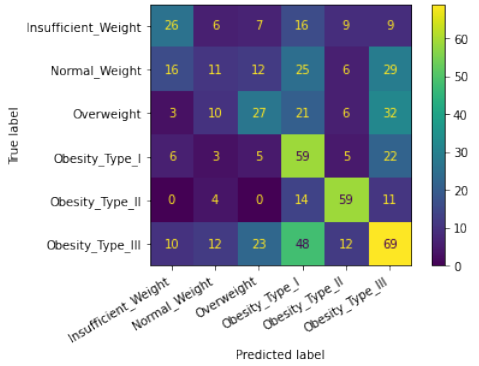
To pick the best splitter nodes, GINI coefficient is used.(See VI.1)

1) Raw testing: Raw Decision Tree Classifier Test Matrix

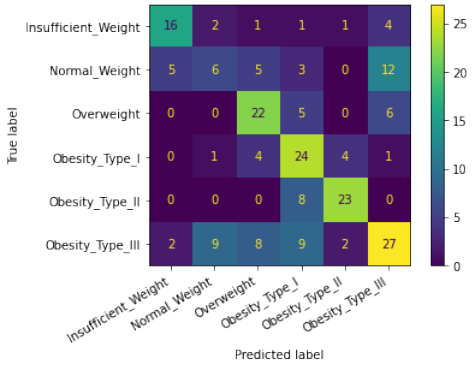


Similar to Random Forest Clasification, Decision Tree Classifier gave better results. May due to categorical values on the features.

2) Important Feature Based Testing: Decision Tree Classifier with Important Feature Based Test Matrix



3) Principle Component Analysis Testing: Decision Tree Classifier Test with Principle Component Analysis Matrix



4) Comparison:

Accuracy Metric	Accuracy
Raw Accuracy	0.60
Gradual Accuracy	0.74
Important Features Based	0.40
Principle Component Analysis	0.56

Very similar to Random Forest Classifier, Decision Tree Classification resulted similar. Backing the claim of tree type classifiers work better with categorical values.

One notable difference between Random Forest and Decision Tree is, in our case, the wrong labels were more scattered.

VIII. Clustering

A. K-Means Clustering

Similar to KNN K-Means algorithm works by computing distances. K-means separates the data space into k clusters. To do so, algorithm suggests clustering classes with closer means together.

1) Based on BMI-Age: Plot of Kmeans clustering based on BMI vs Age

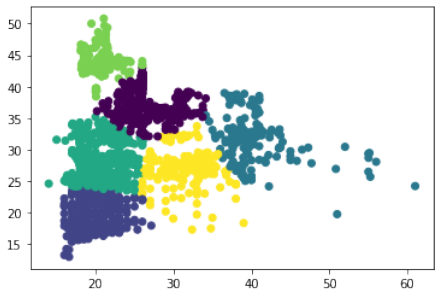


Fig. 13. BMI-Age K-means

2) Bsed on NCP-FAF: Plot of Kmeans clustering based on NCP vs FAF

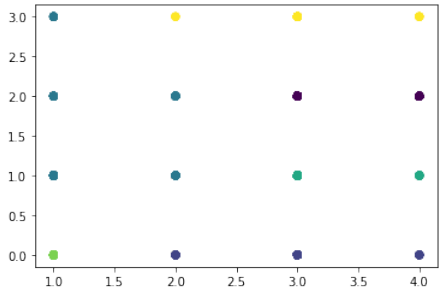


Fig. 14. NCP-FAF K-means

3) Comparison:

Accuracy Metric	Accuracy
BMI-Age	0.81
NCP-FAF	0.69

B. Hierarchical Clustering

Hiearchical clustering starts by treating each observation as a separate cluster. After that it merges the most similar pair of clusters. This iteration is done until all the clusters are merged. After mergers, the hierarchical order is denoted by the order of mergers. First is lower and the last is higher on the chain.

Similarity is measured by distance which is deducted by euclidean distance.

The hierarchical clustering we used in this case is euclidean affinity with complete linkage. Which uses the maximum distances between all observations of the two sets.

1) Based on BMI-Age: Plot of hierarchical clustering based on BMI vs Age

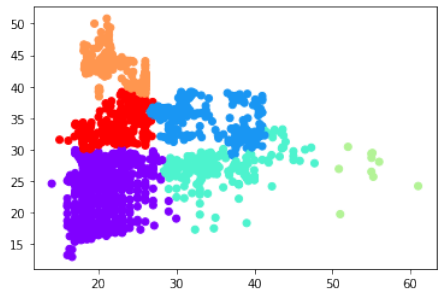


Fig. 15. BMI-Age Hierarchical Clustering

2) Based on NCP-FAF: Plot of hierarchical clustering based on NCP vs FAF

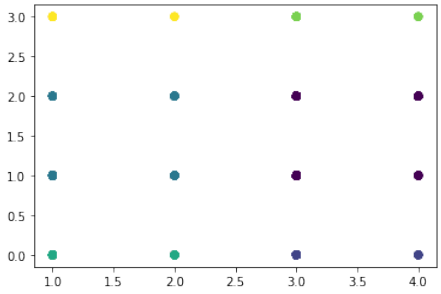


Fig. 16. NCP-FAF Hierarchical Clustering

3) Comparison:

Accuracy Metric	Accuracy
BMI-Age	0.77
NCP-FAF	0.62

C. Agglomerative Clustering

Agglomerative clustering is a subset of Hierarchical clustering. With the difference of linkage. In this part linkage is ward. Which minimizes the variance of the clusters being merged.

1) Based on BMI-Age: Plot of agglomerative clustering based on BMI vs Age

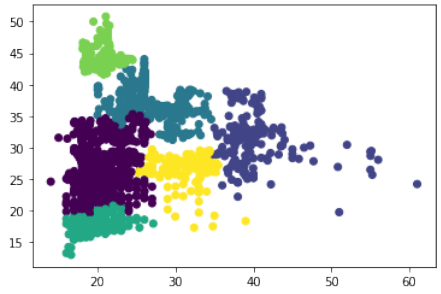


Fig. 17. BMI-Age Agglomerative Clustering

2) Based on NCP-FAF: Plot of agglomerative clustering based on NCP vs FAF

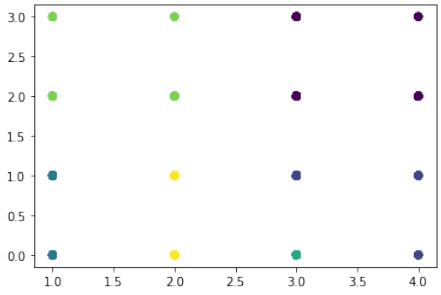


Fig. 18. NCP-FAF Agglomerative Clustering

3) Comparison:

Accuracy Metric	Accuracy
BMI-Age	0.78
NCP-FAF	0.68

IX. Conclusion

Table of Classifier Methods

Method	Normal Accuracy	Gradual Accuracy	Important Feature Accuracy	PCA
Naive Bayes	0.35	0.49	0.25	0.39
K Nearest Neighbor	0.56	0.70	0.30	0.53
Random Forest	0.66	0.78	0.39	0.60
Support Vector Machine	0.50	0.65	0.35	0.46
Decision Tree	0.60	0.73	0.39	0.55

Table of Clustering Methods

Method	BMI-Age Accuracy	NCP-FAF Accuracy
Kmeans	0.81	0.689
Hierarchical	0.77	0.623
Agglomerative	0.78	0.686

The dataset we used in this project had 2117 entries and 17 columns of information. In order to prepare data, outliers and entries with empty columns have been removed. To focus on lifestyle choices on the subject, certain columns were removed. After that, remaining data on the dataset has been plotted to visualize and better understand the distributions. Later on, a feature importance graphic has been implemented to obtain best columns for important feature based approaches. Consequently, a PCA analysis was formed and PCA vector was created for PCA approaches on classifiers. Next, classifiers were trained and tested on the dataset. Discoveries were visualized and put on a table of information. Subsequently, clustering algorithms were trained and tested.

In light of our findings, tree based approaches gave better results while both classifying and clustering.

References

[1] Jiang, S., Lu, W., Zong, X., Ruan, H., Liu, Y."Obesity and hypertension (Review)". Experimental and Therapeutic Medicine 12.4 (2016): 2395-2399.

[2] Tjepkema, Michael. "Adult obesity." Health reports-statistics canada 17.3 (2006): 9.

[3] Booth, M., Hunter, C., Gore, C., Bauman, Adrian., Owen, Neville. (2000). The relationship between body mass index and waist circumference: Implications for estimates of the population prevalence of overweight. International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity. 24. 1058-61. 10.1038/sj.ijo.0801359.

[4] Fabio Mendoza Palechor, Alexis de la Hoz Manotas, Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief, Volume 25,2019, 104344,ISSN 2352-3409, https://doi.org/10.1016/j.dib.2019.104344.

[5] https://www.kaggle.com/code/pmrich/obesitydataset-eda-data-prep-ml-hypertuningFeature-Importance-w/-Random-Forest

[6] https://www.kaggle.com/code/juanfearias/exploring-and-modelling-obesity-dataset

[7] S. Sivagama Sundhari, "A knowledge discovery using decision tree by Gini coefficient," 2011 International Conference on Business, Engineering and Industrial Applications, 2011, pp. 232-235, doi: 10.1109/ICBEIA.2011.5994250.

[8] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." Encyclopedia of machine learning 15 (2010): 713-714.

[9] Wang, L., Zhang, Y., Feng, J. (2005). On the Euclidean distance of images. IEEE transactions on pattern analysis and machine intelligence, 27(8), 1334-1339.

[10] Danielsson, P. E. (1980). Euclidean distance mapping. Computer Graphics and image processing, 14(3), 227-248.

[11] Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.
[12] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
[13] https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575