

**DEPARTMENT: Computer Science**

**Module Code: Machine Learning**

**Academic Year: 2024**

**Submission Deadline: 25/02/2024**

**Actual Submission Date: 25/02/2024**

|  |
| --- |
| **Student Registration Number** |
| ARY23030 |

**Declaration**: *I have completed and submitted this work by myself without assistance from or communication with another person either external or fellow student or any AI type of content generator. I understand that not working on my own will be considered grounds for unfair means and will result in a fail mark for this work and might invoke disciplinary actions. This piece of assessment will be continuously checked for its academic integrity until my graduation and the mark will be revised if it is found to breach the unfair means policy. It is at the instructor’s discretion to conduct an oral examination which will result in the award of the final grade for that particular piece of work.*

**Abstract**

This research studies predicting likelihood of obesity based on several variables by using machine learning algorithms by making useful analysis and visualizations. In the report both supervised and unsupervised learning methods - clustering analysis and decision trees (Random Forest) were applied. The report shows the effects when changing parameters on model performance. Additionally, comparisons between metrics from the training and test sets are used to evaluate the possibility of overfitting. The report also dives deep into how regularization methods work in case of overfitting. For clustering analysis, a lot of techniques such as elbow method, silhouette score was used. In summary, this research gives useful analysis on obesity prediction using machine learning techniques in healthcare and related domains by making decisions.

Python is used for programming with various libraries.

Contents

[**1. Introduction** 4](#_Toc159784347)

[**2. Data Preprocessing** 4](#_Toc159784348)

[**3. Decision tree (Random Forest) metrics** 5](#_Toc159784349)

[**3.1 Checking the Overfitting** 6](#_Toc159784350)

[**3.2 Building the models** 7](#_Toc159784351)

[**3.3 Working on the model** 8](#_Toc159784352)

[**3.4 Working on a leaf from the tree** 9](#_Toc159784353)

[**4. Clustering** 10](#_Toc159784354)

[**4.1 Working on the model** 11](#_Toc159784355)

[**4.2 Model insights** 12](#_Toc159784356)

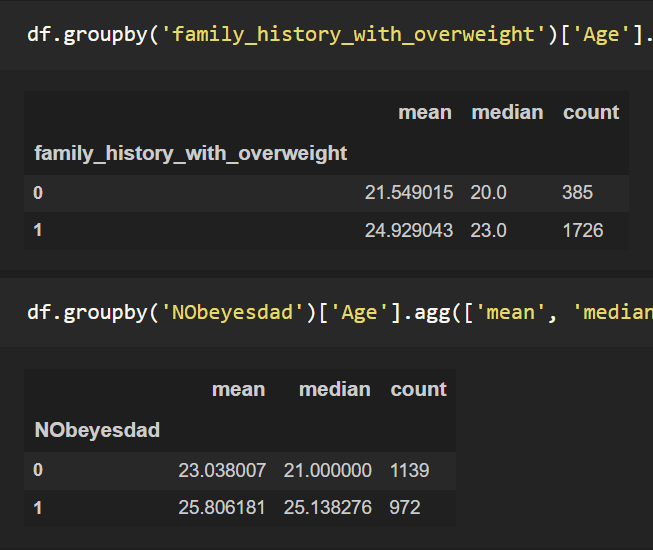
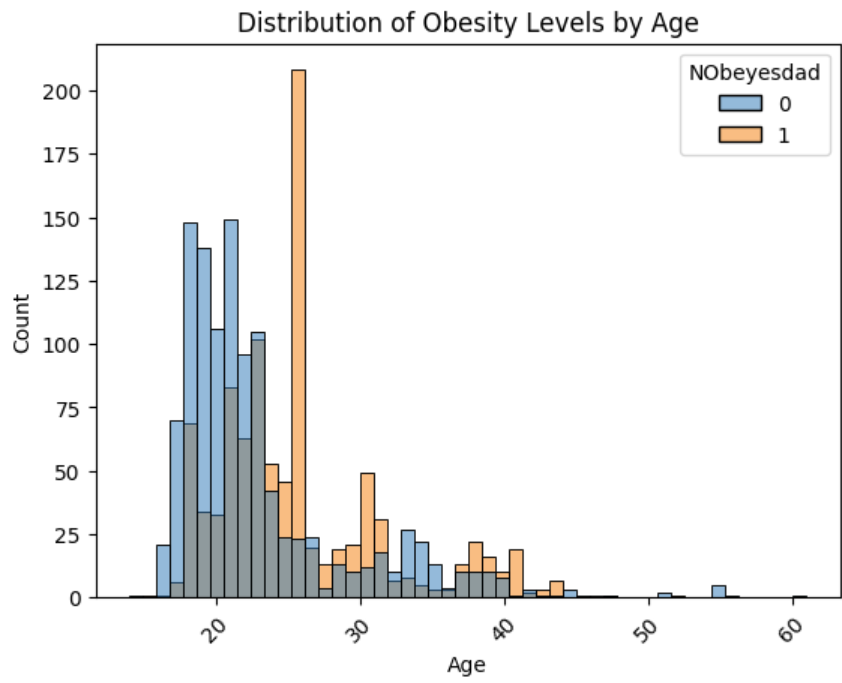
[**5. Conclusion** 13](#_Toc159784357)

# **1. Introduction**

The purpose of the project is to work with machine learning algorithms to determine meaningful information from the dataset considering out y variable. The predictor will show how likely they are to be obese. Both supervised and unsupervised machine learning models are used in the report. Recursively splitting the dataset into subgroups and optimizing each division for a certain criterion, usually impurity reduction or information gain maximization is how the decision tree method operates, and for the second part, clustering analysis will be used to determine the optimal number of K’s and to understand how an unlabeled data works. As it is not the key point to find a fine-tuned model, the data will not be divided into validation set but it might be used.

# **2. Data Preprocessing**

To start with, it is important to check if there are missing values or not in order to decide the best way to deal with them, as well as to see the feature scaling solution if needed. In the dataset provided there are no missing values but there is ‘Age’ variable which differs from all other features by the scale, so it must be normalized to have an improved dataset. It is very important to know the data before moving on. Thus, a lot of visualizations are plotted to understand important patterns and necessary information.



*Fig.1: Obesity levels by age and number of obese people.*

It is clear that most people who have the most obesity are at the age of 26. And there are almost 5 times more people who have family history of overweight. However, those people who have obesity issues are people who have family history of overweight and only 8 people are obese without family history issues shown in figure 2.

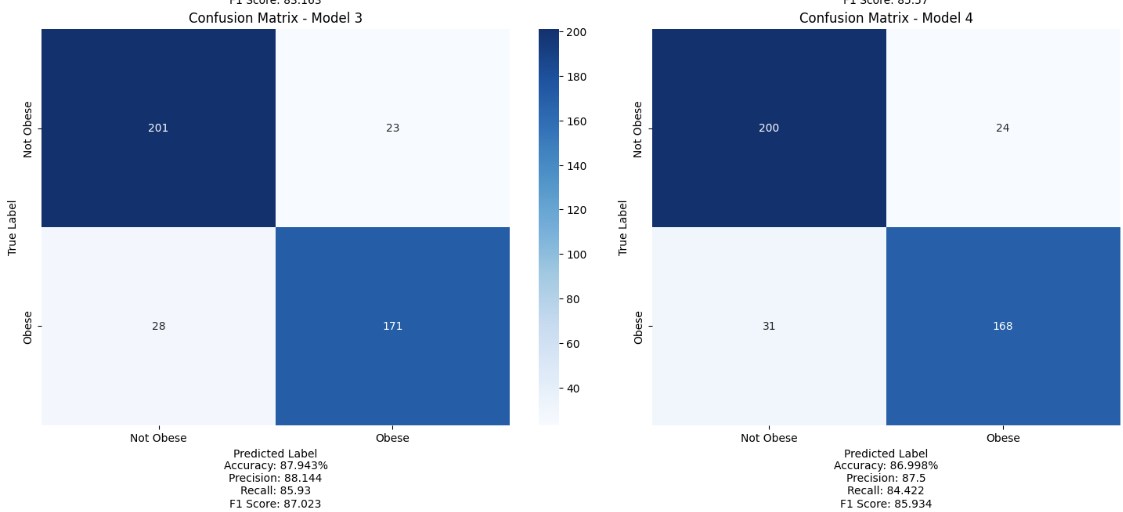


*Fig.2: Number of obese people having family history of overweight.*

Also, it is important to understand the correlations between variables, if there are any strong ones. There is a very strong negative correlation between Public transportation and Automobile having a -0.91 score.

# **3. Decision tree (Random Forest) metrics**

In Decision tree algorithm, different estimators were used to understand the difference in metrics (precision, recall, f1) of the models. The purpose of these matrixes is to see how many obese people were found correctly. The best one was performed with having 130 estimators. Less false negatives and false positives with a better F1 score. Having observed a lot of estimators it is clear that having more than 150 estimators doesn’t change anything meaningful in metric scores. Even with 2000 estimators the model gave almost the same result (less than 1% of difference in the metrics of models). It was also performed in different leaf nodes for all of the models, with 30 leaf nodes model works better than it would work with 16 leaf nodes. Here is the visualization of the confusion matrix having 130 and 2000 estimator [1] [2].

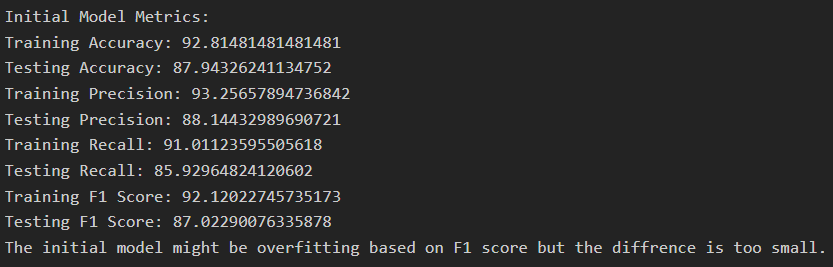


*Fig.3: Confusion matrix of 2 different models.*

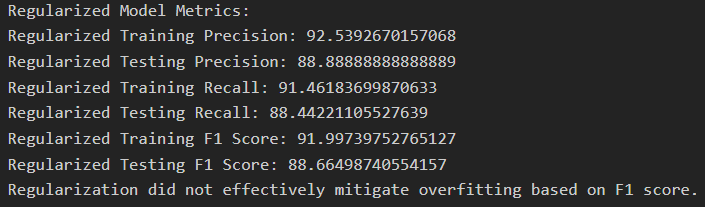
Precision shows how often model is correct when predicting the class, it shows the accuracy of positive predictions, where Recall identifies all positive instances out of all actual positive instances in the dataset. Meanwhile, F1 score is the harmonic mean of these 2 measures. For accuracy, it shows how often the model is predicted correctly, however it is not such a significant measure to use when the data is unbalanced. For this case, precision and recall and both quite high which is either overfitting or very good model. Most cases have different score of precision and recall as one of them goes up the other goes down. It will be shown later in this report [3].

## **3.1 Checking the Overfitting**

It is important to understand if the model might be overfitting. In this case precision and recall are both quite high and have almost the same score, so it must be checked before moving on. By comparing the performance metrics between the training and test sets will help to assess whether the trained model is overfitting or generalizing well to unseen data.



Then it must be checked whether the regularization applied to the random forest model effectively reduced overfitting and improved performance on unseen data. If the difference is too small than most probably regularization did not help and therefore it is not overfitting.

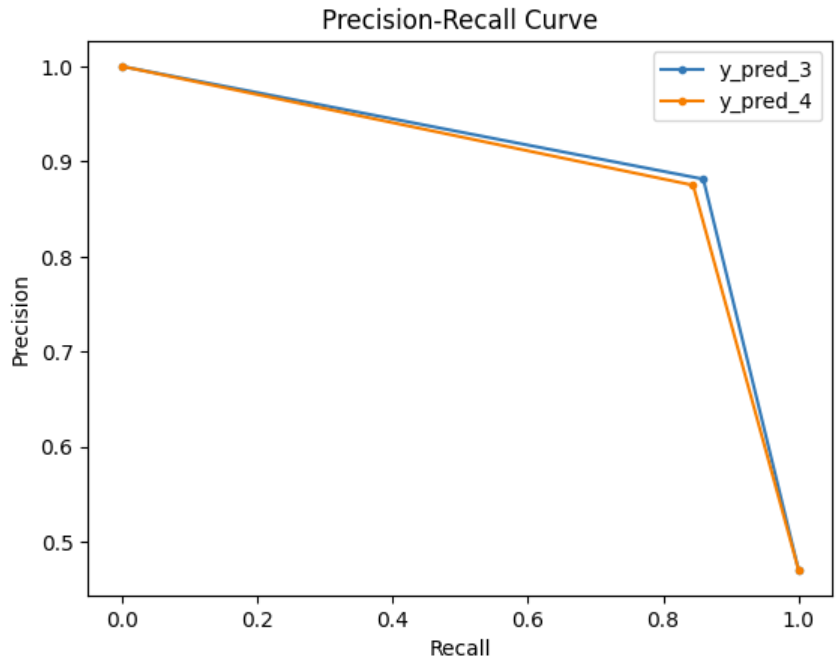


*Fig.4-5: Checking overfitting and checking difference after regularization.*

A scalar value indicating the classifier's prediction accuracy on the test dataset is returned by the F1 score. Better performance is indicated by a greater F1 score, whereas the classifier may require more optimization if the score is lower. As the recall and precision scores are similar to each other and are quite high some would say the model is overfitting and some regularization methods should be used to better define the model. As it is seen in the figures even with regularization method the model did not change dramatically. Therefore, it can be said that the model is very well fine-tuned and it is not overfitting.

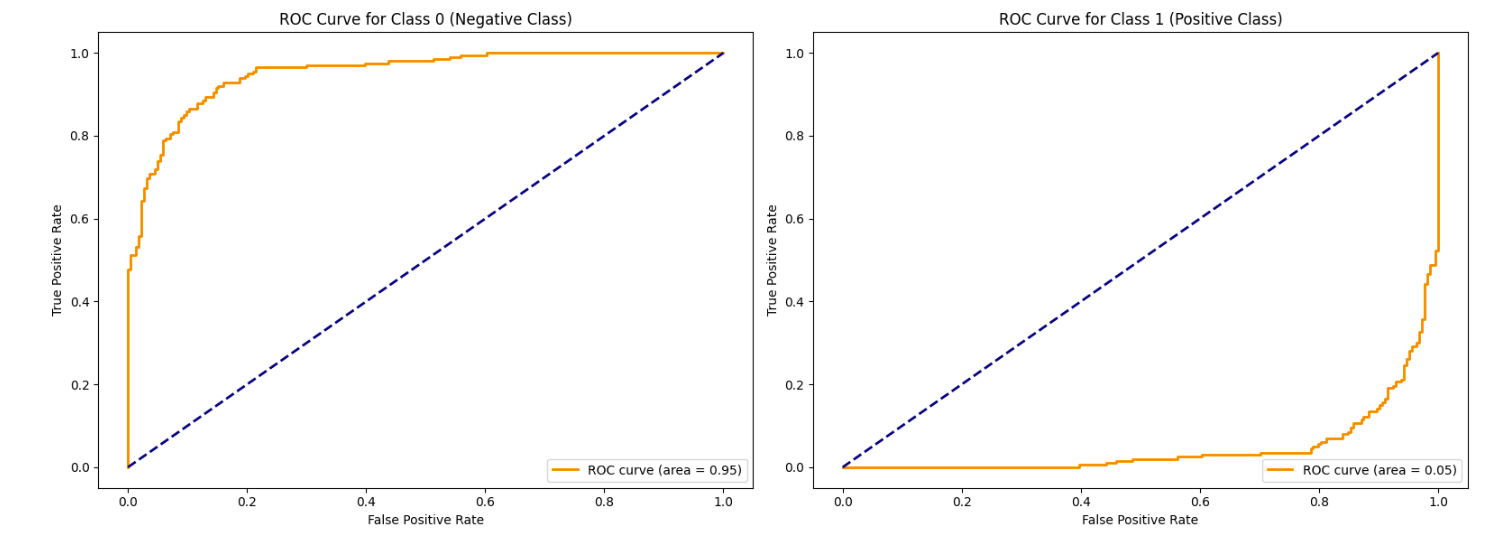
## **3.2 Building the models**

Firstly, dataset shall be splited into train and test tests, so later it can be tested on the training data to see how the model is working. Also, it is better to see a visualization of Precision and Recall for the model\_3. As, it is seen here, one parameter starts going down when the other goes up and the best value will be the harmonic mean F1 score [4].

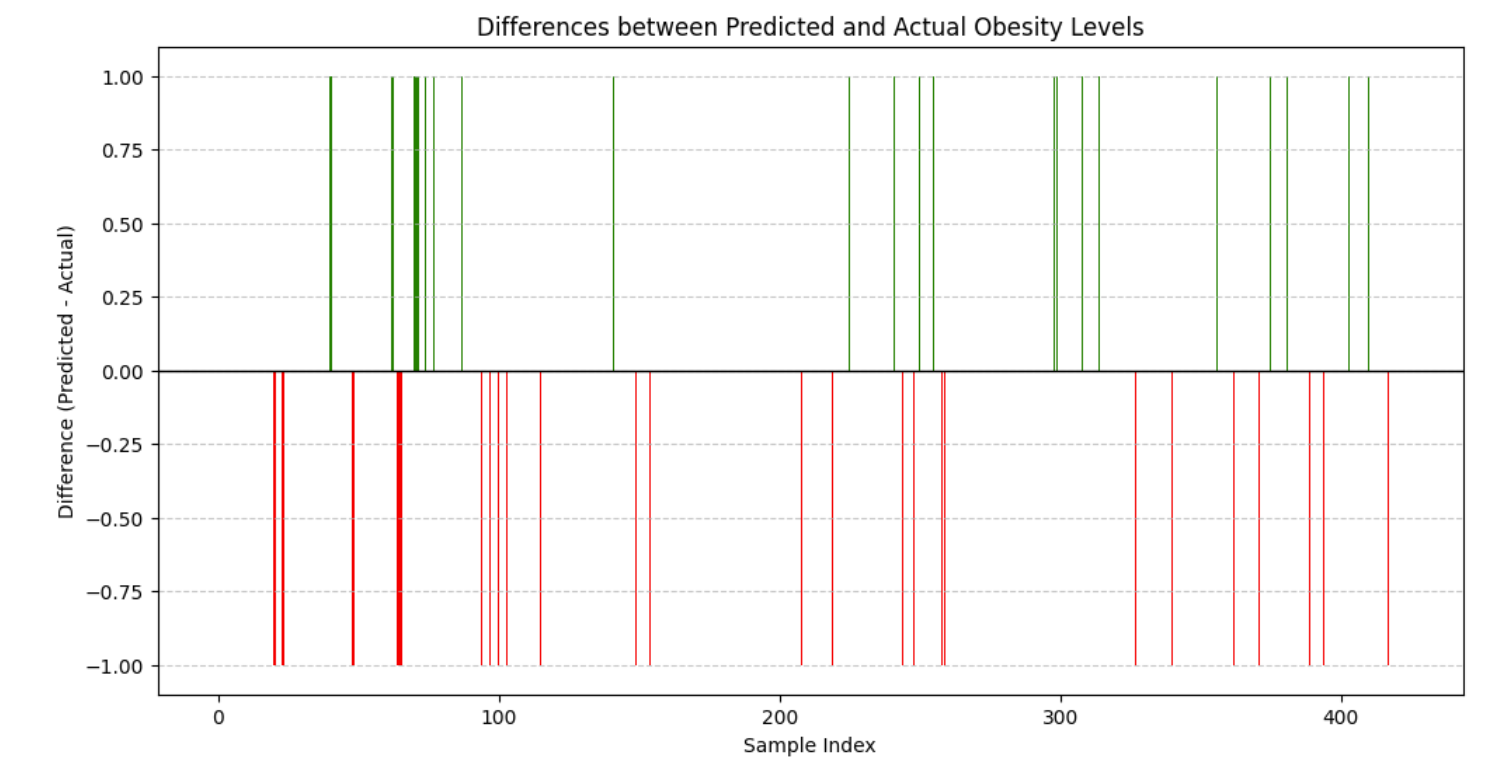


*Fig.6: Recall & Precision for model3 & model4*

Then, Roc was used to understand the performance of classification in rates – FP rate and TP rate. Plotting the Roc curve showed that the model is working very well as it has 95% of accuracy for the True positive. And only 5% for False Positive.



*Fig.7: Classification rates using ROC curve*



*Fig.8: Difference between Predicted and Actual obese levels.*

The differences between the predicted and actual obesity levels for each sample in the test set are visually seen, making it easier to identify where the model performs well or poorly.

The x index indicates the position of the sample test set. The y represents the difference between the predicted obesity level and the actual obesity level for each sample. A positive value indicates that the model overestimated the obesity level, while a negative value indicates that the model underestimated the obesity level.

## **3.3 Working on the model**

BaggingClassifier is a type of ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and aggregates their predictions.

n\_estimators=130 specifies the number of base estimators (decision trees) to create in the ensemble.

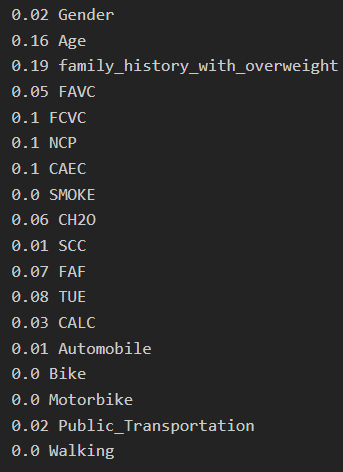
max\_samples=200 indicates the maximum number of samples to draw from the training set to train each base estimator. In this case, it's set to 200 (randomly).

n\_jobs=-1 utilizes all available CPU cores to parallelize the training process, which can speed up computation for large datasets.

random\_state=42 ensures reproducibility by setting a fixed random seed for the random number generator.

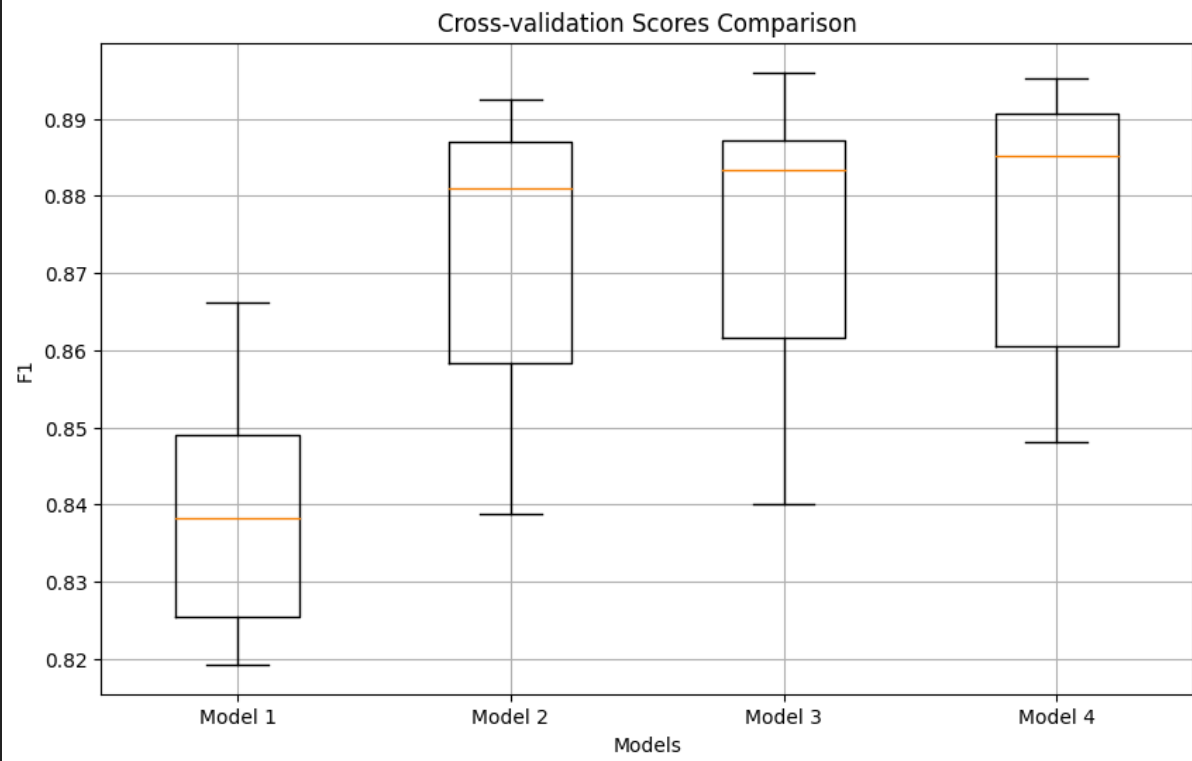
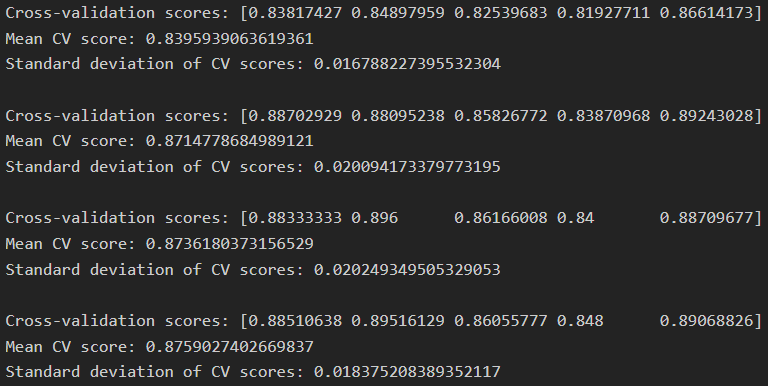
After, a BaggingClassifier ensemble model with 130 decision tree base estimators is defined, it enables OOB scoring during training, and then computes the OOB score as an estimate of the model's performance on unseen data. OOB score is a performance metric which is quite high in this case.

This shows how to retrieve feature importance’s, train a random forest classifier, and determine which features have the greatest influence on classification decisions. So, if there is a need to fine-tune the model or if it is overfitting, surely this would help a lot. It shows that the Age and family history of overweight have the highest impact.



*Fig.9: Impacts on the model from each feature.*

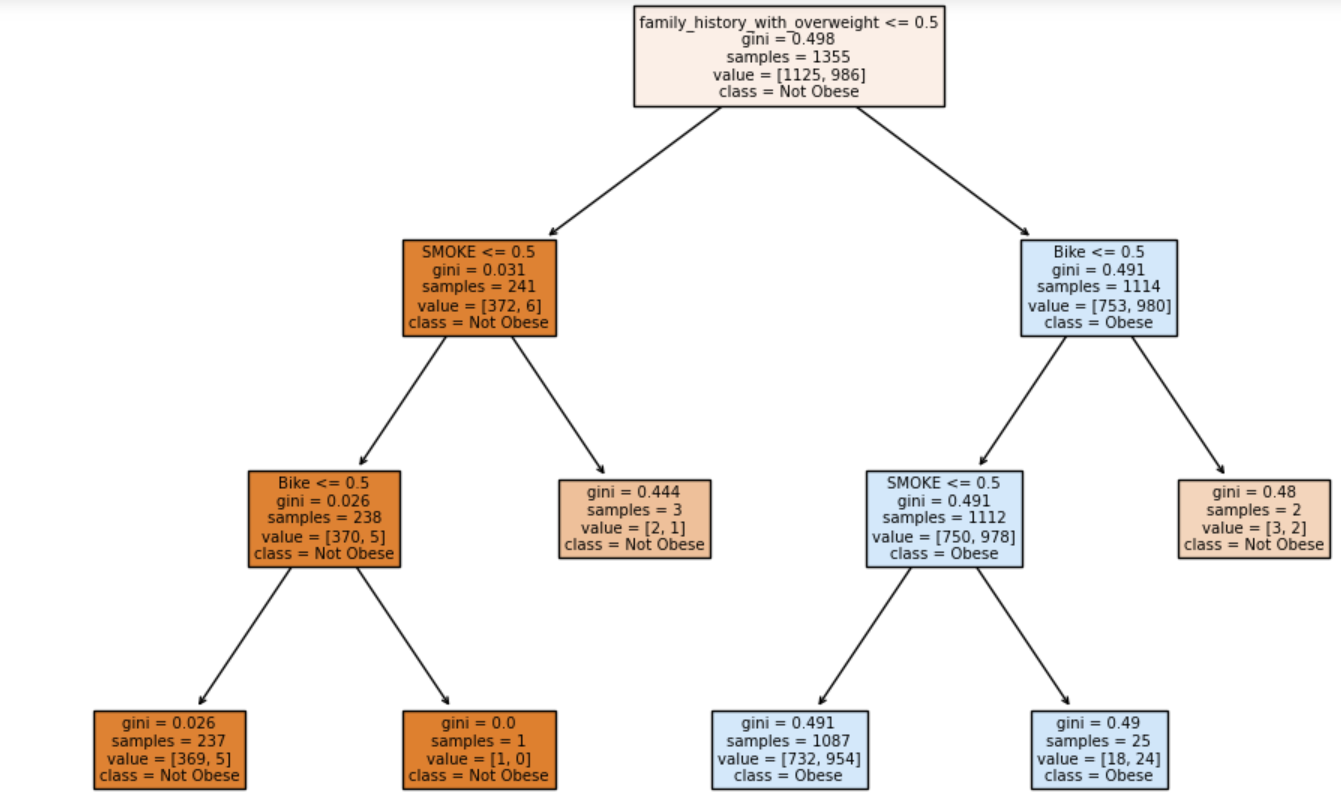
Another important thing to consider is the cross-validation. Cross-validation is used to assess the performance of different Random Forest models using the F1 score as the evaluation metric. The mean cross-validation score gives an estimate of how well each model generalizes to unseen data, and the standard deviation of the scores indicates the variability of the model's performance across different folds. Compared to employing a single train/test split, the primary goal of cross-validation is to get a more accurate estimate of how well a model will perform on unknown data. The boxplot shows the distribution of the cv scores for each model as well as metrics like the median, quartiles, and outliers. This helps to compare the performance of different models and select the one that performs the best on average.

*Fig.10-11: The visualization shows boxplots of the models that enables comparison of the f1 scores using cross-validation*

## **3.4 Working on a leaf from the tree**

It is important to see a decision on a tree depending on some features. It makes easier to understand with the help of which leaf we shall make decisions.



*Fig.12: The visualization shows how the tree splits the data into nodes based on the selected features ('family\_history\_with\_overweight', 'SMOKE', 'Bike').*

Each node presents a decision based on a feature. Features closer to the top of the tree, are considered more important in the classification process.

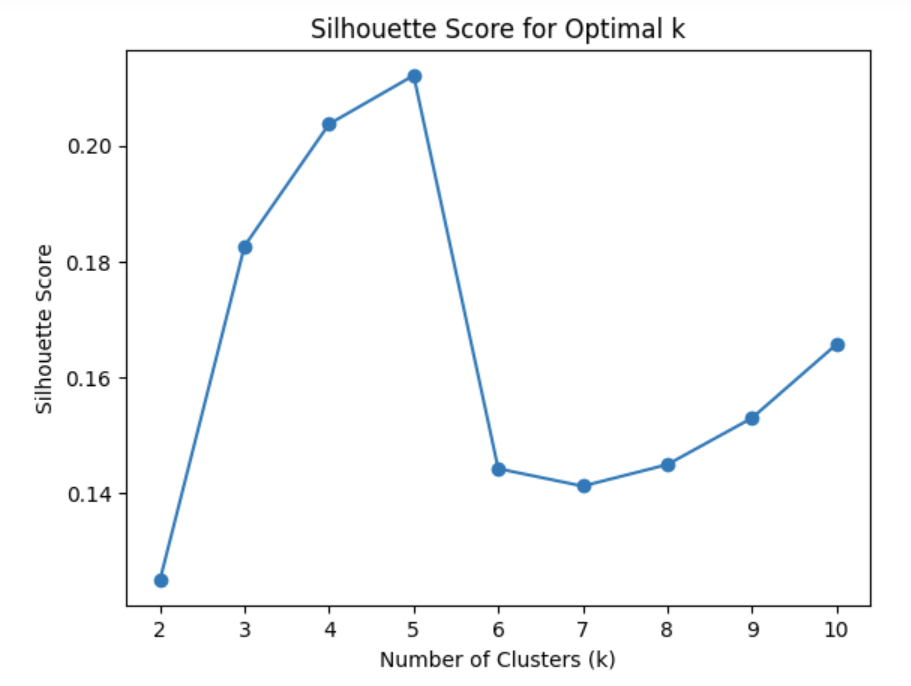
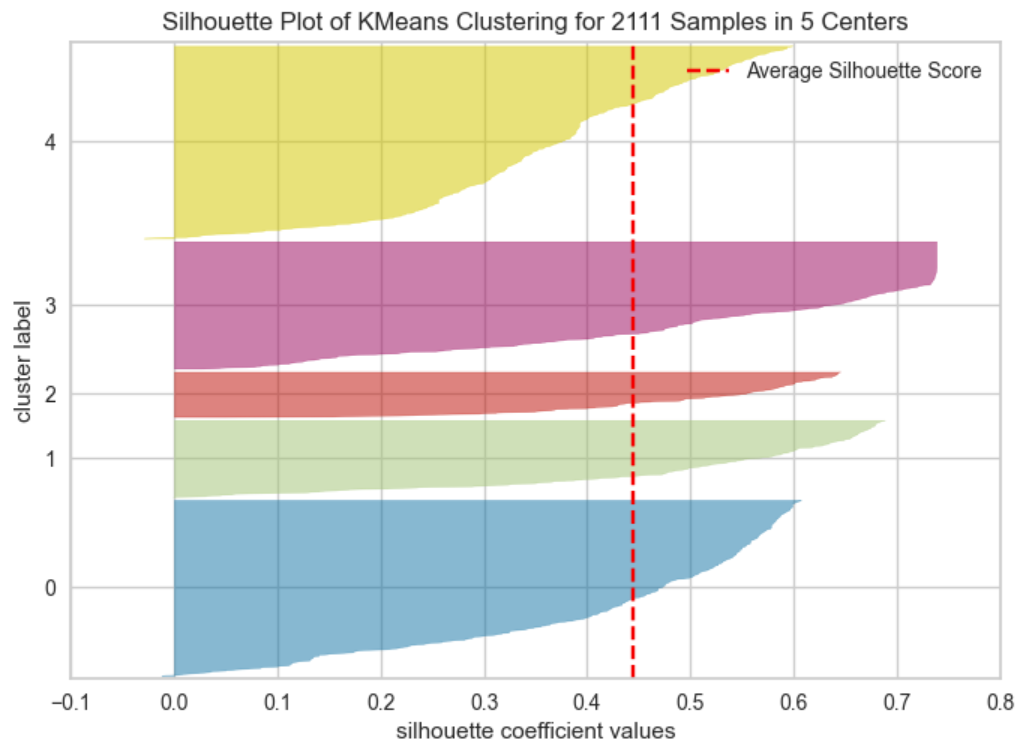
The leaves of the tree represent the final classification outcomes. Each leaf node corresponds to a class label ('Not Obese' or 'Obese') based on the majority class within the node. A node might split the data based on whether a person has a family history of overweight, whether they smoke, or whether they ride a bike.

Understanding how the decision tree is used to select attributes for classifying instances into the intended classes ('Not Obese' or 'Obese') can be beneficial. It gives the decision-making process an easier way with the visualization, which makes it much less complex to understand and explain the behavior of the model [5], [6].

# **4. Clustering**

As clustering is an unsupervised algorithm there is no need to split the data into train and test datasets unless there is a classification problem. Here it is very important to define the number of k’s so when the new data is tested it can identify which cluster it belongs to. If the number of k’s are wrongly defined than the testing data will provide wrong outputs.

One way to find number of k is with elbow method. Elbow method is used to help determine the appropriate number of clusters for K-means clustering based on the inertia values calculated for different values of k. But it isn’t always the best way to identify the number of k. That is why silhouette visualization and score are used for better understanding.

*Fig.13-14: Silhouette score and visualization.*

K = 5 is the best option as for other numbers the visualization looks bad because it has a lot of different shapes, either too long or too fat. Of course this is not the perfect one but a better choice from others. Also, the silhouette score is the highest when k = 5 [7], [8].

**4.1 Working on the model**

The scatter plot shows how the data points are distributed across the 'Age' and 'FCVC' features. There is some overlap between the clusters when there are different colored points among the other colored points and vice versa. This overlap indicates that there are sections in the data where its properties are not well defined. Unfortunately, this has been the best one so far [9].



*Fig.15: Data distribution in Age and FCVC.*

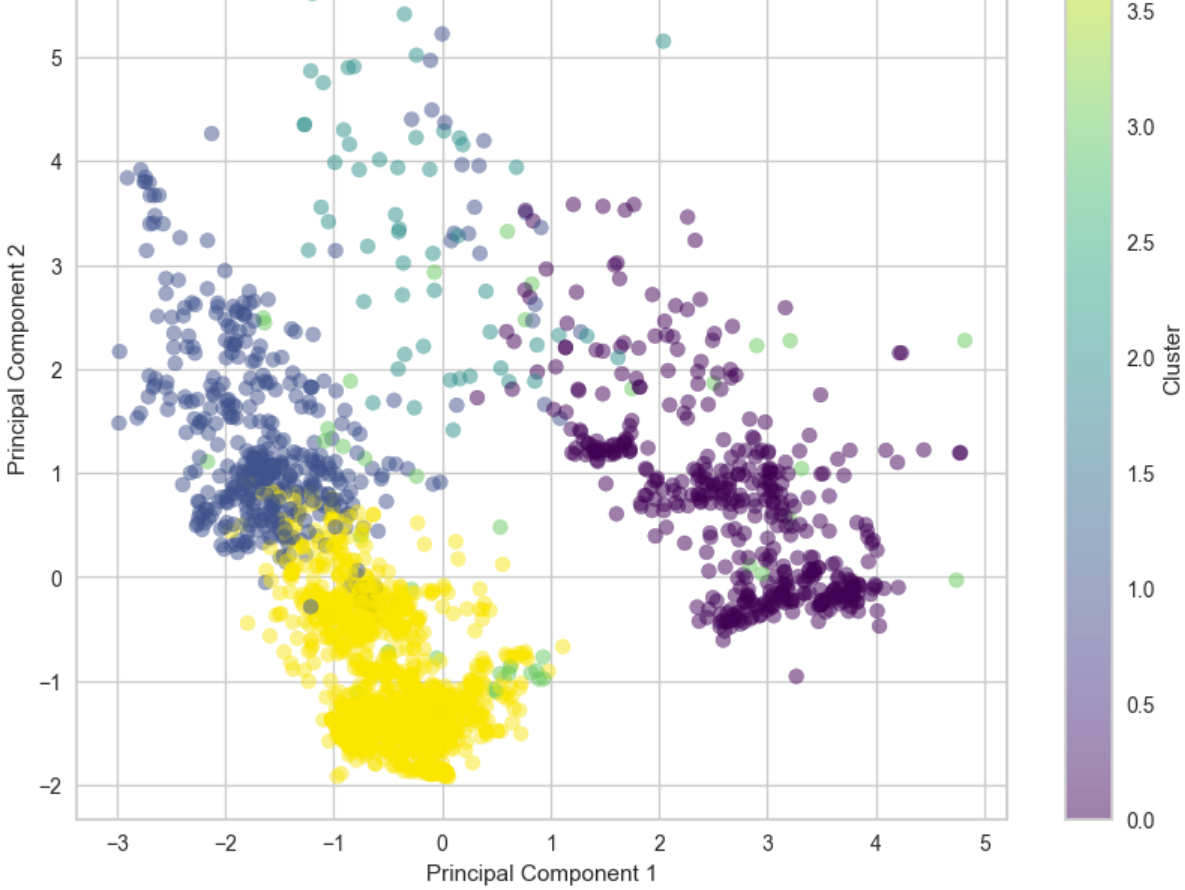
The clustering visualization can help to decide weather the clusters make sense and if the chosen number is the right number of groups. If the clusters match how the data are naturally grouped and if they show useful patterns, it means the clustering is working well. There might be a need to adjust things if the clusters don't seem to represent the data accurately. From the visualization it is clear that people aged more than 35 (green) have a high FCVC and there is another group (purple) that is clearly seen for the age of 28 – 35 having smaller mean of FCVF as there are a lot of points spread form 0 to 4 FCVC. So, if the new unseen data are tested on this data, they will be devided into 5 groups.

## **4.2 Model insights**

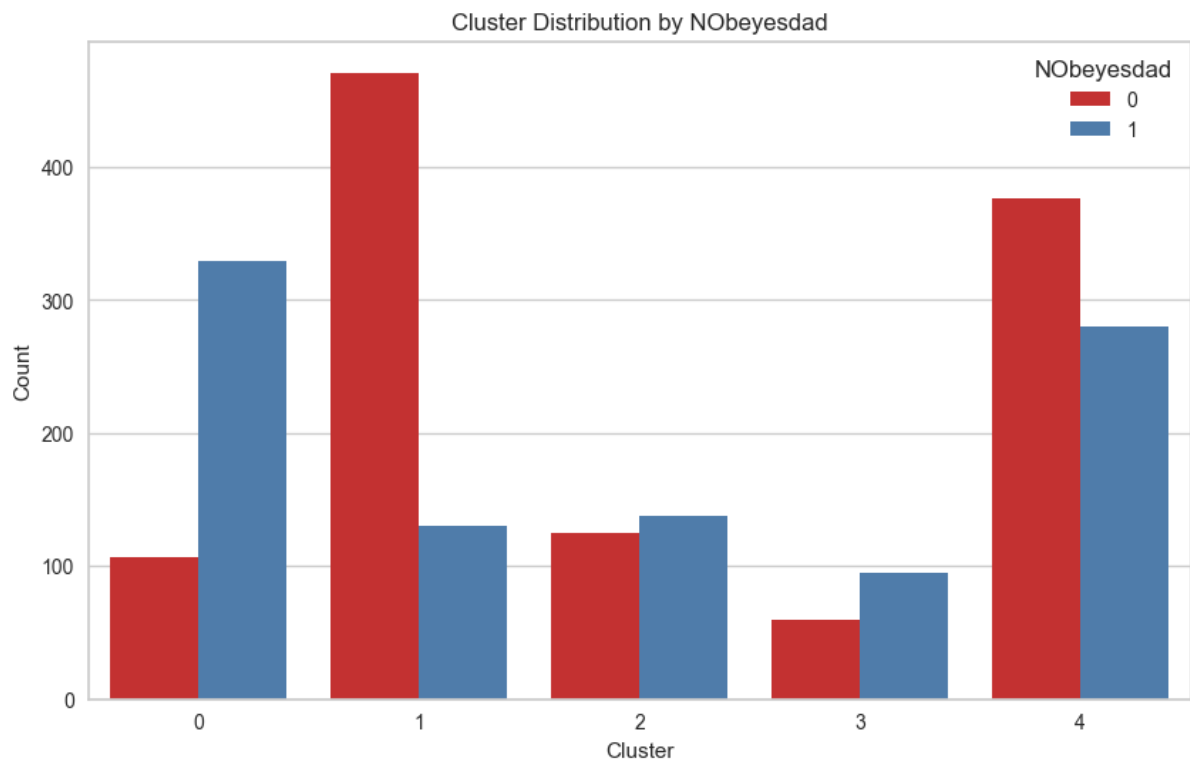
It is performed K means with PCA to have a better visualization of the data with reduced dimensionality into two dimesnions.

Initially, the K Means algorithm was applied to the scaled data with the specified number of clusters, which was set to 5 as Silhouette score was the highest.

PCA was used again to the scaled data to reduce its dimensionality while preserving the most important information. In this case, PCA reduced the data to two principal components having n\_components=2 [10].



*Fig.16: Using PCA to all fetures for clustering.*



*Fig.17: It shows a vizualization of distribution of people who are obese/not obese within our clusters.*

It is clear that there are more people with obesity in cluster 1 and cluster 5. Cluster 3 and cluster 4 are similar to each other which indicates that we could have picked initially 4 clusters to see if anything would have changed if we were to fine tune it.

# **5. Conclusion**

In conclusion, this project is about how to use machine learning to predict the likelihood of obesity using supervised (decision trees and Random Forest) and unsupervised learning (clustering analysis and PCA). The investigation began with thorough data preprocessing, where missing values were checked, and feature scaling was applied to ensure the dataset's integrity and usability. Visualizations were part of a very important and useful information making decisions with the analysis. Through meticulous evaluation of metrics like precision, recall, and F1 score, the study aimed to strike a balance between model complexity and predictive accuracy while guarding against overfitting. Furthermore, the examination of overfitting revealed that the model, despite high precision and recall scores, required consideration of regularization methods. The results of clustering techniques, such as the silhouette score, helped to determine the optimal number of clusters for the data. In summary, the study continues to contributes in public health sector for it to be improved.

References

[1]<https://www.geeksforgeeks.org/decision-threshold-in-machine-learning/?fbclid=IwAR1iT1t7cbrJoZTkJiCILrltg-9LErvt3TmLdQ8WczYqi0vK2ElrXyzzK44>

[2] <https://medium.com/@ttc4best/decision-tree-assumptions-advantages-disadvantages-applications-46f7e5807517>

[3] <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>

[4] <https://scikit-learn.org/stable/modules/tree.html>

[5] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106 available: <https://hunch.net/~coms-4771/quinlan.pdf>

[6] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. CRC press, available: <https://rafalab.dfci.harvard.edu/pages/649/section-11.pdf>

[7] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics,28(1), 100-108. Available: <https://www.stat.cmu.edu/~rnugent/PCMI2016/papers/HartiganKMeans.pdf>

[8] <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>

[9] <https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html>

[10] <https://neptune.ai/blog/k-means-clustering>