UNIVERSITY
OF LONDON

ST3189 Machine Learning

Student Number : 200579973

Table of Contents

1. Introduction

This paper examines how to predict Body Mass Index (BMI) and categorize obesity ("NObeyesdad") according to lifestyle characteristics such food intake, use of technology, and physical activity. In order to identify important factors controlling BMI and comprehend their influence on health outcomes, we will employ both supervised and unsupervised learning techniques.

2. Dataset Information

Based on eating patterns and physical characteristics, this dataset provides information for predicting the prevalence of obesity in people from Mexico, Peru, and Colombia between the ages of 14 and 61. It has 2111 data and 17 attributes that were taken from the UC Irvine Machine Learning Repository (Fabio & Alexis , 2019).

2.1 Features

Additionally, added BMI column by $\frac{weight}{height^2}$.

Categorical	**Description**	**Numerical**	**Description**	**Target variable**	**Description**
Gender	Female/Male	FCVC	Vegetables Frequency	NObeyesdad	Obesity Level
Family history	Family obesity history	NCP	Main meals	BMI	
FAVC	High-caloric Food	FAF	Exercise Frequency		
CAEC	Between-mean intake	TUE	Electronic Time		
SMOKE	Do you smoke?	Age	numeric		
SCC	Caloric tracking	Height	In metres		
MTRANS	Transportation mode	Weight	In kilograms		
CALC	Alcohol Intake	CH20	Water Intake		

3. Exploratory Analysis

3.1 Research Questions

1. Given the significance of the respondents' demographics in shaping the dataset's findings, what specific traits are present among the individuals included in this dataset?

2. Is it feasible to utilize Body Mass Index (BMI) as a quantitative proxy for the qualitative weight classification category?

3.2 Analysis

1. The dataset is loaded and checked for duplicates and missing values. Any duplicated rows are removed to ensure data integrity.

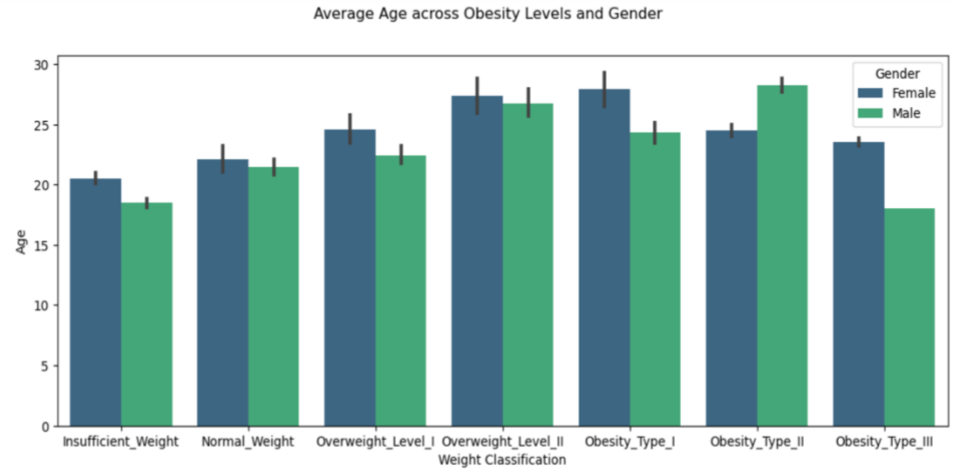


Figure 1. Average Age Across Weight Categories

In Figure 1, as we progress from the lowest to the highest weight categories, the average age within each category tends to increase. This indicates a likely correlation between age and being overweight, suggesting that as individuals age, they may be more inclined to higher weight categories.

Furthermore, we noted a slight variation in average age between males and females across the weight categories. Except for type II obesity, females generally showed a slightly higher average age compared to males in most weight categories. This insight suggests that gender might also influence weight tendencies, with females tending to have a higher average age across different weight categories.

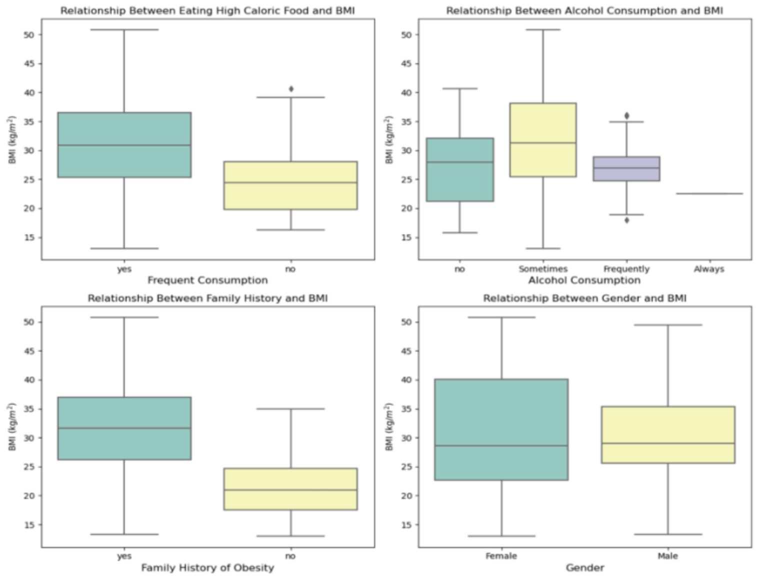


Figure 2. Relationship Between BMI and Various Factors

An examination of the correlation between BMI, an indicator of obesity, and different category characteristics yielded important information on influential factors. Based in figure 2, the first subplot shows a discernible difference in the median BMI between people who regularly eat foods high in calories and people who don't. Moving to the second subplot, it is notable that there seems to be minimal association between alcohol consumption and BMI. However, suggesting that the lack of responses in this category might be affecting the observed relationship. This indicates a need for more data responses to understand better the impact of alcohol consumption on BMI. The third subplot, which displays a significant variation in median BMI, emphasizes the impact of family history on obesity. A median BMI that is around 10 kg/m² higher among those with a family history of obesity suggests that hereditary factors play a major role in an individual's weight status. In contrast, the fourth subplot reveals no discernible relationship between gender and BMI, with both genders displaying similar median BMIs.

2.

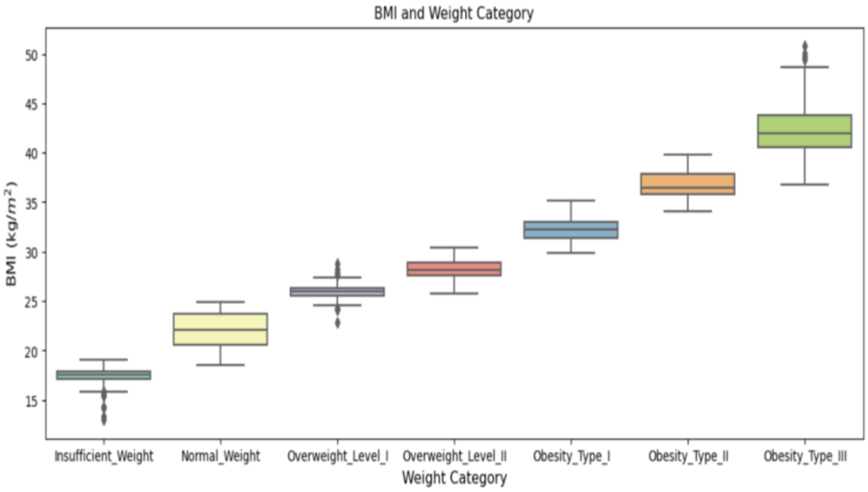


Figure 2. Relationship Between BMI and Various Factors3

Moreover in Figure 3 highlights a distinct correlation between BMI levels and the various weight categories, affirming the use of BMI in categorizing weight classes. Thus, it emphasizes the efficacy of BMI in characterizing different weight groups. This reinforces the notion that BMI serves as a practical tool for classifying individuals based on weight, particularly evident by the minimal presence of outliers. The strong correlation between BMI levels and weight categories further supports BMI's utility as a quantitative variable.

4. Unsupervised Learning

4.1 Research Questions

1. What are the key factors contributing to obesity status, and how do these variables interplay according to the Principal Component Analysis?
2. Can we accurately predict the obesity type of individuals based on their eating habits?

4.2 Analysis

1. PCA is a dimensionality reduction technique that helps us identify the most important variables that explain the variance in the data. To start off, PCA is sensitive to the scale of the variables, so it's important to standardize the data, and create dummy variables for the numerical variables.

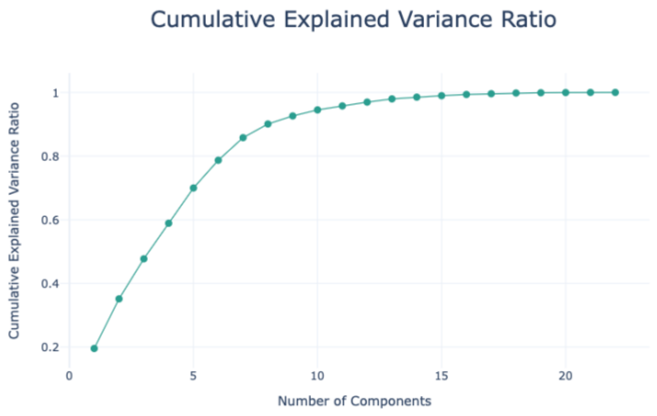
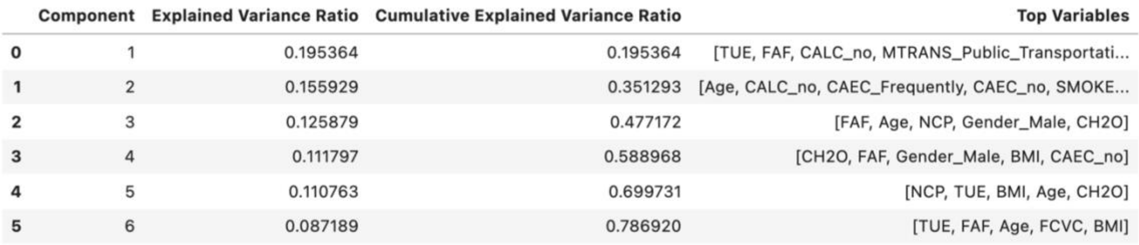


Figure 4. Cumulative Explained Variance

It focuses on the first five components which collectively explain over 70% of the variance. Component 0, with an explained variance ratio of 19.54%, highlights variables like "Time Using Technology" (TUE) and "Physical Activity Frequency" (FAF) as significant contributors. In Component 1, explaining 15.59% of the variance, "Age" emerges as a prominent factor, indicating a correlation between age and obesity.

Component 2, accounting for 12.59% of the variance, showcases the importance of "Physical Activity Frequency" (FAF), "Age," and "Number of Main Meals" (NCP). It suggests that individuals with higher FAF and NCP values might be more prone to obesity.

Moving to Component 3, which explains 11.18% of the variance, we observe a focus on "Water Consumption Daily" (CH2O), "Physical Activity Frequency" (FAF), and "Gender" (Male). This component suggests a potential gender-based difference in obesity factors. Finally, Component 4 with an explained variance of 11.08% emphasizes variables such as NCP, TUE, BMI, and Age. These variables indicate a potential association between the frequency of meals, sedentary behaviour, and obesity status.



	Component	Explained Variance Ratio	Cumulative Explained Variance Ratio	Top Variables
0	1	0.195364	0.195364	[TUE, FAF, CALC_no, MTRANS_Public_Transportati...
1	2	0.155929	0.351293	[Age, CALC_no, CAEC_Frequently, CAEC_no, SMOKE...
2	3	0.125879	0.477172	[FAF, Age, NCP, Gender_Male, CH2O]
3	4	0.111797	0.588968	[CH2O, FAF, Gender_Male, BMI, CAEC_no]
4	5	0.110763	0.699731	[NCP, TUE, BMI, Age, CH2O]
5	6	0.087189	0.786920	[TUE, FAF, Age, FCVC, BMI]

Figure 5. PCA Summary Table

To sum up, among these five elements, the most crucial variables are TUE, FAF,NCP, Age, CH2O,all of which have been shown to have a substantial impact on the state of obesity.

2. Based on the three clustering models - K-means, Hierarchical Clustering, and DBSCAN. Based on similarities, K-means clustering divides people into groups. First, use the Elbow Method to determine the number of clusters.

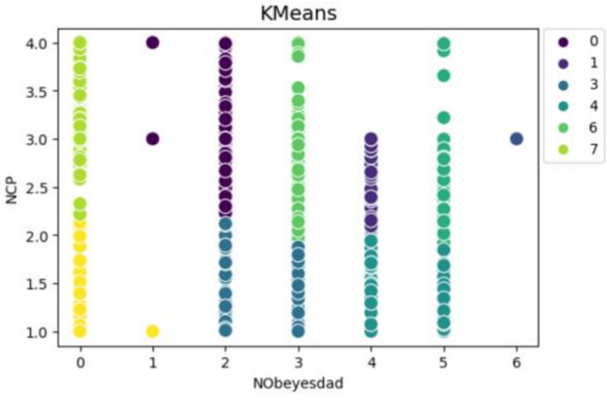


Figure 6. K-means Scatter Plot

Based on Figure 6, it can be shown that group of individuals with the highest levels of obesity and lowest frequency of main meals are cluster 2 in K-means model. The degree to which the clusters are compact and well-separated is indicated by the Silhouette Score. The score is -1 to 1 with the higher score the better. Hence, based on the provided silhouette scores and the best number of clusters for each method,

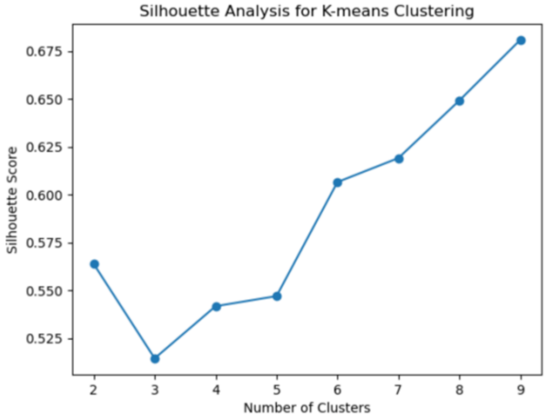


Figure 7. Silhouette Analysis for K-means

The silhouette score attained using K-means clustering is 0.6809 indicates that the clusters are well-separated. The silhouette score for Hierarchical is 0.6769 and silhouette score of DBSCAN is 0.5362. In summary K-means clustering shows clear and distinct clusters, indicating significant differences in obesity levels based on meal frequency. Hierarchical clustering offers a hierarchical view of cluster relationships, providing insights into varying levels of similarity among groups with different meal frequencies and obesity levels. Yet, DBSCAN, despite having a lower silhouette score, is effective in identifying outliers, which may represent unique patterns of meal frequency and obesity levels.

5. Regression Model

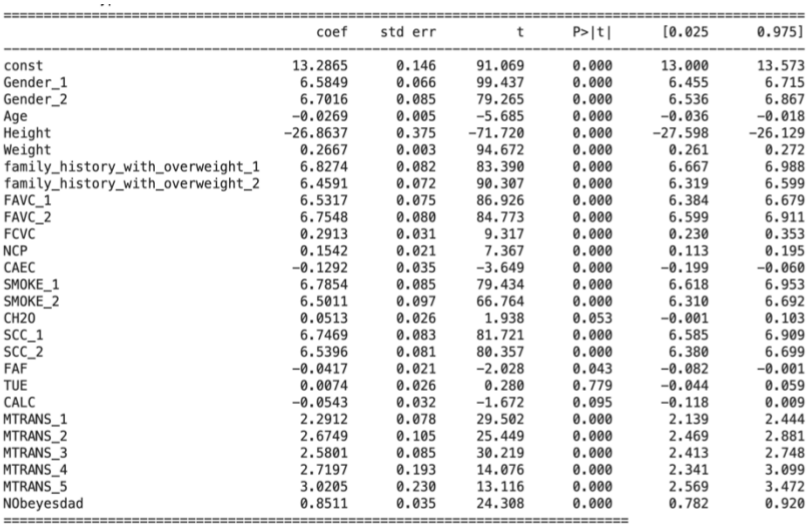
5.1 Research Questions

1. What features contribute significantly to the prediction of obesity levels, and how does the model's performance compare after excluding non-significant features?
2. How accurately can we predict Body Mass Index (BMI) using different regression models?

5.2 Analysis

1. The "BMI" (body mass index), a continuous numerical variable that indicates a person's body weight in relation to their height, is the target variable of interest in the regression analysis.

First, examine the association between a dependent variable (obesity level) and one or more independent variables (eating habits and lifestyle factors) using the Ordinary Least Squares (OLS) approach. The OLS model demonstrates a high R-squared value of 0.993, indicating that approximately 99.3% of the variance in the obesity level (dependent variable) is explained by the features (independent variables). This suggests a strong fit of the



```
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                               13.2865      0.146     91.069      0.000      13.000      13.573
Gender_1                             6.5849      0.066     99.437      0.000       6.455       6.715
Gender_2                             6.7016      0.085     79.265      0.000       6.536       6.867
Age                                 -0.0269      0.005     -5.685      0.000      -0.036      -0.018
Height                             -26.8637      0.375    -71.720      0.000     -27.598     -26.129
Weight                               0.2667      0.003     94.672      0.000       0.261       0.272
family_history_with_overweight_1     6.8274      0.082     83.390      0.000       6.667       6.988
family_history_with_overweight_2     6.4591      0.072     90.307      0.000       6.319       6.599
FAVC_1                               6.5317      0.075     86.926      0.000       6.384       6.679
FAVC_2                               6.7548      0.080     84.773      0.000       6.599       6.911
FCVC                                 0.2913      0.031      9.317      0.000       0.230       0.353
NCP                                  0.1542      0.021      7.367      0.000       0.113       0.195
CAEC                                -0.1292      0.035     -3.649      0.000      -0.199      -0.060
SMOKE_1                              6.7854      0.085     79.434      0.000       6.618       6.953
SMOKE_2                              6.5011      0.097     66.764      0.000       6.310       6.692
CH2O                                 0.0513      0.026      1.938      0.053      -0.001       0.103
SCC_1                                6.7469      0.083     81.721      0.000       6.585       6.909
SCC_2                                6.5396      0.081     80.357      0.000       6.380       6.699
FAF                                 -0.0417      0.021     -2.028      0.043      -0.082      -0.001
TUE                                  0.0074      0.026      0.280      0.779      -0.044       0.059
CALC                                -0.0543      0.032     -1.672      0.095      -0.118       0.009
MTRANS_1                             2.2912      0.078     29.502      0.000       2.139       2.444
MTRANS_2                             2.6749      0.105     25.449      0.000       2.469       2.881
MTRANS_3                             2.5801      0.085     30.219      0.000       2.413       2.748
MTRANS_4                             2.7197      0.193     14.076      0.000       2.341       3.099
MTRANS_5                             3.0205      0.230     13.116      0.000       2.569       3.472
NObeyesdad                           0.8511      0.035     24.308      0.000       0.782       0.920
==============================================================================
```

Figure 8. OLS Regression Results

model to the data, indicating that the selected independent variables can effectively predict obesity levels based on eating habits and lifestyle factors

With p-value is less than 5% significance level, we reject the null hypothesis (the coefficient is not significant) and conclude that the coefficient is statistically significant. TUE (Time Using Technology) is not statistically significant as p-value is 0.779 meaning it does not have a significant impact on the target variable. CH2O (Consumption of Water Daily) and CALC (Consumption of Alcohol) are borderline significant, meaning they are close to being statistically significant but not definitively so.

Upon examining the statistical significance of individual features, we found that TUE was not statistically significant. Additionally, CH2O and CALC showed borderline significance. This insight prompts us to consider whether these variables should be retained in the model or dropped. We will now analyse the model performance metrics before dropping the non-significant features and after removing TUE, CH2O, and CALC.

Before dropping		After dropping	
Mean Absolute Error (MAE)	0.2939	Mean Absolute Error (MAE)	0.2960
Mean Squared Error (MSE)	0.1447	Mean Squared Error (MSE)	0.1456
R-squared (R2)	0.9651	R-squared (R2)	0.9649

The comparison between the two sets of metrics shows a minimal change in performance after dropping the non-significant variables. The MAE increased slightly from 0.2939 to 0.2960, the MSE increased from 0.1447 to 0.1456, and the R-squared decreased marginally from 0.9651 to 0.9649. These changes suggest that dropping the non-significant features did not significantly impact the model's predictive capability. This means that the simplified model, after removing TUE, CH2O, and CALC, still retains a high level of accuracy in predicting obesity type based on eating habits. The slight changes in performance metrics are within an acceptable range, indicating that the model's overall effectiveness remains largely unchanged.

Eventually, the elimination of the non-significant features appears to have been a wise option because it results in a more understandable and straightforward model without compromising predictive accuracy. This result is consistent with the Occam's razor approach, which favours simpler models where performance is still similar. Therefore, we can state with confidence that, using the simplified model that still includes important properties, we can reliably predict the kind of obesity based on the modified study.

2.

Regression Models	MAE	MSE	R-squared	RMSE
Linear regression	0.5457	0.4948	0.9922	0.7034
Gradient Boosting	0.4808	0.3963	0.9937	0.6295
Random Forest	0.2914	0.1727	0.9972	0.4156
Neural Network	0.9831	1.5364	0.9759	1.2395

Based on the performance regression models table, Random Forest exhibits the best performance among the models with an R-squared value of 0.997, suggesting that 99.7% of the variance in BMI is explained by the features. The MAE, MSE, and RMSE are the lowest among the models, indicating the highest accuracy and precision. Followed by gradient boosting with an R-squared value of 0.994, thus shows slightly better performance than Linear Regression. It has a lower MAE, MSE, and RMSE, indicating improved accuracy in predicting BMI. Third position is Linear Regression, this model fits a linear equation to the data, making it simple and easy to interpret. It assumes a linear relationship between the input features and the target variable which is BMI and lastly is Neural Network model has the highest MAE, MSE, and RMSE among the models, suggesting that it may not perform as well as the other models in predicting BMI. Therefore, Random Forest emerges as the top-performing model for predicting BMI, with the lowest errors and highest R-squared value.

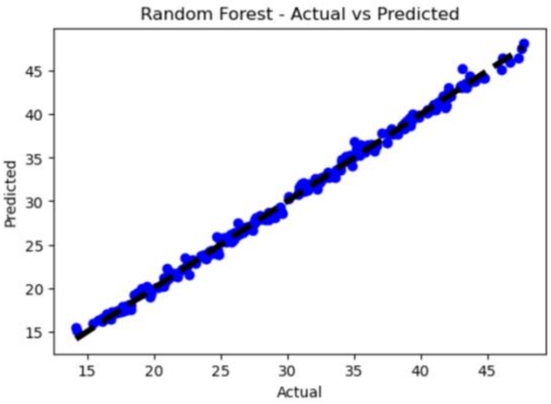


Figure 9. Random Forest Graph

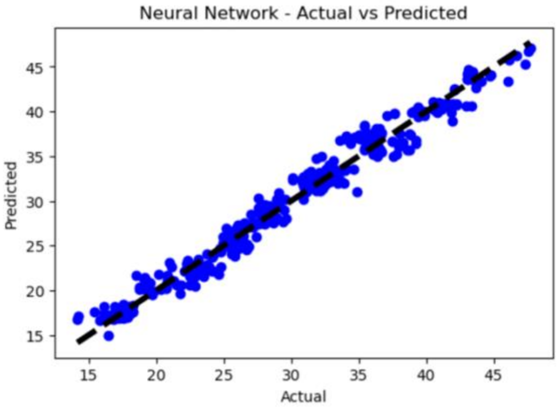


Figure 10. Neural Network Graph

To evaluate the performance and reliability of the regression models, we will further compare the difference from the strongest and weakest model which is Random Forest and Neural Network in this case to see the predictions that are closer to the actual values. The Random Forest model's scatter plot demonstrates an accurate linear correlation between the observed and expected BMI values. Given that the majority of the points are strongly concentrated around the diagonal line (black dashed line), the model's predictions appear to be very accurate. The few spots that stray from the diagonal line indicate some situations in which the model's predictions and the actual BMI values don't match. As a result, Random Forest model predicts BMI levels rather well overall. Meanwhile, the actual and projected BMI values similarly exhibit a roughly linear connection in the scatter plot for the Neural Network model.

In contrast to the Random Forest plot, there is a greater dispersion of points, with a greater number of points situated beyond the diagonal line. Between predicted and actual BMI values, the Neural Network model shows an adequate linear connection. The occurrence of outliers and a greater range of points, however, indicate that the model's predictions are not as reliably accurate as those of the Random Forest model. Moreover, higher prediction errors for some cases may arise from the model's difficulty interpreting specific patterns in the data. To enhance its functionality and lower forecast mistakes, it might need more fine-tuning or modifications.

6. Classification Model

6.1 Research Questions

1. Which classification algorithm performs best in predicting obesity levels using selected features, based on key performance metrics?
2. Does hyperparameter tuning improve the Random Forests model's performance?

6.2 Analysis

The target variable of interest in the classification analysis is "NObeyesdad," which stands for an individual's obesity status. The "NObeyesdad" variable has several categories, including "Insufficient Weight," "Normal Weight," "Overweight Level I," "Overweight Level II," "Obesity Type I," "Obesity Type II," and "Obesity Type III."

Classification Models	Accuracy	Recall	Precision	F1 Score	Accuracy
Random Forest Classifier	0.972	0.971	0.973	0.971	0.009
Decision Tree Classifier	0.969	0.968	0.969	0.968	0.018
KNeighbors Classifier	0.958	0.957	0.960	0.956	0.009
SVC	0.922	0.921	0.931	0.918	0.012
Logistic Regression	0.726	0.721	0.714	0.710	0.029
Linear SVC	0.632	0.636	0.682	0.607	0.050
AdaBoost Classifier	0.553	0.551	0.339	0.441	0.045

Beforehand, There are importance of metrics in this Classification task which are accuracy where it shows how well the model's predictions are predicted overall, recall where it evaluates how well the model identifies true positives, or positively anticipated cases. Precision where it indicates how well the model can detect positive predictions without producing false positives, and F1-score where it is a balance between precision and recall calculated as the harmonic mean of the two measures (Analytics Vidhya, 2021).

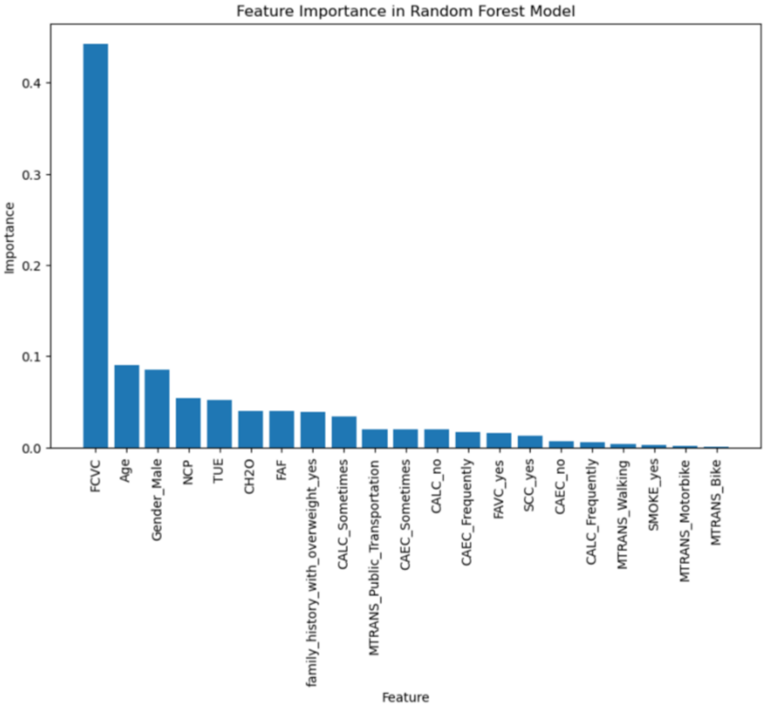


Figure 11. Feature Importance Random Forest

The Random Forest Classifier has superior performance with the highest accuracy, recall, precision, and F1 score, as determined by these measures. It accurately classified 97.2% of the occurrences, with an accuracy of 97.2%. The results provide a balanced performance in predicting the levels of obesity, as seen by the strong recall, precision, and F1 score. The accuracy standard deviation is minimal (0.009), indicating that the model performs consistently. Notably, the 'FCVC' (Frequency of Consumption of Vegetables) feature stands out as the most important, suggesting that individuals' vegetable consumption habits significantly influence their obesity levels according to the model.

Following closely after with comparable high ratings in accuracy, recall, precision, and F1 score is the Decision Tree Classifier. It does a good job of identifying the various degrees of obesity, with an accuracy of 96.9%. Conversely, AdaBoost Classifier performs the worst out of all the models, successfully classifying 55.3% of the examples with an accuracy of 55.3%. Compared to other models, it may have more false positives because to its low precision of 39.9%. In conclusion, the Random Forest Classifier outperforms the Decision Tree Classifier in terms of accuracy when it comes to predicting obesity levels based on specific variables.

2.

Before Hyperparameter Tuning		After Hyperparameter Tuning	
Accuracy	0.981	Accuracy	0.983
Precision	0.982	Precision	0.984
Recall	0.980	Recall	0.983
F-1	0.980	F-1	0.983

It's seen from comparing multiple models that the Random Forest model generally performs better than the others. Prior to tuning, the RF model most likely suffered from overfitting, which is akin to learning by heart certain instances instead of comprehending the underlying patterns. Generalization suffers by this overfitting. As such, the model would probably perform poorly on fresh, unseen data even though it may have scored perfectly on known data. This overfitting problem has probably been reduced by adjusting the RF model's hyperparameters, which has improved the model's ability to generalize to new situations.

Thus, Following hyperparameter modification, the model's accuracy increased, suggesting that it is now correctly predicting a greater percentage of the test data. Increased precision indicates the classifier's capacity to avoid labelling a negative sample as positive. Less false positives are indicated by this. The classifier's recall, which gauges its capacity to identify every positive sample, also went up. Less false negatives are indicated by this. The precision and recall harmonic mean, or F1-score, also went up. This suggests that the model has generally been better at striking a balance between recall and precision.

Furthermore, we can determine the prediction accuracy on the matching OOB data for each tree in the Random Forest and compare the OOB score and Random Forest accuracy to? . The model performs well on unseen data points, as indicated by the OOB score of 0.9639, which offers a reliable measure of its generalization capabilities. As a result, the model appears to perform extremely well on certain data samples that were not included during training, based on the high test set accuracy of 0.9835. The model's performance on unseen data OOB is fairly similar to its performance on the test set, as evidenced by the tiny difference 0.0196 between the OOB score and the test set accuracy. Following hyperparameter adjustment, the Random Forest model exhibits a high degree of accuracy on both the test set and unseen data (OOB samples).

7. References

Analytics Vidhya. (n.d.). Metrics to Evaluate Your Classification Model to Take the Right Decisions. [online] Available at: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right decisions/#:~:text=Classification%20Metrics%20like%20accuracy%2C%20precision,in%20evaluating%20the%20model%20performance [Accessed 8th March 2024].

ScienceDirect. (n.d.). Predicting obesity levels using machine learning techniques. [online] Available at: https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub [Accessed 8th March 2024].

UCI Machine Learning Repository. (n.d.). Estimation of Obesity Levels Based on Eating Habits and Physical Condition. [online] Available at: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition [Accessed 8th March 2024].