Obesity Classification

Main objective

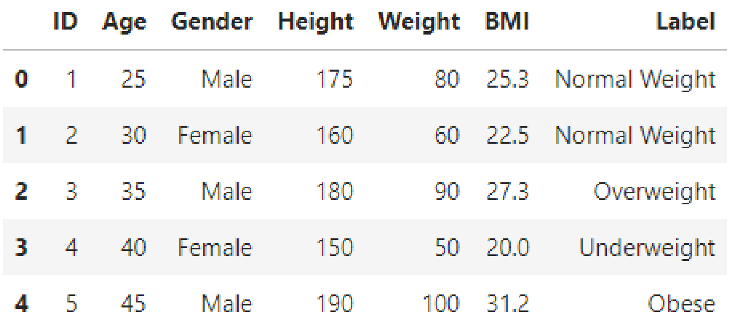
In this analysis we gain a deeper understanding of the effect of various features such as Weight, Age and Gender, on obesity. In addition to this, we create 3 models which, given relevant features, classify an individual as Underweight, Normal Weight, Overweight or Obese.

This analysis provides insight into the factors that affect obesity, providing medical practitioners with background information which can be used to classify patients' obesity levels and administer the required prescription to remedy the situation if there is a problem.

Dataset

The dataset for this analysis was found on Kaggle. It contains information about the obesity classification of individuals. The data was collected from a variety of sources, including medical records, surveys, and self-reported data. The dataset includes the following columns:

- ID: A unique identifier for each individual.
- Age: The age of the individual.
- Gender: The gender of the individual.
- Height: The height of the individual in centimetres.
- Weight: The weight of the individual in kilograms.
- BMI: The body mass index of the individual, calculated as weight divided by height squared.
- Label: The obesity classification of the individual, which can be one of the following:
  1. Underweight
  2. Normal Weight
  3. Overweight
  4. Obese



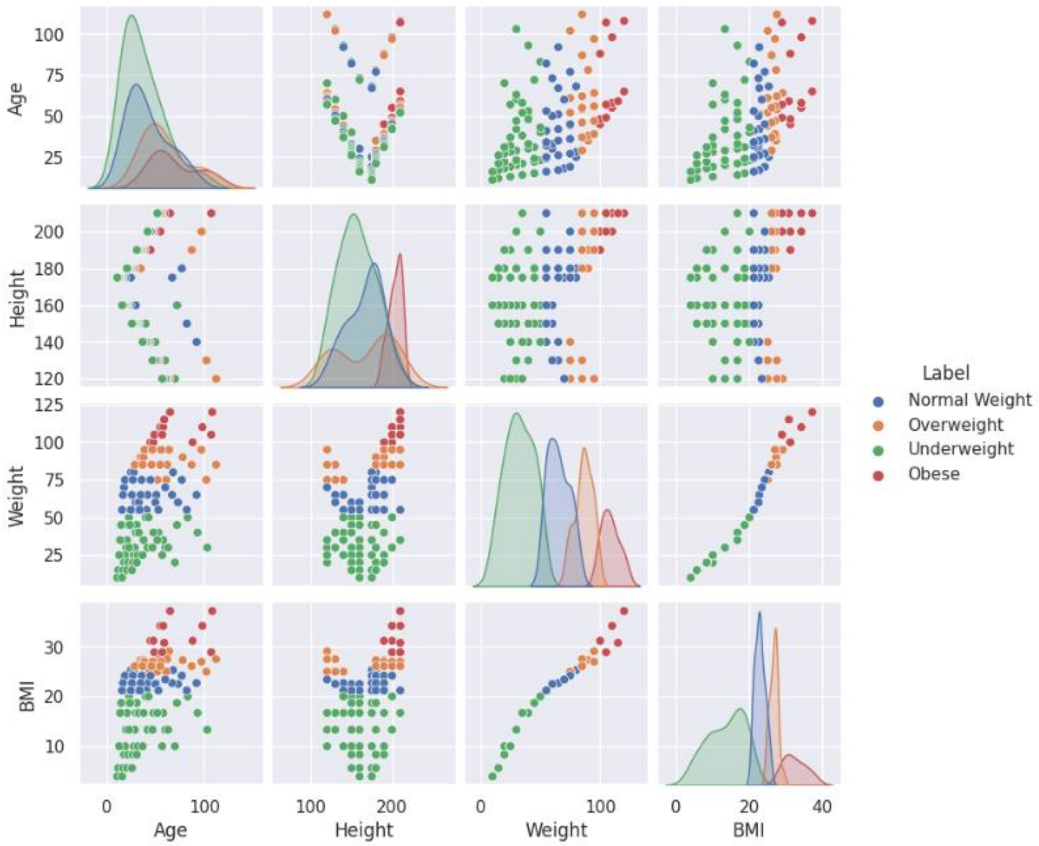
	ID	Age	Gender	Height	Weight	BMI	Label
0	1	25	Male	175	80	25.3	Normal Weight
1	2	30	Female	160	60	22.5	Normal Weight
2	3	35	Male	180	90	27.3	Overweight
3	4	40	Female	150	50	20.0	Underweight
4	5	45	Male	190	100	31.2	Obese

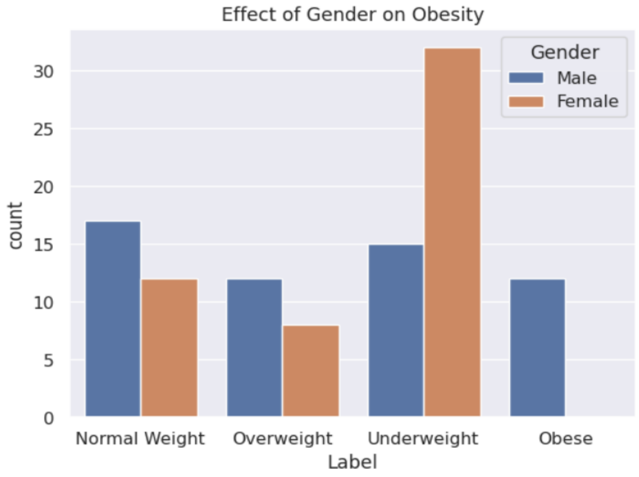
Data exploration

Before the analysis, the following data cleaning steps were taken:

- ✓ Removal of null values (there were none)
- ✓ Removal of outliers (there were none, verified by histograms)

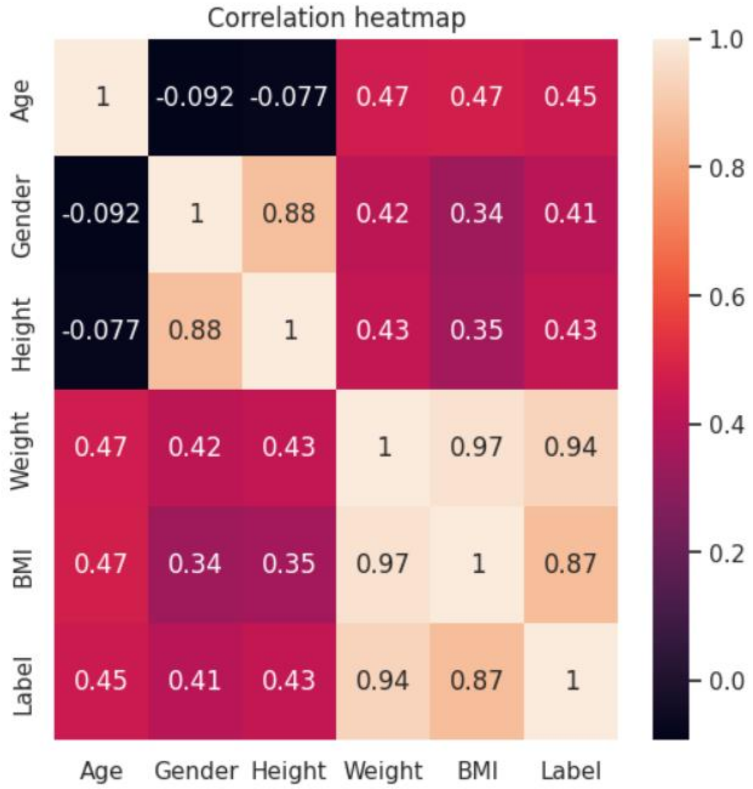
After cleaning the data, the dataset was picturized with scatter plots, histograms, and bar plots to understand the relationships between the features of the dataset and how each feature affected the obesity of an individual.





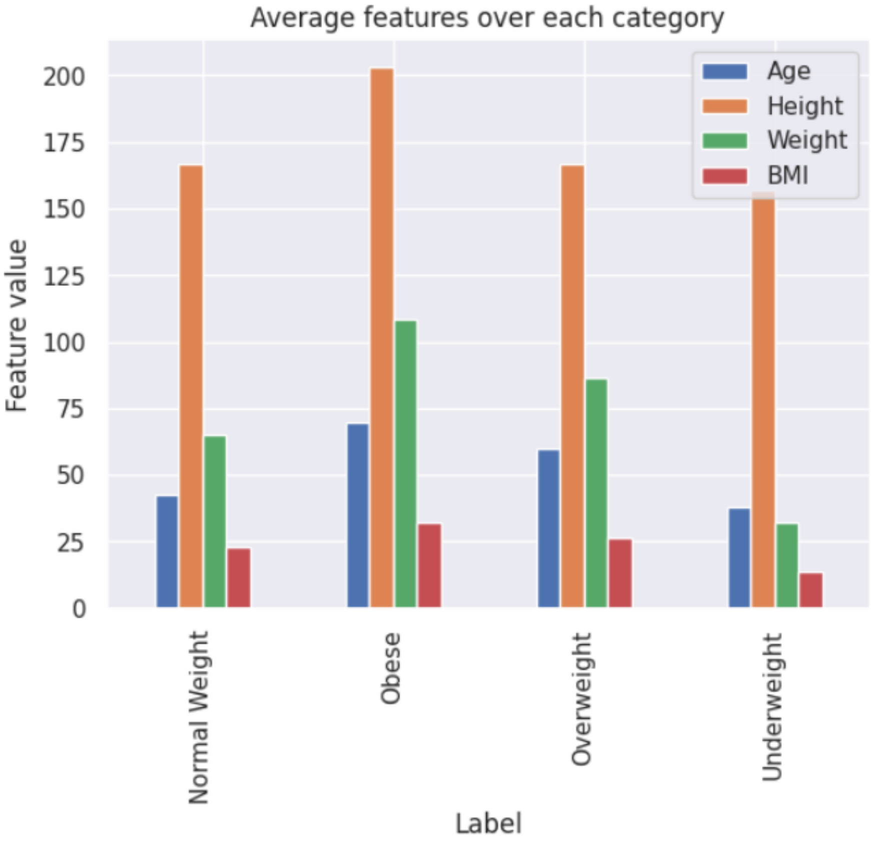
As shown by the pair plot above, **Weight** and **BMI** seem to be the driving factors of obesity.

The graph to the left shows that men are more likely to be obese and overweight, while women are more likely to be underweight.



The heatmap to the left shows the correlations of each feature. Evidently, Weight and BMI have extremely high correlations with the target variable (Label).

This suggests that they would be useful predictors of obesity.



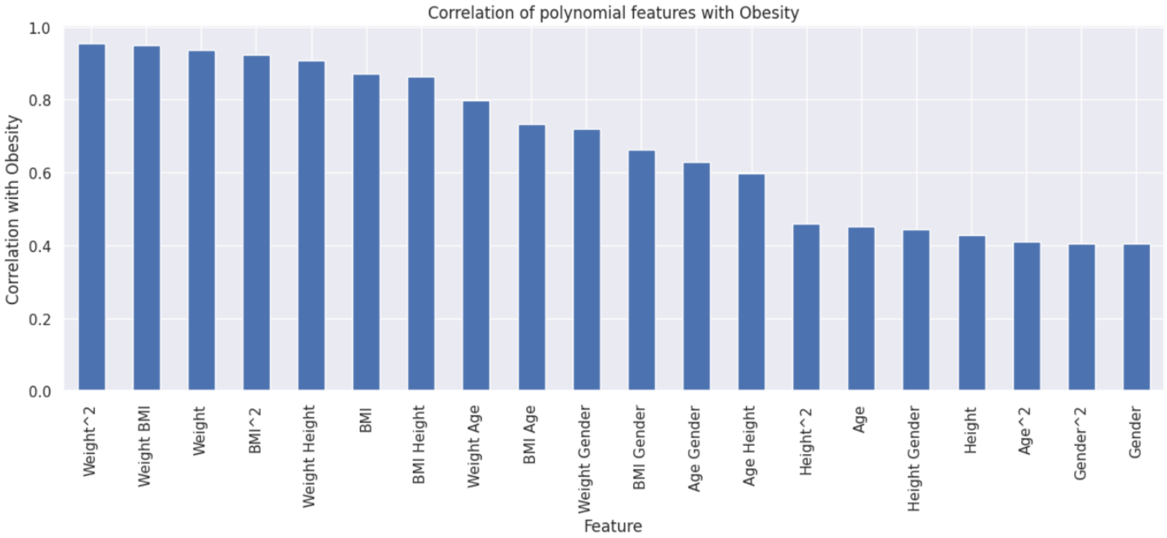
From the chart to the left, it can be observed that Age, Height, Weight, and BMI, are on average, higher form obese individuals.

This suggests that older individuals tend to have a higher chance of being obese.

Feature Engineering

To derive further information from the dataset, **polynomial features** along with interactions were used to understand the relationships between the features, and how these relationships affected obesity.

The above chart shows different interactions between the features, and the correlations between these interactions and the target variable (obesity).



Models

Following this analysis, 3 models were fitted to attempt to predict whether an individual is obese given their Age, Gender, Weight, Height and BMI.

Before training each model, the data was scaled using Standardization.

Each model was fitted using cross-validation (5 folds). Grid Search was used for the Decision Tree and Random Forest to find the optimal hyperparameters.

Model 1: Logistic Regression

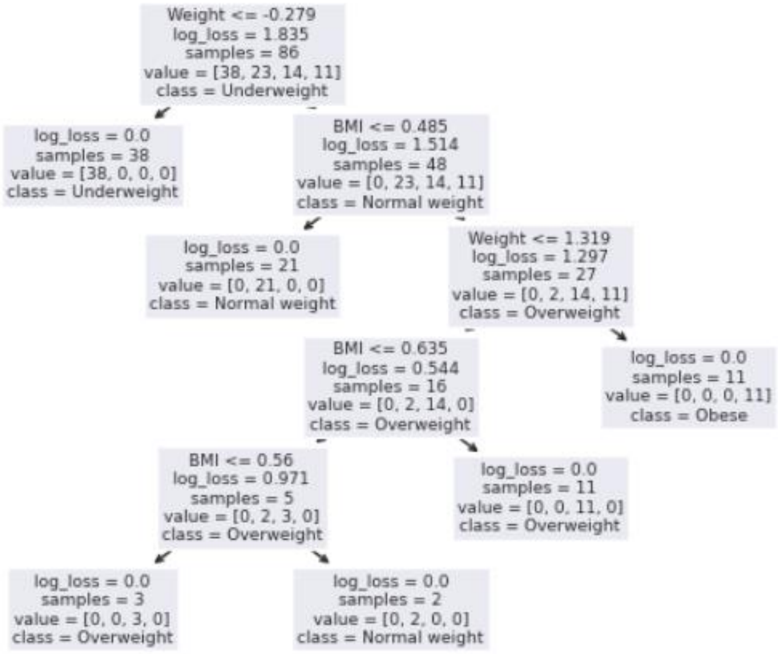
Logistic regression outputs the probability the observation belongs to each class. The class with the highest probability is the class that is predicted.

Performance was astounding: Logistic Regression achieved perfect precision, recall and accuracy on the test set. This is most likely due to the extremely strong correlations between the features and the target variable.

Model 2: Decision Tree

Decision trees use Boolean logic (similar to how humans think) to make decisions based on features.

Performance was also perfect for the Decision Tree model. One advantage, however, is that the Decision Tree offers higher interpretability than the Logistic Regression, as shown by the diagram below:



Note that the values are scaled, which is why you see “Weight <= -0.279”.

Model 3: Random Forest

Random Forest uses an ensemble of Decision Tree Classifiers in an attempt to reduce overfitting and improve accuracy.

Since the Decision Tree Classifier achieved perfect performance, it’s no surprise that the Random Forest also achieved perfect performance. However, just like Logistic Regression, it loses on interpretability as there are too many trees to classify.

Recommendation

Each model achieved perfect performance, so they all win in that respect. However, the Decision Tree offers increased interpretability, so that model is recommended. Not only does it accurately predict whether a patient is obese, but it also enlightens the medical practitioner with the reasoning behind that prediction.

Summary of Key Findings

- Data Quality: The dataset was clean, with no missing values or outliers, ensuring the reliability of the analysis.
- Gender Disparities: Men were more likely to be obese or overweight, while women were more likely to be underweight, indicating a gender disparity in obesity rates.
- Weight and BMI: Weight and Body Mass Index (BMI) had strong correlations with obesity levels, making them significant predictors of obesity.
- Age Influence: The analysis suggested that older individuals tend to have a higher rate of obesity than younger people.
- Feature Engineering: Polynomial features and interactions were used to explore relationships between features and their impact on obesity, enhancing the understanding of feature interactions.
- Logistic Regression: The Logistic Regression model achieved outstanding performance, suggesting that the features had strong predictive power for obesity classification.
- Decision Tree: The Decision Tree model also performed exceptionally well and offered high interpretability, making it a valuable tool for understanding the factors contributing to obesity.
- Random Forest: The Random Forest model, an ensemble of Decision Trees, also achieved perfect performance but lacked the interpretability of a single Decision Tree.
- Model Consistency: All three models (Logistic Regression, Decision Tree, and Random Forest) achieved perfect performance, indicating a robust and reliable prediction for obesity classification.
- Recommendation: The Decision Tree model was recommended due to its interpretability, making it a valuable tool for both accurate predictions and explaining the reasoning behind those predictions to medical practitioners.

Possible Improvements

- Data Collection: Collect more information on various natural factors that may affect obesity such as genetic and environmental information.
- Model Evaluation Metrics: While the models achieved perfect performance, more diverse evaluation metrics such as ROC curves and AUC could have been used to further assess model performance.
- Imbalanced Classes: Investigation into class imbalance and using techniques such as oversampling and undersampling.
- Interpretability: Utilize techniques like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) to provide more detailed insights into why models made certain predictions. This could be especially useful for Random Forest.
- Domain expertise: Collaboration with health professionals can allow a richer analysis of the factors affecting obesity and thus more informed decisions can be made regarding the predictions of the models.