8 September 2025

Faculty of Science

Health Analytics

Decision tree

Molleth Ndlovu - 1718251

1. Introduction………………………………………….1
2. Aim…………………………………………………..1
3. Method
   1. Data description……………………………….…1
   2. Tools…………………………………………..….2
   3. Model justification……………………….............2
   4. Procedural steps for analysis……………………..2
4. Results
   1. Descriptive statistics……………………………...2
   2. Correlation analysis…………………………...….3
   3. Obesity categories distribution.…………………..3
   4. Obesity category by gender..……………………..3
   5. BMI distribution………………………………….4
   6. Model Evaluation………………………………...4
5. Discussion………………………………….………...5
6. Conclusion……………………………………………5
7. References…………………………………………....6
8. **Introduction**

One of the biggest health problems in the world is obesity, which has almost tripled in frequency since 1975 and it is still going up (World Health Organization, 2021). It is closely associated with non-communicable diseases that lower life expectancy and raise healthcare expenses, such as cardiovascular conditions, type 2 diabetes, and several types of cancer (Blüher, 2019). Obesity is a chronic, complex condition defined by an excess or abnormal buildup of body fat that poses a health risk (World Health Organization, 2021). It is typically characterized as having a body mass index (BMI) of 30 kg/m² or greater (Blüher, 2019).

Obesity is caused by a complex interaction of elements such as energy imbalance, genetic predisposition, lifestyle and behaviour, and environmental and socioeconomic factors. Early therapies and policy decisions can be guided by the ability to predict and categorize obesity. Machine learning technologies are becoming increasingly significant in health analytics given that they enable the detection of patterns in big and complicated datasets (Topol, 2019).

Decision tree algorithms, in particular, have gained popularity in clinical prediction due to their ease of understanding and capacity to handle both categorical and numerical data (Song and Ying, 2015). Decision tree is a type of supervised machine learning algorithm used in classification and regression applications. It creates a tree-like structure of decision rules by iteratively dividing the dataset into subsets according to feature values (Song and Ying, 2015). A feature test is represented by each internal node, a test result by each branch, and a final prediction or class by each leaf node. Their interpretability makes them particularly useful in obesity research, where transparency is critical for therapeutic decision-making.

1. **Aim of this report was to:**

* Investigate and define the dataset to better understand its characteristics and distributions.
* Prepare the data adequately for decision tree modelling.
* Develop and support the use of a decision tree classifier to predict obesity categories.
* Assess the model's performance in terms of accuracy, interpretability, and constraints.
* Document methodological decisions and evaluate the results of the analysis.

1. **Method**

**3.1 Data description**

The obesity dataset (obtained from Kaggle) used had no missing values and consisted of 1000 records with seven attributes:

* Numerical – Age (years), height (cm), weight (kg), BMI (kg/m2), physical activity level
* Categorical – Gender (Male or Female), Obesity category (normal weight, overweight and obese)

Descriptive statistics – Mean, standard deviation (std), minimum (min) 25th percentile, median, 75th percentile and maximum (max)

**3.2 Tools used**

* Pandas and NumPy- used for data loading, manipulation, and preprocessing.
* Scikit-learn- used for regression modelling and assessment.
* Matplotlib and Seaborn- used for data visualization.
  1. **Model justification**

Decision tree was used for this report because it can handle both categorical and numerical data types without considerable preparation (Song and Ying, 2015). It also provides interpretable structures, which is critical in health-care applications as transparency promotes decision making based on evidence (Lundberg et al., 2020). Finally, it enables feature importance analysis to help find characteristics that are more strongly connected with obesity.

* 1. **Procedural steps for Analysis**
     1. Import libraries and load data

-libraries were imported and heart attack dataset(csv) was loaded into google colab (Python 3)

* + 1. Exploratory data analysis (EDA)

-Illustrated descriptive statistics, correlation analysis, and class distributions (Obesity categories distribution, obesity category by gender and BMI distribution)

* + 1. Data Preprocessing

-Encoded categorical variables (such as gender and obesity category).

-Divided the dataset into training (20%) and test (80%) subgroups.

* + 1. Model Training

-Trained a decision tree classifier on the training data.

-Optimized hyperparameters (e.g., tree depth, splitting criterion)

* + 1. Evaluation

-Evaluated predictive performance using the confusion matrix, and classification report (accuracy, recall, precision and F1- score).

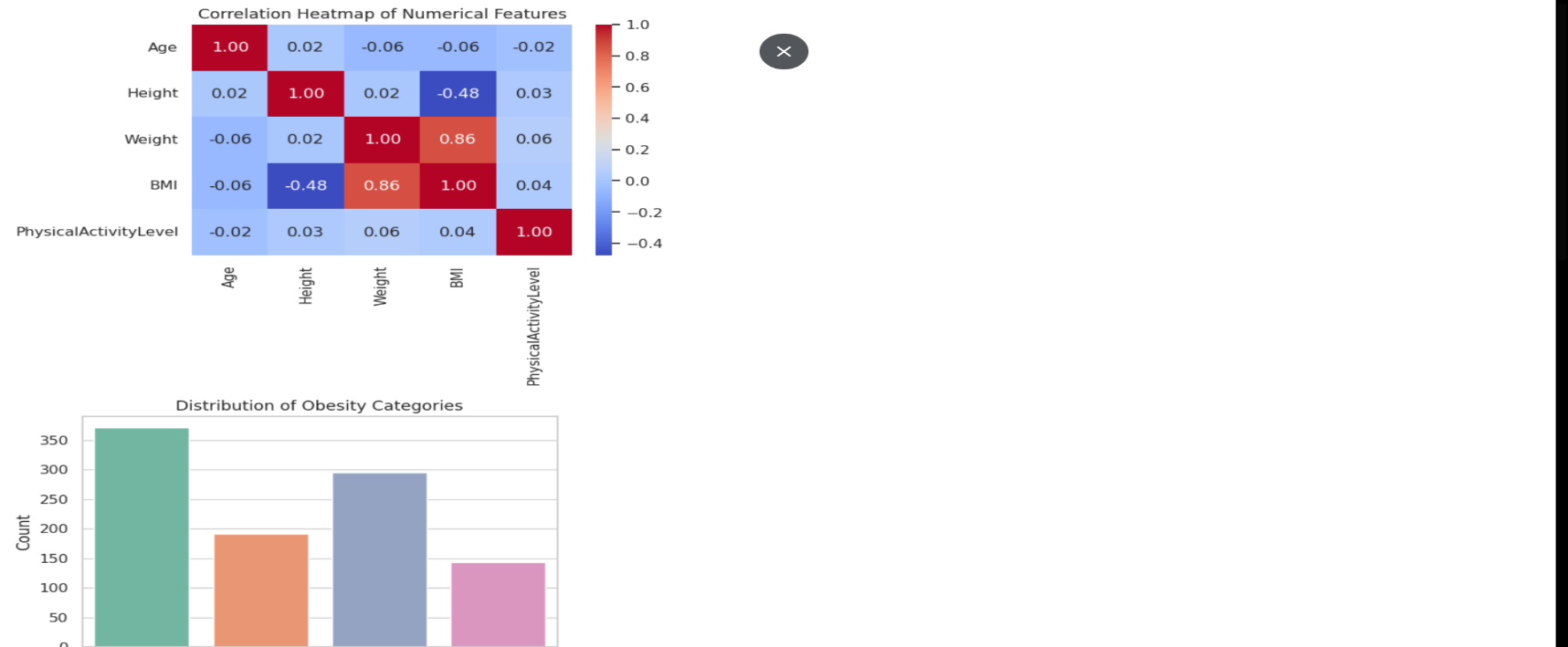
**4. Results**

**4.1 Table 1. Descriptive statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Age | Height | Weight | BMI | PhysicalAL |
| Count | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 | 1000.000000 |
| Mean | 49.8570000 | 170.052417 | 71.205769 | 24.888317 | 2.534000 |
| Std | 18.114267 | 10.309971 | 15.509849 | 6.193917 | 1.116284 |
| Min | 18.000000 | 136.115719 | 26.065730 | 8.470572 | 1.000000 |
| 25% | 35.000000 | 163.514205 | 61.129629 | 20.918068 | 2.000000 |
| Median | 50.000000 | 169.801665 | 71.929072 | 24.698647 | 3.000000 |
| 75% | 66.000000 | 177.353596 | 81.133746 | 28.732132 | 4.000000 |
| Max | 79.000000 | 201.419670 | 118.907366 | 50.791898 | 4.000000 |

Description: The mean age was in the late 40s. While BMI readings ranged across normal, overweight, and obese categories, height and weight followed normal distributions . Physical activity levels varied across individuals.

**4.2 Figure 1. Correlation Analysis**

****

Description: BMI was strongly association with weight (0.86) and a negative link with height (-0.48).

**4.3 Figure 2. Obesity categories distribution**

****

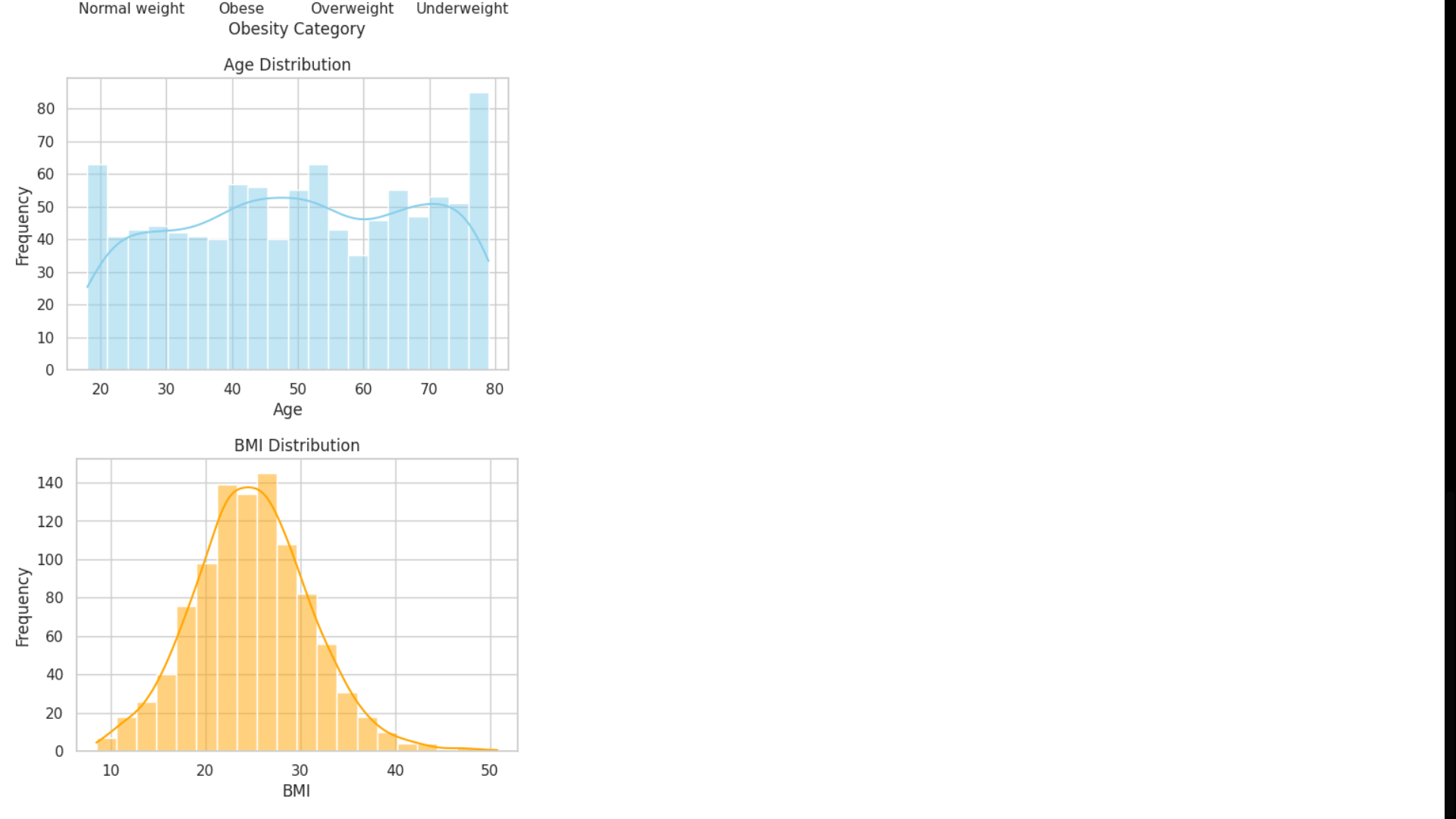
Description: The dataset included underweight (n=140), normal weight (n=370), overweight (n=290), and obese (n=180). This suggested a nicely balanced sample with representation from all classes.

**4.4 Figure 3. Obesity category by Gender**

****

Description: Males and females were evenly distributed throughout all categories.

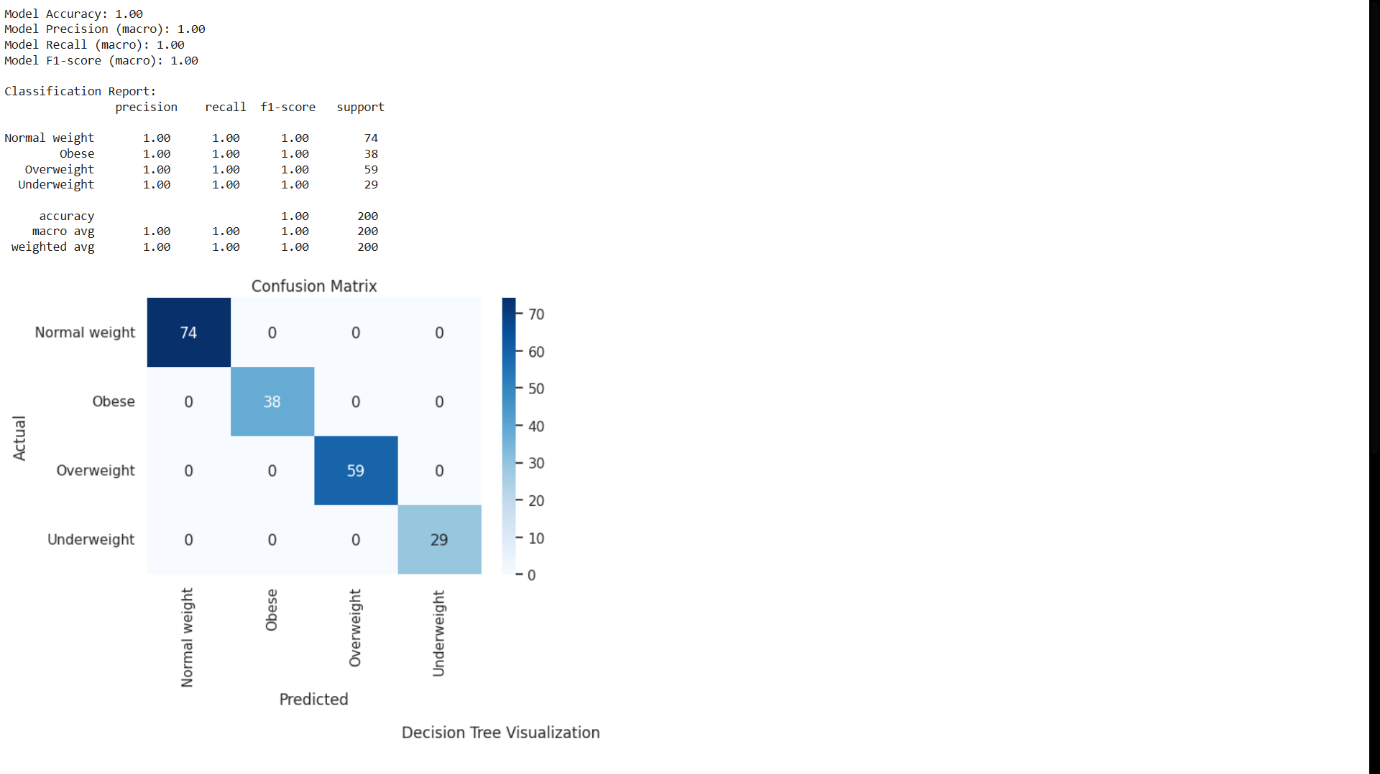
**4.5 Figure 4. BMI distribution**

****

Description: Values were clustered around standard cut-offs, indicating a clear difference between normal, underweight, overweight, and obese categories.

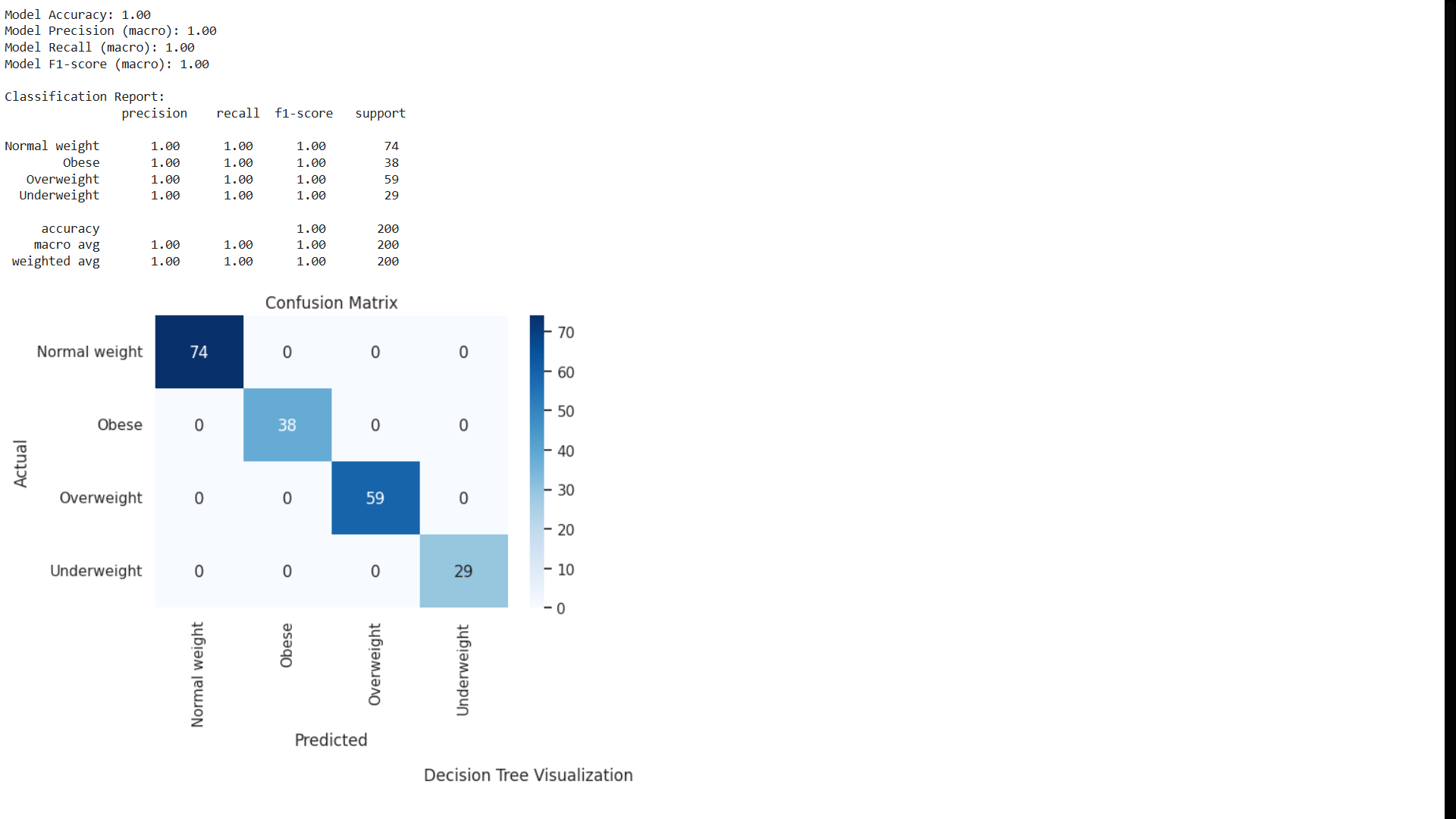
**4.6 Model evaluation**

Figure 5. Confusion matrix



Description: The confusion matrix revealed that all cases were properly classified, with no misclassifications.

Figure 6. Classification report



Description: The classification report revealed precision, recall, and F1-scores of 1.00 for all four obesity categories (normal weight, underweight, overweight, and obese)

**5. Discussion**

The findings emphasize the effectiveness of decision tree algorithms in obesity classification. The classification report showed accuracy, precision, recall, and F1-scores of 1.00 in all four categories (normal weight, underweight, overweight, and obese). While this suggests perfect performance with no misclassifications, such results are extremely exceptional in real-world health datasets, implying that the model is overfitting.

The model's flawless performance is most likely attributed to the inclusion of BMI, which directly determines the obese categories. This allowed the decision tree to split groups based on defined BMI thresholds, resulting in perfect classification. While this indicates the model's capacity to learn from strong predictors, it also reveals its limited generalizability. In more complicated datasets where obesity classification depends on a broader variety of variables—such as food patterns, genetic predisposition, or socioeconomic factors—the model may not reach comparable accuracy.

Decision trees persist to be extremely helpful because of their interpretability. The visualizations revealed that the most influential parameters were BMI, weight, and physical activity level, which is congruent with clinical knowledge of obesity factors. This transparency makes decision trees appropriate for healthcare analytics, as model decisions must be explicable to physicians and policymakers (Song and Ying, 2015; Lundberg et al., 2020).

Nonetheless, the ideal performance metrics require caution in interpretation. In practice, models should be assessed on more difficult datasets or with stronger constraints, such as omitting BMI as a predictor. Furthermore, strategies such as tree pruning could simplify the model and prevent overfitting, whilst collaborative approaches such as random forests would improve stability and accuracy by averaging across numerous trees (Fernández-Delgado et al., 2019; Biau & Scornet, 2016). These approaches are likely to produce more realistic performance while maintaining interpretability.

**6. Conclusion**

This study showed how to predict obesity categories using anthropometric, behavioural, and demographic characteristics using a decision tree classifier. Since BMI was included as a predictor, the model produced flawless results. However, these results show the risk of overfitting and low generalizability, even though it also emphasizes the significance of BMI in the definition of obesity.

Decision trees are often suitable for clear and understandable classification in health analytics. However, Future research should test the model on more sophisticated datasets and incorporate ensemble techniques such as random forests or pruning to guarantee resilience and wider applicability in real-world obesity prediction.

**References**

Biau, G. and Scornet, E. (2016). A random forest guided tour. Statistics Surveys, 2, pp. 145–195.

Blüher, M. (2019). Obesity: global epidemiology and pathogenesis. Nature Reviews Endocrinology, 15(5), pp.288–298.

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2019). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 20(55), pp.1–55.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), pp.56–67.

Song, Y.Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry, 27(2), pp.130–135.

Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), pp.44–56.

World Health Organization (2021). Obesity and overweight. [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> [Accessed 1 September 2025].