Comparison of Random Forest and Decision Tree for Predicting Obesity Levels Based on Lifestyle Factors

1. Introduction/Problem Statement

Obesity is a global health concern influenced by a numerous dietary and physical habits. Our main objective in this small research is to analyze a dataset with total 16 different factors and train two different ML methods for predicting the obesity based on those. Decision Trees and Random Forest were used for predicting the classification on the UCI archive Obesity dataset[3]

2. Data set overview and manipulation

2.1 Data exploration

- This dataset includes data for estimating obesity levels in individuals from Mexico, Peru, and Colombia.
- The data is based on eating habits and physical condition with 16 attributes, 1 target column and 2111 different records, with no missing values.
- Target column is classified to 7 different outcomes (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III).
- 23% of the data was collected directly from users via a web platform, while the rest of it was synthetically generated using the Weka tool and the SMOTE filter.

Feature	Additional Feature Description	Data Type	Range/Categories
Gender		Categorical	Female, Male
Age		Numeric	Min: 14.0, Max: 61.0
Height		Numeric	Min: 1.45, Max: 1.98
Weight		Numeric	Min: 39.0, Max: 173.0
FHWO	Family history with overweight	Categorical	no,yes
FAVC	Frequent high caloric food intake	Categorical	no, yes
FCVC	Frequence of vegetables intake	Numeric	Min: 1.0, Max: 3.0
NCP	Daily meals	Numeric	Min: 1.0, Max: 4.0
CAEC	Food between meals	Categorical	Sometimes, Frequently, Always, no
SMOKE		Categorical	no, yes
CH2O	Daily water intake	Numeric	Min: 1.0, Max: 3.0
SCC	Calories Monitoring	Categorical	no, yes
FAF	Frequence of physical activity	Numeric	Min: 0.0, Max: 3.0
TUE	Frequence of technology usage	Numeric	Min: 0.0, Max: 2.0
CALC	Alcohol intake	Categorical	no, Sometimes, Frequently, Always
MTRANS	Most used transportation	Categorical	Public_Transportation, Walking, Automobile, Motorbike, Bike

Figure 1: Brief description of dataset features

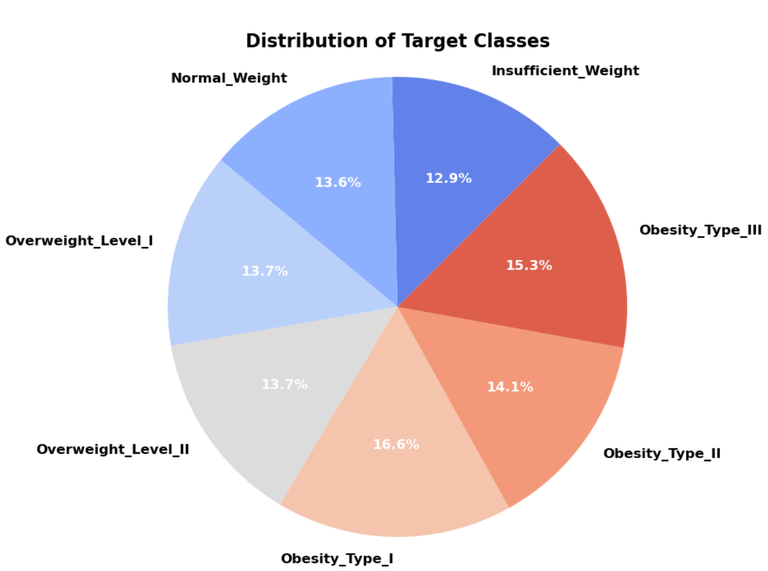


Figure 2: Distribution of target column classes

- Features table shows that we have a split of 8 categorical and 8 numerical variables
- The classes of the target variable are equally distributed across the dataset, mainly due to the SMOTE technique that was applied.
- To be able to apply SMOTE to the dataset, missing values and outliers were handled before that as explained in the paper[4]

2.2 Data Preprocessing

- Categorical features converted to numeric for uniformity and normalized to range [0,1]
- Plotted the distribution of obesity classes to assess balance.
- Derived a BMI feature to explore its correlation with obesity levels.
- Found overlapping BMI ranges across classes, making it unsuitable for classification.
- Considered k-means and mixture models to refine BMI ranges but opted for simpler approaches due to time constraints. So opted for Decision Trees over Logistic Regression for categorical variables and robust classification capabilities.[6]

3. Models Selected

3.1 Random Forest

Description: Random Forest is an ensemble learning method that builds multiple decision trees using bagging (bootstrap aggregation). It combines predictions from these trees to achieve higher accuracy and generalization.

Pros:

- Handles non-linear relationships effectively and performs well on imbalanced data.
- Reduces overfitting by aggregating multiple trees yielding robust generalization
- Handles missing data without significant loss of performance.

Cons:

- Computationally expensive due to ensemble training.
- Training time is significantly longer compared to single models like Decision Tree.
- Less interpretable compared to individual decision trees.

3.2 Decision Trees

Description: Decision Tree is a simple, interpretable model that splits data into subsets based on feature thresholds. It builds a tree-like structure where each leaf node represents a class label.

Pros:

- Easy to interpret and explain, results are transparent.
- Fast to train and computationally inexpensive, suitable for real-time applications
- Requires minimal data preprocessing and works well with categorical variables.

Cons:

- Prone to overfitting (deep trees), sensitive to small changes in data (high variance).
- Has lower accuracy compared to ensemble methods like Random Forest.
- Struggles with imbalanced datasets and complex relationships.

4. Hypothesis

Given the balanced nature of the dataset and the categorical target variable, the difference in accuracy between Random Forest and Decision Tree models is expected to be moderate. Random Forest's ensemble approach should provide better generalization and slightly higher accuracy by reducing overfitting, while Decision Tree, being a simpler model, should still perform well due to its suitability for categorical data and the absence of significant class imbalance.

5. Model Training

Dataset Split: Stratified split into 80% training and 20% testing to preserve class balance.

Grid Search with Cross-Validation:

- Performed grid search with 5-fold cross-validation to tune hyperparameters.
- Selected the model with the smallest average cross-validation error across the folds to ensure better generalization, since we prioritize it over accuracy.[5]

Hyperparameter Tuning and experimental results:

- Random Forest (RF): Tuned: NumTrees [250-350], MinLeafSize [1-3], MaxSplits [350-450].
- Decision Tree (DT): Tuned: MinLeafSize [1-5], MaxSplits [50-150].
- We began by exploring a broad range of distinct hyperparameter values to generate intermediate results. The initial tuning gaps were intentionally large to quickly identify promising ranges, ensuring efficient use of computation time. Based on these results, we refined the ranges to the ones above and increased the granularity of the values tested for more precise tuning. The optimal hyperparameters were then identified, saved, and exported for use in the final model.

Best Hyperparameters:

- **RF**: NumTrees = 350, MinLeafSize = 1, MaxSplits = 400
- **DT**: MinLeafSize = 2, MaxSplits = 125

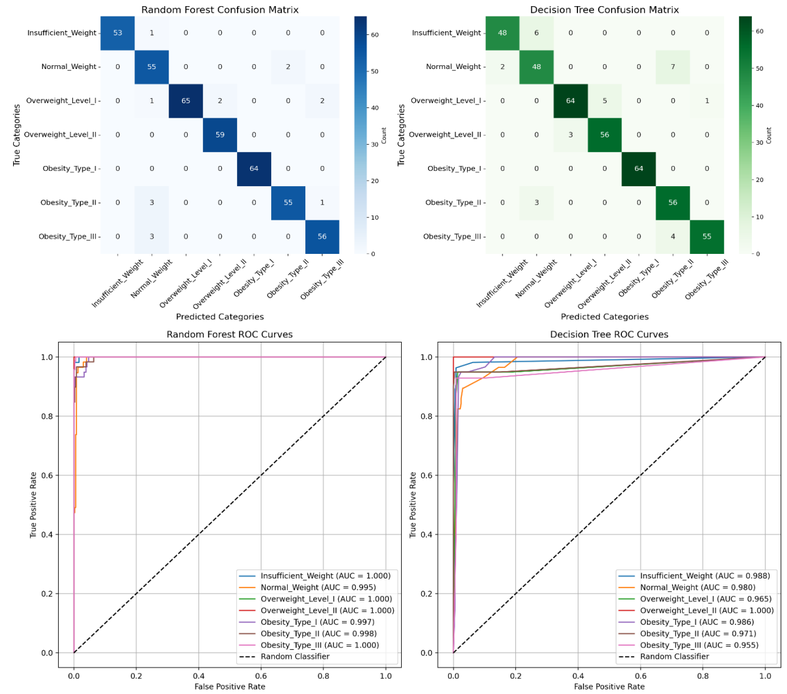
Training Process:

- **RF**: Ensemble method reducing overfitting, suitable for complex relationships (Breiman)[1].
- **DT**: Fast, interpretable, but prone to overfitting (MathWorks Documentation)[2].

6. Evaluation Metrics and Visuals

RF: Accuracy: 96.45%, Precision: 0.965, Recall: 0.965, F1-Score: 0.964, Training Time (s): 1.438

DT: Accuracy: 92.65%, Precision: 0.928, Recall: 0.925, F1-Score: 0.925, Training Time (s): 0.011



Visuals Interpretation:

The confusion matrices reveal that the Random Forest (RF) model has fewer misclassifications compared to the Decision Tree (DT), particularly in distinguishing between similar obesity classes. The ROC curves further illustrate this performance gap, with RF demonstrating a higher Area Under the Curve (AUC), indicating better overall classification performance. RF's superior recall is critical for minimizing false negatives in health predictions, while DT's faster computation makes it suitable for real-time applications despite slightly lower accuracy.

7. Results Analysis and Critical Evaluation

- The analysis of the obesity dataset using Decision Trees (DT) and Random Forest (RF) highlighted the strengths and limitations of each model. RF outperformed DT with an accuracy of 96.45% compared to 92.65%, and demonstrated higher precision, recall, and F1-score values. RF's ensemble nature allows it to capture complex relationships, making it more accurate and generalized. However, its increased computational cost and reduced interpretability are important trade-offs when model transparency and efficiency are required.
- The Decision Tree, while slightly less accurate, provides faster results and is easier to interpret, making it ideal for real-time applications where transparency is critical. However, it is more prone to overfitting and struggles with complex relationships between features.
- In practice, the choice between models depends on the application context. If accuracy and generalization are paramount, Random Forest would be the better option, but if computational efficiency and interpretability are prioritized, Decision Tree is a suitable choice. These results highlight the trade-off between model complexity and performance, with Random Forest offering better overall classification performance, while Decision Tree remains a viable option for faster, interpretable predictions.

8. Lessons Learned

This analysis reinforced that Decision Trees (DT) is a strong choice when working with a balanced and tidy dataset, especially when time efficiency is a priority. Given the simplicity and speed of Decision Trees, they can provide valuable results without the need for extensive computational resources. However, when seeking higher accuracy and generalization, Random Forest outperforms DT. This reinforces the importance of carefully selecting the model based on the problem's needs, balancing between speed and predictive power.

9. Future Directions

For future work, it would be beneficial to collect additional data, such as body fat and muscle percentage, and consider dropping less influential features to improve accuracy and reduce model complexity. A promising approach would be to use k-means clustering combined with Gaussian Mixture Models (GMM) to define appropriate BMI ranges for each target class. This would allow BMI, a continuous variable to be treated as target, making it more suitable for Logistic Regression and potentially improving model performance. By refining the target variable and simplifying the model, we could enhance both accuracy and interpretability.

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
2. MathWorks Documentation: TreeBagger and fitctree Functions.
3. "Estimation of Obesity Levels Based On Eating Habits and Physical Condition ," UCI Machine Learning Repository, 2019. [Online]. Available: https://doi.org/10.24432/C5H31Z.
4. F. M. Palechor and A. de la H. Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," Data in Brief, vol. 25, p. 104344, Aug. 2019, doi: https://doi.org/10.1016/j.dib.2019.104344..
5. M. Kirk, Thoughtful machine learning : a test-driven approach. Beijing: O'reilly Media, 2014.p.10-11.
6. C. Molnar, Interpretable machine learning : a guide for making Black Box Models interpretable. Morisville, North Carolina: Lulu, 2019.p.102.