# Executive Summary– Build and Evaluate Tree Models and Participate in the Multi-Class Prediction of Obesity Risk Kaggle Competition (Late Submission).

By Samuel Mbah Nde – DDS 8555 – Assignment 6

## Project Overview.

This week, I performed multiclass classification with tree models and answered conceptual questions 1 and 7 from chapter 8 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023). This executive summary is divided into 2 sections. In section 1, I present the solutions to the conceptual questions and in section 2, I implement 4 multiclass classification tree model models and submit the most performant model in the Multi-Class Prediction of Obesity Risk Kaggle (Reade & Chow, 2024). The code used for this analysis can be found in my GitHub repository https://github.com/ndesamuelmbah/DDS-8555/tree/main/AssignmentFiles/Week6 (Nde, 2024).

**Section 1: Conceptual Questions.**

Question 1: This is question 1 extracted from page 363 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.

A close-up of a text

AI-generated content may be incorrect.  
**Solution**

I created an imaginary scenario using a decision tree to predict house prices based on the number of bedrooms and the year built. Houses built before 2005 are split into two groups: those built before 1978 and those built in or after 2014. Homes built before 1978 are priced between $160,000 and $220,000 depending on whether they have fewer than 4 rooms or 5 or more. Similarly, homes built in or after 2014 are priced between $200,000 and $300,000, with higher prices for houses with more rooms. The tree and regions used in my problem are below.  
**Figure 1**

Imaginary Decision Tree and Boundary Diagram for using prices Rooms and Year to predict Price.  
A diagram of a tree

AI-generated content may be incorrect. A graph with different colored lines

AI-generated content may be incorrect.

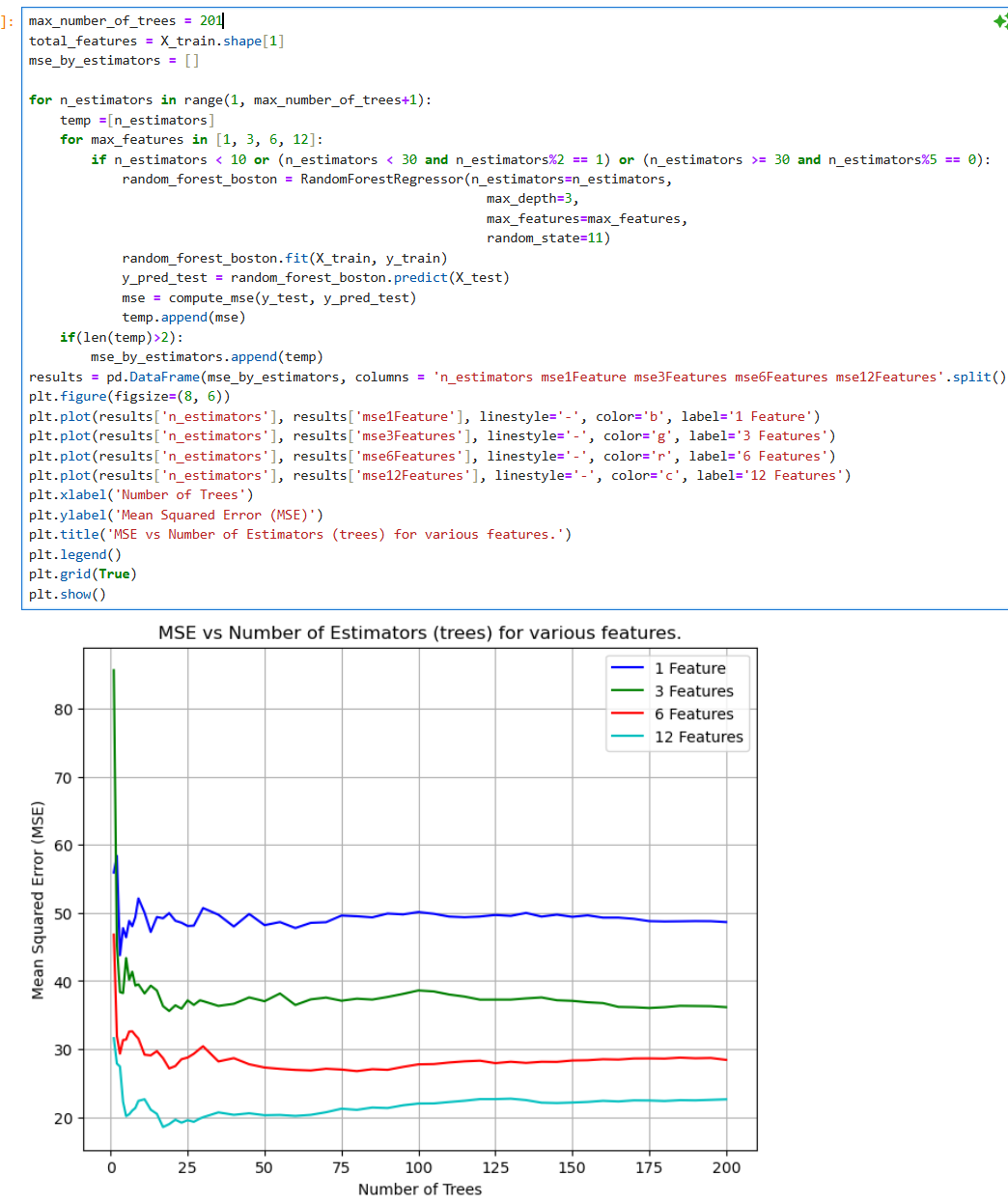
Note: Imaginary Decision Tree (left) and Boundary Diagram (right) for predicting house prices using the number of rooms and year built. The decision tree splits houses based on the year built and number of rooms to estimate prices. The boundary lines in dashed lines identify regions labeled 1 to 8. The regions correspond to the leaves of the decision tree for example, R1 represent less 3 or fewer bedroom homes built before 1978. This imaginary model would predict prices of such homes as $160,000.

Question 2: This is question 13 extracted from page 196 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below. A close-up of a text

AI-generated content may be incorrect.

**Solution**

To provide test errors for random forest regressor fitted on different parameters sets on the dataset, I split my data into test and training sets and then looped over different parameter sets as shown in Figure 2 below, computing and storing MSE for every pair of parameters. I implemented a filter to skip over certain values in order to improve the speed of the job.

**Figure 2.**  
Test mean square error on Test dataset for various combinations of number of trees used to train random forest regressor on the training dataset.  
**Note**: *As the number of features increases, the test set's mean squared error (MSE) declines quickly to its minimum, then slightly rises, suggesting overfitting. Models with more features yield lower MSE, indicating most features are important. The model with one feature shows unstable performance, lacking enough information to capture the trend.*

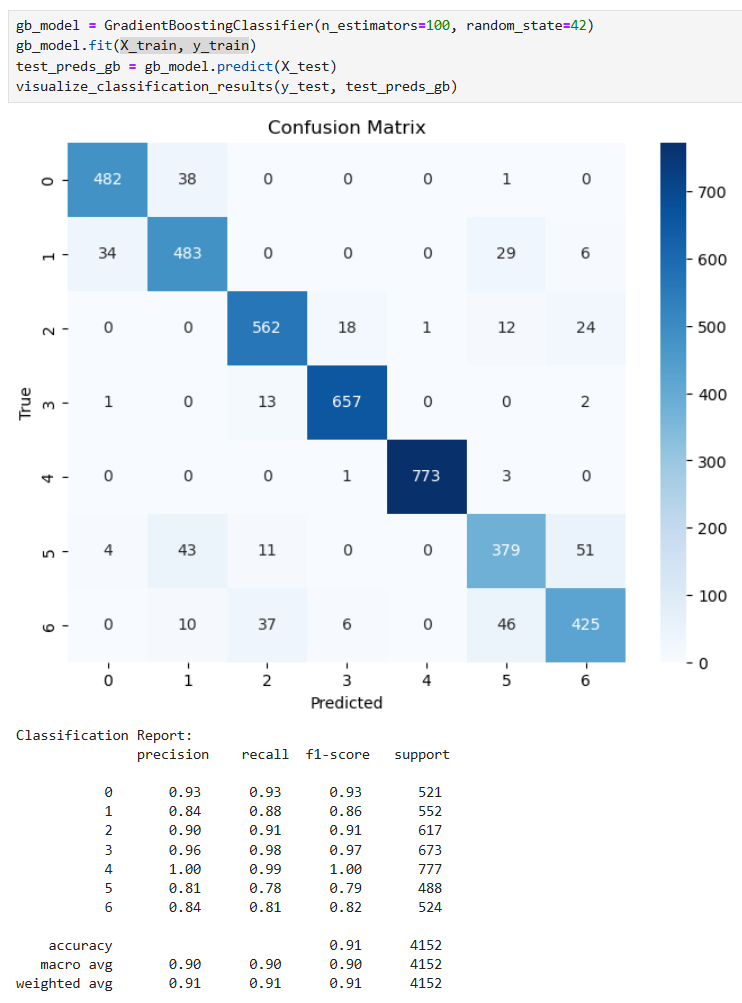
**Section 2.** Participation in Multi-Class Prediction of Obesity Risk Kaggle competition(Reade & Chow, 2024)**.**

I trained four tree-based classification models: Decision Tree, Random Forest, Bagging, and Gradient Boosting classifiers using all dataset variables. Categorical variables were label-encoded, with the unseen CALC category Always in the test set merged into Frequently due to their synonymous meanings (Mumuni & Mumuni, 2025). Unlike last week when I scaled data because I was building models like KNNs and SVMs that rely on distances between points, I did not scale the data since tree models are not sensitive to feature scales (Ahsan, Mahmud, Saha, Gupta, & Siddique, 2021). The dataset was split into 80% training and 20% testing. The best model, Gradient Boosting, achieved a weighted recall score of 91%. The reason I decided to use a weighted recall score for evaluating my model is because there was data imbalance and using a weighted score would take that into consideration. The screenshot in Figure 3 shows proof taking part in the Kaggle competition while the confusion matrix and classification report of the best model are displayed in the snippet in Figure 4.

**Figure 3**  
Proof of participation in the Kaggle competition. A computer screen shot of a computer

AI-generated content may be incorrect.

**Figure 4**  
Classification report for Gradient boosting classifier on the test dataset.

  
**Note:** The Gradient Boosting model achieves an overall macro weighted recall of 90%, with the highest recall (99%) for class 4. Lower performance is observed in class 5, indicating that further tuning might improve the model's performance on this class.

Finally, to understand how the model got its predictions, I generated the feature importance plot from the Gradient Boosted Classifier. The plot indicates the relative contribution of each feature to the model's predictions. The most influential feature is **Weight**, which has the highest importance score, significantly larger than all other features. This suggests that weight is the primary determinant in the classification task. Other features such as **FCVC (Frequency of Consumption of Vegetables), Gender, and Height** also contribute meaningfully, but to a lesser extent. Features like **Age, CH2O (Daily Water Intake), and CALC (Alcohol Consumption)** have minor influence, while several others, including **family history with overweight, MTRANS (Mode of Transportation), and SMOKE (Smoking Habit)**, have negligible impact on the model's predictions. This insight can guide feature selection and further model optimization by potentially removing low-impact features.

**Figure 5**  
Feature Importance plot from Gradient Boosted Classifier – The best tree model.

A graph with blue bars

AI-generated content may be incorrect.

# Bibliography

Ahsan, M. M., Mahmud, M. A., Saha, P., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies, 9*(52). doi:10.3390/technologies9030052

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python (1st ed.).* Springer Nature.

Mumuni, A., & Mumuni, F. (2025). Automated data processing and feature engineering for deep learning and big data applications: A survey. *Journal of Information and Intelligence, 3*(2), 113-153. doi:10.1016/j.jiixd.2024.01.002

Nde, S. M. (2024, 02). *DDS-8555*. Retrieved from Github: https://github.com/ndesamuelmbah/DDS-8555.git

Reade, W., & Chow, A. (2024). *Multi-Class Prediction of Obesity Risk*. Retrieved from Kaggle: https://kaggle.com/competitions/playground-series-s4e2