| **DATA 430 Technical Report Assignment 3: Decision Trees** | **<enter student name here>** |
| --- | --- |
| **<enter descriptive project title here>** | |
| **URL to dataset:** | |

| **Overview** |
| --- |
| **Problem Domain**: |
| With an increased focus on physical fitness as a measure of health, professionals are exploring automated classification methods for physical performance levels. These methods are critical in supporting a proactive approach to health and wellness, allowing for personalized assessments that go beyond traditional one-size-fits-all fitness recommendations. This project aims to develop a robust classification model that will assist healthcare providers, fitness trainers, and wellness coaches in evaluating and predicting fitness classes based on individual performance metrics.By accurately categorizing individuals into different fitness levels, this model can provide actionable insights that enhance personalized training programs and preventive care strategies. For healthcare providers, such insights are invaluable for identifying early signs of potential health issues, tracking progress, and tailoring interventions based on a person's unique physical capabilities. For trainers and coaches, the model serves as a tool to optimize workout plans, reduce injury risks by adjusting intensity levels appropriately, and set realistic goals aligned with each client’s fitness level.Moreover, automated fitness classification supports large-scale fitness and wellness initiatives by enabling remote monitoring and assessment. This can be particularly beneficial for corporate wellness programs, physical therapy practices, and telehealth platforms, where individual assessments may not be feasible in person. The data-driven insights derived from the model can also help in fostering long-term adherence to healthy habits by showing measurable progress and motivating individuals to reach their health goals.In essence, this project seeks to bridge the gap between data science and health by building a scalable solution that can dynamically adjust to individual needs, thus promoting a more data-informed, preventive approach to health and fitness. |
| **Objective**: |
| To classify individuals into performance categories (A-D) based on physical and physiological attributes using a Decision Tree model. This project’s main goals are:   * To identify significant factors influencing fitness levels. * To provide an interpretable decision framework for healthcare practitioners. * To evaluate class prediction accuracy without balancing techniques such as SMOTE. |
| **Analysis** |
| **Exploratory Data Analysis**(EDA) : |
| **Exploratory Analysis**   * **Dataset & Variables:** The dataset includes 13,393 samples, with attributes like age, gender, height, weight, body fat percentage, and fitness metrics such as grip force, sit-ups count, and broad jump distance. Key insights from the exploratory analysis are as follows:   + **Age Distribution:** Primarily between 25-45 years, skewed toward younger participants.   + **Body Fat and Fitness Levels:** Lower body fat correlates with higher performance classes.   + **Performance Metrics Correlation:** Grip strength and broad jump performance correlate positively, hinting at strength as a predictor of higher fitness classes. * **Data Visualizations:**   Visual tools used to analyze data patterns include:   * + **Histograms:** Highlighting class-wise distributions of physical attributes.   + **Heatmaps:** Showing correlations among physical and physiological metrics. * **Summary Analysis:** |
| **Preprocessing**: |
| * **Missing Data:** No missing values were detected. * **Encoding:** Gender was one-hot encoded. * **Scaling:** Continuous variables were standardized. * **Class Balancing:** No balancing techniques, such as SMOTE, were applied |
| **Model Fitting**: |
| * Cross-Validation Scores for each fold: [0.63307204 0.63045913 0.62075401 0.63256161 0.63890963] * Mean Cross-Validation Score: 0.6311512856474206 * Standard Deviation of CV Score: 0.005906820922341248 |
| **Results** |
| **Model Properties:** |
| 1. Tree Structure and Depth:  * **Root Node**: The starting point of the decision tree. * **Depth**: The longest path from the root node to a leaf node. Higher depth generally means a more complex model but can risk overfitting. * **Number of Leaves**: The count of terminal nodes (leaves) that determine the final prediction for each path. |
| **Output Interpretation**: |
| **Evaluation**: employ appropriate metrics to quantitatively evaluate the performance of the fitted model. For supervised classification, this includes simple accuracy, precision & recall (or sensitivity & specificity), all of which can be generated from a confusion matrix, or ROC. |
| **Conclusion** |
| **Summary**: |
| The final output of this decision tree model reflects a structured approach to data classification, breaking down complex decisions into a hierarchy of simpler decisions. If the model achieves high purity in its leaves, it can confidently meet the stated classification objective. However, if class distributions at the leaves remain mixed, additional tuning or feature engineering may be necessary to improve its predictive power. |
| **Limitations & Improvement areas**: discuss the limitations of the analysis and identify potential improvement areas for future work. This could be related to the data, algorithm, or a combination of the two. |
| Here , all the classes have the same height.  **Feature Limitations**: The effectiveness of a decision tree depends on the quality and relevance of features in the dataset. If key predictors are missing, the tree may fail to effectively classify certain observations, reducing model accuracy and interpretability. |

| **Appendix** |
| --- |
| 1. **Confusion Matrix:** The confusion matrix visually represents the performance of the Decision Tree model across different classes (A, B, C, D). It highlights the true positives, false positives, false negatives, and true negatives for each class, showing where the model is most and least accurate. This matrix reveals the challenges in distinguishing Class B and Class C, reflecting areas where model adjustments may be beneficial. 2. **Decision Tree Visualization:** A visualization of the trained Decision Tree model illustrates the decision paths based on features such as age, body fat percentage, and grip strength. Each node represents a split on a particular feature, and the leaves indicate the predicted class. This visualization serves as an interpretative tool, offering healthcare professionals insights into how different metrics influence performance classifications. 3. **Key Code Snippets for Preprocessing and Modeling:** The model-building process involved key preprocessing steps, including encoding categorical variables (e.g., gender), scaling continuous features (e.g., age, weight), and handling class imbalance using SMOTE to ensure balanced representation in the training data. The Decision Tree model was then trained with parameters set to balance interpretability and performance, specifically using a controlled depth to avoid overfitting. These steps ensured that the model remained both interpretable and robust across the dataset.  References  * Breiman, L. (1984). *Classification and Regression Trees.* Routledge. * Quinlan, J. R. (1986). *Induction of Decision Trees.* *Machine Learning*, 1(1), 81-106. * Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825-2830. * Lemaître, G., et al. (2017). *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning.* Journal of Machine Learning Research, 18(1), 559-563. * Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment.* Computing in Science & Engineering, 9(3), 90-95. * Waskom, M. (2021). *Seaborn: Statistical Data Visualization.* Journal of Open Source Software, 6(60), 3021. |