|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 3: Decision Trees** | **Monique Reed** |
| **Should Male and Female Athletes Really Be Held to Different Standards?** | |
| **URL to dataset: https://www.kaggle.com/datasets/kukuroo3/body-performance-data** | |

This template should be used in conjunction with the assignment instructions. The size of the text area below will expand to the length of your response; the area should not be interpreted as a required or suggested length of response. Responses within the text area should be single-spaced with Times New Roman 12pt font. The body of the document will likely be 6-9 pages, not including the Appendix; length may vary depending on the specifics of the analysis and the dataset. As needed, APA format in-text citations should be included, along with a full references list at the end of the document.

|  |
| --- |
| **Overview** |
| **Problem Domain**: give some background and context about the problem domain (application area). For instance, if you are doing the analysis for predicting heart disease, provide some context about the disease and include some interesting statistics about it. Also, discuss how the method is relevant for the chosen problem. |
| The main objective of the decision tree is to determine whether the athlete is a male, or a female based upon the fitness scores. There has always been stated that there are physically limitations and differences between the two genders in sporting. For example, the qualification times for male and female athletes differ in sporting events such as the 2025 NYC marathon. The difference between the inherent femininity and masculinity creeps into many discussion and research articles. But the application of such can be clarified further with a decision tree classifier applied to statistics and measurement of athletes. The model then decides if the athlete is a male or female based upon the statistics that are given.  References:  <https://jensanfilippo17.medium.com/beyond-the-binary-looking-past-gender-stereotypes-in-sports-a7648ea92dce>  https://www.nyrr.org/tcsnycmarathon/runners/marathon-time-qualifiers |
| **Objective**: clearly state the objective of the analysis in relation to the kind of algorithm you are employing. Use specific language as to what question(s) you are trying to answer using the specific analysis/modeling type. |
| The objective of the decision classification model is to determine whether an athlete is either male or female based upon their body statistics and performance measurements, therefore denying or agreeing if the strong distinction between male and female athletes should be supported. |
| **Analysis** |
| **Exploratory Analysis**: describe the data including the source, the collection method, and variables. Perform exploratory analysis. Also, select few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc. |
| The dataset was downloaded from Kaggle, but the actual source of the data was provided from the Korea Sports Promotion Foundation. Some preprocessing was done to visualize some of the data analysis but after that, a heatmap was printed to see the data’s visualization. |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
| Preprocessing for this dataset had to be conducted before a visualized analysis was performed. Two of the columns in the dataset were class values that were strings. These values were passed through a label encoder, although that is referred to as a bad practice by some data scientists, to further classify and correlate the data. But li9ke mentioned earlier, thankfully there were two different genders and three different classes to encode. Anything more would result to a reconsideration of using an encoder.  From there, a model selection was made on the columns that were going to be our features and targets, since our main goal was to determine a gender based upon the performance measurement s of the individual. After these features are selected, a split of the data for training and testing purposes were made. There was a standard 80% of the data placed to the side for training and another 20% allocated for testing. The random state algorithm was used for a more accurate labeling. |
| **Model Fitting**: explain the key steps and activities you perform to fit the model. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate ones is the amount of performance tuning performed. |
| A total of four iterations was made on the data with several variables adjusted for finer algorithm modeling and prediction analysis. The first iteration was based upon the original amount of data a total of ten features and a single target variable. The second iteration was made based upon the same amount of target variables and the singular feature variable but with a shorter entropy, essentially a shorter tree than the original. The third and fourth iterations were made with a smaller amount of target variables and entropy concurrently. Each of these methods were visualized with the export\_graphiz package. |
| **Results** |
| **Model Properties:** explain the components of the fitted model and their characteristics. Leverage functions to summarize the model properties. Also, leverage visualization as required. |
| Without a defined entropy for the model, the extensive decision tree seemed a little difficult to decipher on a standard monitor screen. A defined entropy led top more seeable nodes. But a smaller entropy gave way to a decreased accuracy score of the model and a slightly skewed confusion matrix as visualized in the associated Jupyter Notebook file. |
| **Output Interpretation**: explain the result and interpret the final model output using terms that reflect the application area and in relation to the stated objective. This is where you check whether or not the stated objective is met. |
| The stated objective is met. There is a clear distinction between the physical measurements and performance in male and female athletes based upon the given dataset. In each of the following iterations, with the spread of different entropies and feature variables, the model had a high accuracy score of predicting the determining factors of an individual athlete. |
| **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. |
| The following are confusion Matrices from the iterations  Iteration 000:    Iteration 001:    Iteration 002:    Iteration 003:    In each of the iterations, the True positives and False Negative largely outweighed their counterparts. Which proves the methods and algorithms were proper for the analysis and objective at hand. |
| **Conclusion** |
| **Summary**: highlight the main findings in relation to the stated objective. You don’t need to discuss the details of the analysis and the model such as accuracy here, just focus on the key findings. |
| The models, in each iteration, both large and small entropies had an accuracy score of above 90%. It can be accurately noted that a male and female athlete in the following dataset can be categorized based upon their performance and body measurements. |
| **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. |
| I believe the classification of decision trees could be based upon the soley on one race of males and females. This case, depending on the cultural standards of a male and females’ role, could be a catalyst for differentiation factor between the draw of as female and male athlete. How often did these athletes train? What king of sports were these athletes participating in? The unknows of the people behind the study give a larger drive in the trustworthiness of the data. But, maybe all the sports separation is necessary for the health of the different athletes. |

|  |
| --- |
| **Appendix** |
|  |

**References**

*Marathon time qualifiers*. New York Road Runners. (n.d.). https://www.nyrr.org/tcsnycmarathon/runners/marathon-time-qualifiers

Navlani, A. (2024, June 27). *Python decision tree classification tutorial: Scikit-Learn Decisiontreeclassifier*. DataCamp. <https://www.datacamp.com/tutorial/decision-tree-classification-python>

Sanfilippo, J. (2017, June 15). *Beyond the binary: Looking past gender stereotypes in sports*. Medium. https://jensanfilippo17.medium.com/beyond-the-binary-looking-past-gender-stereotypes-in-sports-a7648ea92dce