|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 4: Clustering** | **Monique Reed** |
| **Should Male and Female Athletes Really Be Held to Different Standards? - Clustering** | |
| **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 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 on the fitness scores. It has always been stated that there are physical limitations and differences between the two genders in sports. For example, the qualification times for male and female athletes differ in sporting events such as the 2025 NYC marathon. The difference between inherent femininity and masculinity creeps into many discussions and research articles. However, such can be clarified further with a decision tree classifier applied to statistics and measurement of athletes. Based on the statistics, the model then decides if the athlete is a male or female. In this research, a K-means clustering algorithm is utilized. |
| **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 male or female based on their body statistics and performance measurements, denying or agreeing if the substantial 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 and a seaborn pair plot were printed to see the data’s correlation visualized. |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
| This dataset had to be preprocessed 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 classify further and correlate the data. But as mentioned earlier, thankfully, there were two genders and three classes to encode. Anything more would result in a reconsideration of using an encoder. |
| **Cluster Development**: explain the key steps and activities you perform to develop the clusters. 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. |
| K-means clustering is different than other machine learning models. In K-means clustering, splitting the data into training and testing sets is unnecessary to find clusters. Instead, all the data that is selected is used for this analysis.  Like most other algorithms, the first part of the development is the feature selection.  The next part is to define the K-means algorithm. The parameters in that function were the number of clusters and the random state number. In Iteration 000, for example, the number of clusters and centroids designated to the model. The second parameter is the random state model number, or the algorithm used to shuffle the model.  The model is then fit, and another variable predicts clusters for each point. |
| **Results** |
| **Cluster Properties:** explain the properties of the clusters by leveraging distance measures and discuss the clusters characteristics (differences and similarities). Produce appropriate cluster plots and discuss the output. |
| In two of the clusters, the centroids are distanced from the actual scatter plots.  Iteration 000:    Iteration 001:    Iteration 002:    Iteration 003: |
| **Output Interpretation**: explain the result and interpret the overall clusters 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. |
| All the iterations were visually evaluated by the elbow methods and scatter plots with defined centroids. Based upon the elbow method evaluation,  **Iteration 000:**  Elbow Method    **Iteration 001:**  Elbow Method    **Iteration 002:**  Elbow Method    **Iteration 003:**  Elbow Method |
| **Evaluation**: employ appropriate metrics (measures) to quantitatively evaluate the performance of the clusters. For unsupervised classification, this primarily involves distance metrics. |
| After a visualization, the completeness, silhouette and homogeneity scores were calculated for each model iteration.  The silhouette score is to evaluate the results of the cluster’s quality. This score typically ranges between -1 to 1. The higher the number, the better defined the clusters are. Similarly, the homogeneity measures the extent to where clusters only contain data from only one category. Lastly, the completeness score measures where all data points of the same category are assigned to the right clusters. Low scores are related to low levels of model performances.  **Iteration 000:**  **Completeness Score:** 0.2806899657320523  **Silhouette Score:** 0.273789162271316  **Homogeneity Score:** 0.452711678257412  **Iteration 001:**  **Completeness Score:** 0.4009149044905597  **Silhouette Score:** 0.19482100717463413  **Homogeneity Score:** 0.4203357564006076  **Iteration 002:**  **Completeness Score:** 0.39214859896262655  **Silhouette Score:** 0.3831520767017062  **Homogeneity Score:** 0.41010644630423376  **Iteration 003:**  **Completeness Score:** 0.3812296604147419  **Silhouette Score:** 0.3817550455849169  **Homogeneity Score:** 0.398669683761429 |
| **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. |
| Based on the methods of K-means clustering algorithms, it is not plausible to determine whether or not an athlete is male or female based on their fitness scores.  In all iterations, the model accuracy and validation scores were low, and the amount of data that could be used was also limited because of the different metrics of measurement used in the data. |
| **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. |
| Given this model and the data evaluation, the best thing for this model is the number of clusters, which was in iteration 000, the first iteration.  This model is complicated with this dataset; measuring the data points along with the proper centroids could have its challenges; maybe the need would be to categorize the data and, from there, plot points based upon label encoding.  Further, K-means clustering is primarily suitable for numeric values, but only values based upon. Thankfully, many measurements are in centimeters on the same numeric scale, so some values needed to be left out. That would pose complications for comparisons that involve more than measurements. |

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

**References**

Abdullahi, T. (2023, December 31). *Developing a clustering model: Utilizing the K-means algorithm*. Medium. https://medium.com/@aaltanim/constructing-a-clustering-model-employing-the-k-means-algorithm-546411fd9733

*Find and visualize clusters with K-means: DataLab*. Find and Visualize clusters with K-Means | DataLab. (n.d.). https://www.datacamp.com/datalab/templates/recipe-python-k-means

Rogel-Salazar, D. J. (2024, July 16). *K-means clustering in Python: A step-by-step guide*. Domino Data Lab. https://domino.ai/blog/getting-started-with-k-means-clustering-in-python

“Marathon Time Qualifiers.” *New York Road Runners*, www.nyrr.org/tcsnycmarathon/runners/marathon-time-qualifiers. Accessed 7 Dec. 2024.

Sanfilippo, Jen. “Beyond the Binary: Looking Past Gender Stereotypes in Sports.” *Medium*, Medium, 15 June 2017, jensanfilippo17.medium.com/beyond-the-binary-looking-past-gender-stereotypes-in-sports-a7648ea92dce.