|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 4: Clustering** | **Sulchan Yoon** |
| **KMeans Clustering with Iris Species** | |
| **URL to dataset: https://www.kaggle.com/datasets/uciml/iris?datasetId=19** | |

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. |
| As the world continues to evolve, flowers continue to create a variety of variations. Researchers and gardeners alike look to be able to understand and appropriately classify each type of variation present within flowers. In this dataset, we are presented with a variety of size measurements for the iris plant. Iris’s are known to have two types of “petals” as show in the photo below, that defines what makes an iris present its beautiful and unique shapes. The first part will be the Sepal which shows in a more drooping or falling manner. This is used by bumblebees as a type of landing pad. The petals are the attraction piece, showing in a typically brighter color and looking towards the sky, showing where bumblebees can enjoy the nectar of the flower.  blue flag iris flower. The parts of the flower are labeled. |
| **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. |
| Our goal with this analysis is to focus on K Means Clustering with python for purposes of classification through grouping. With the data types we have, we will want to correctly define the specific species defined by the sepal and petal size. Our data is broken out into the types below:   |  |  |  | | --- | --- | --- | | Column Number | Column Heading | Definition | | 1 | ID | Object Identifier | | 2 | SepalLengthCm | Length in Cm of the Sepal of the Iris | | 3 | SepalWidthCm | Width in Cm of the Sepal of the Iris | | 4 | PetalLengthCm | Length in Cm of the Petal of the Iris | | 5 | PetalWidthCm | Width in Cm of the Petal of the Iris | | 6 | Species | Categorical Species of the Iris |   \*\*\*\*\*\*\*\*\*\*WRITE ON K MEANS Clustering\*\*\*\*\*\*\*\*\*\*\* |
| **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 contains a total of 150 iris’ that have been observed. Each iris has had the sepal length, width, and petal length and width measured. On an initial review of the sepal measurements, there is a clear distinction between iris-selosa versus the other two species, as shown in chart 1.  What we can also find when reviewing the petal measurements, is that the overall sizes are significantly different, however there is still a clear distinction between all three species.    The Sepal lengths go down to a minimum of 4.3 cm, and max to 7.9 cm. Comparing that to the petal length, we see that there is a minimum of 1.0 cm, and max of 6.9 cm. We find that the size range is a lot larger for the petal lengths than the sepal length. When looking at the widths, we have sepal min at 2 cm and max at 4.4 cm. The petals show 0.1 cm and max at 2.5 cm. The ratios show petals have much more variability and are also significantly smaller than the sepal sizes. The high variability is what also shows to be one of the main reasons between each of the species type. |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
| Our main focus is to separate our x and y variable datasets. The x values will be focused on the size measurements for both sepal and petal sizes. The y will be the species. We will be dropping the ID column as it does not play a part in this algorithm. For simplicity of calculations, we will also be labeling our y values using a simple label encoder. Lastly, as part of the preprocessing, we will create our training and testing data split using a general standard split. My preference is to always use a common random state (10) and a initial test size of 20%. |
| **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. |
| One of the best methods to build out the KMeans clustering is figuring out the number of clusters. This is where the most amount of errors can be taken care of and also over/under fitting occurs. A common method to find the best number of clusters is to use the Elbow Method. This method is a graphical representation of finding the optimal “K” in KMeans clustering. This uses a process called WCSS – Within Cluster Sum of Squares which is the sum of square distances between the centroids and each of the points. To start off our process, we push a model without defining the hyperparameters and pull the Elbow calculator:  model = KElbowVisualizer(kmeans,k=(1,11))  This will provide us the optimal K values given the inputted range (in our case 1 to 11). The default method for scoring this value will be a distortion method which computes the sum of squared distances from each point to its centroid. |
| **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. |
| Our estimate through the elbow method has an output of k = 3    Thus, we can say that there are three clusters built for an ideal KMeans clustering. |
| **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. |
| Our output using the three clusters provides us with the following:    We can see that there are three distinct clusters that have been grouped. Through there are some outliers that seem to have been blended between the Setosa and the Virginica classes. Overall, using the KMeans Clustering with an Elbow method for calculating the ideal number of clusters provided us with a strong candidate for a model on categorizing the dataset. |
| **Evaluation**: employ appropriate metrics (measures) to quantitatively evaluate the performance of the clusters. For unsupervised classification, this primarily involves distance metrics. |
| We will be able to use our Centroid graph to see the our values. We ended with the following three points   1. Iris Versicolour (5.006, 3.418) 2. Iris Virginica (5.902, 2.748) 3. Iris Setosa (6.85, 3.074) |
| **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. |
| Overall, our key findings in this data is that the dataset has proven to show highly distinct classifications. We see there are three distinct class types, and through our modeling with KMeans Clustering, we also find the Elbow method to show the ideal clustering fixed at the three distinct classification types. |
| **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 find that KMeans clustering is a fun method to categorize and classify objects. I think that there would have been a drastically different result and more mixing if we had a much larger dataset. The limitation of finding the ideal model was limited to the small dataset. Overall we could see bigger changes with the hyperparameters if we had more data to work with. |

|  |
| --- |
| **Appendix** |
| Images: |

**References**

UCI Machine Learning (2016). Iris Species. Retrieved July 2023 from https://www.kaggle.com/datasets/uciml/iris?datasetId=19

U.S. Forest Service. Forest Service Shield. (n.d.). https://www.fs.usda.gov/wildflowers/beauty/iris/flower.shtml

Yoon. (2023). *Data\_430\_GC Iris* [ipynb]. Retrieved from https://github.com/sulchan/Data\_430\_GC