|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 3: Decision Trees** | **Sulchan Yoon** |
| **Decision Trees 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 decision trees with python for purposes of classification. 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 |   Decision Trees is a machine learning algorithm with high levels of versatility. It focuses mostly on classifications and regression tasks, which is a perfect fit for our initial analysis on classifying the species for the iris. Decision Trees can also be broken down into one of the most popular and powerful machine learning algorithm, the Random Forest, though we will continue to focus on decision trees for this project.  A few of the advantages for decision trees include simplicity of the overall algorithm allowing for it be easily understood. Decision trees can also handle both numerical and categorical variables, thus separate labeling would not necessarily be required as part of the data pre-processing. Included in the data pre-processing and overall data cleanup, decision trees are a lot less affected by outliers which will support a larger and messier dataset. |
| **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. |
| As mentioned, decision trees are overall a forgiving algorithm, thus minimum preprocessing will be required. 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%. |
| **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. |
| Now that we have finalized our data set preprocessing and created the training and test set, we will be able to fit our data into the decision tree model:    We continue to use our random state at (10). For the splitter, we use a random such that it takes the feature randomly but within the same distribution. Though it does not select the best feature, by using a splitter = best. This allows for a little less overfitting and adds a little more randomness overall to the model. |
| **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. |
| The model we chose to use is the Decision Tree Classifier. Here we entered default variables under the training set for x and y.    For the purposes of reproducibility in our research and analysis, we chose a random\_state of 10 and selected a splitter of random to further randomize our approach as explained in the model fitting. |
| **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. |
| Our main goal was to find a suitable decision tree algorithm that would help classify iris’ by the measurements gained through the sepal and petal sizes. Using the Decision Tree method and a variety of randomized testing scenarios, we were able to pull an accuracy score of 96.67%. I would recommend this method to a high degree of accuracy overall. However, through the evaluation stage, we will be able to find further details into the decision tree’s algorithm to go deeper into the definition of accuracy. |
| **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. |
| Below is the first 3 depths that have been produced from the modeling through decision tree. To see the full code please see the appendix.    The Samples: This defines how many instances is being applied, 82 instances have petal length greater than 1.851 cm. Then 59 samples is smaller than 5.423 cm at depth 2.  Value: This will show of the instances, how many each class is applied, so in purple, we see there are 0 iris-setosa, 0 iris-veriscolor, and 23 iris-virginica.  Gini: This is the measure of impurity, where a true pure is defined at gini = 0. The depth 2 left of gini = 0.456 can be defined as:    Entropy: Is a measure of impurity where entropy of 0 is when a set contains instances of only one class. The purple box, depth 2 right, would be considered pure and thus entropy is 0. The depth 3 left, with a gini of 0.208 would have an entropy of 0.362 equal to: |
| **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. |
| Through this exercise, we have been able to find a suitable model to classify our iris plants given the measurements. We were able to calculate an accuracy score of the model to be a high value of 96.67% accurate. Though, to dig further into each of the situations, we also looked into measuring information gain on purity and impurity within each node of our tree. Through our model, we found a generous gini scores throughout along with a calculated entropy score as well. We are glad we did not get pure gini throughout as that would mean we may be looking at a situation of overfitting. That being said, given our conditions and output readings for the our modeled tree, I would say we are successful in finding a well versed classification model using the Decision Tree method. |
| **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. |
| One of the major limitations when it comes to trees is determining depth based on the random training and test sets. I find that this dataset may be a smaller dataset, and thus gives little options to be able to test with different scenarios when splitting the data. I would also like to explore the significant differences when going through the Decision Tree Classifier when setting a splitter to random versus to best given a much larger dataset. Current status, the splitter of best would give an accuracy score of 100%. I would also like to redefine the Decision Tree Classifier with a selected max\_depth parameter to see how that may affect overall performance. |

|  |
| --- |
| **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