Programming Assignment 03

Assignment:
You are the new junior member of a data mining consulting firm. Your newest client has uncovered a large number of unknown animals and has asked you to build a classification system to help their field agents classify these new animals as mammal or non-mammal from the observational data they have collected.

The client feels their field agents can reliably collect the following observational data: presence of hair, presence of feathers, if they lay eggs, if they nurse their young with milk, if they are airborne, if they are aquatic, if they are predatory, if they have teeth, if they have a backbone, if they breath air, if they are venomous, if they have fins, the number of legs they have, if they have a tail, if they are domesticated, if they are cat sized, and approximate gestation period.

Your client, audience, can have technical skills that range from introductory (the field agents) to highly skilled (the clients in house scientists). Keep this in mind when determining what to document and whether to use visualizations or not.

The client has provided a data set of known animals to train and test your classification system with. The animal-taxonomy.csv can be found in Canvas, the company document management system. You will need to separate the data set into a training and test data set.

Follow the corporate report template (submission guide) to format your work.

The details:

Use the data to create a 1R classifier to classify input data as mammal or non-mammal.

In the example description section describe the attributes that have been provided and any attributes that need to be generated. Include whether the attribute is nominal, ordinal, interval, or ratio. Document the process used to generate the generated attributes.

In the data import and wrangling section import the data and discretize any interval or ratio attributes.

In the data import and wrangling section create a two-class classification with the class mammal and non-mammal.

You will need to separate the provided data into a training set and a testing set. Be sure to use and document the seed you use for reproducibility.

There should be a good Exploratory Data Analysis section where you look for data features and issues.

In the mining and analytics section use the training data set to manually create a 1R classifier to classify each example as mammal or non-mammal. Manually means you can use code to build a frequency table of the classes of each attribute and/or a table of probabilities. It also means you will use the tables to determine the classification rule. When manually generating the classifier I don't want you to use a one package does it all method (like the one in R). I want you to look at the tables and determine your own classification rule.

In the evaluation section use the test data set to create a confusion matrix for the manually created 1R classifier. Then determine the Accuracy and F1-Score for the model. Show any intermediate calculations. Use the Accuracy and F1-Score to evaluate which classifier is better.

In the results section clearly describe the attribute you have chosen at your classifier and discuss how successful the model is..

In the reference section document any references you need.

Use visualizations where appropriate.

The following is the only metadata about the attributes in the data set..

Summary of provided data:
Number of Attributes: 18 (animal name, 15 Boolean attributes, 2 numerics)
Classification (7 classes)
Attribute Information: (name of attribute and type of value domain)

1. animal name: Unique for each instance
2. hair Boolean
3. feathers Boolean
4. eggs Boolean
5. milk Boolean
6. airborne Boolean
7. aquatic Boolean
8. predator Boolean
9. toothed Boolean
10. backbone Boolean
11. breathes Boolean
12. venomous Boolean
13. fins Boolean
14. legs Numeric (set of values: {0,2,4,5,6,8})
15. tail Boolean
16. domestic Boolean
17. catsize Boolean
18. gestation Numeric (in days)
19. type Label (amphibian, arthropod, bird, fish, insect, mammal, reptile)