Zoo Analysis

DSCI310-group7

Apr 09, 2022

CONTENTS

Summary

The data set we will be using is Zoo (1990) provided by UC Irvine Machine Learning Repository. It stores data with 7 classes of animals and their related characteristics including animal name, hair, feathers and other attributes. In this project, classification is the main method on predicting a most likely type of a given animal.

Below is a table of contents

CHAPTER

ONE

INTRODUCTION

The earth is an amazing planet that cultivates branches of animals. In general, scholars split them into 12 classes including mammals, birds, reptiles, amphibians, fishes, insects, crustaceans, arachnids, echinoderms, worms, mollusks and sponges [BioExplorer.net, 2022]. The traditional way in animal classification is manually identifying the characteristics and attributing it the mostly close class [N. Manohar and Kumar, 2016]. However, it is tedious and time consuming, especially when the data set is very huge. A question hereby comes to us, if we can apply K-nearest neighbors (KNN) algorithms in predicting the type an animal belongs to given its related characteristics, such as hair, feathers, etc.? Therefore, in this project, we will show how we use KNN to do classification in animals based on data set [Repository, 1990] which contains 1 categorical attribute, 17 Boolean-valued attributes and 1 numerical attribute. The categorical attribute appears to be the class attribute. Detailed breakdowns are as follows:

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. `type`: Numeric (integer values in range [1,7])

CHAPTER

TWO

METHODS & RESULTS

We are going to use multiple analysis to classify the type of the animals using 16 variables including hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, catsize as our predictors. To predict the class of a new observation, the algorithms of each type will be further explained before implementation.

The first thing is to import the data. The data set is downloaded from UCI repository. It is then saved as a csv file in this project repository. Some exploratory data analysis needs to be run before running the actual analyses on the data set. Here is a preview of pre-processed data set:

```
   index  hair  feathers  eggs  milk  airborne  aquatic  predator  toothed  \
0      0     1         0     0     1         0        0         1        1
1      1     1         0     0     1         0        0         0        1
2      2     0         0     1     0         0        1         1        1
3      3     1         0     0     1         0        0         1        1
4      4     1         0     0     1         0        0         1        1

   backbone  breathes  venomous  fins  legs  tail  domestic  catsize
0         1         1         0     0     4     0         0        1
1         1         1         0     0     4     1         0        1
2         1         0         0     1     0     1         0        0
3         1         1         0     0     4     0         0        1
4         1         1         0     0     4     1         0        1
```

It is checked that there aren't missing values in the data set, we can clearly deduce that the data set is clean according to the data summary we generated above. Since most features are binary and categorical, there is no need to do normalization and standardization.

As shown in *fig.1*, the histograms of each feature are generated. The ones with skewed distribution might be more decisive in the prediction. However, since the data set is relatively small, all the features except the `animalName` are going to be used to predict. In the next part, we are going to split the data, into the training set and testing set. After that, different classification models will be trained and evaluated.

2.1 Classification

Now we will use the training set to build an accurate model, whereas the testing set is used to report the accuracy of the models. Here is a list of algorithms we will use in the following section:

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine

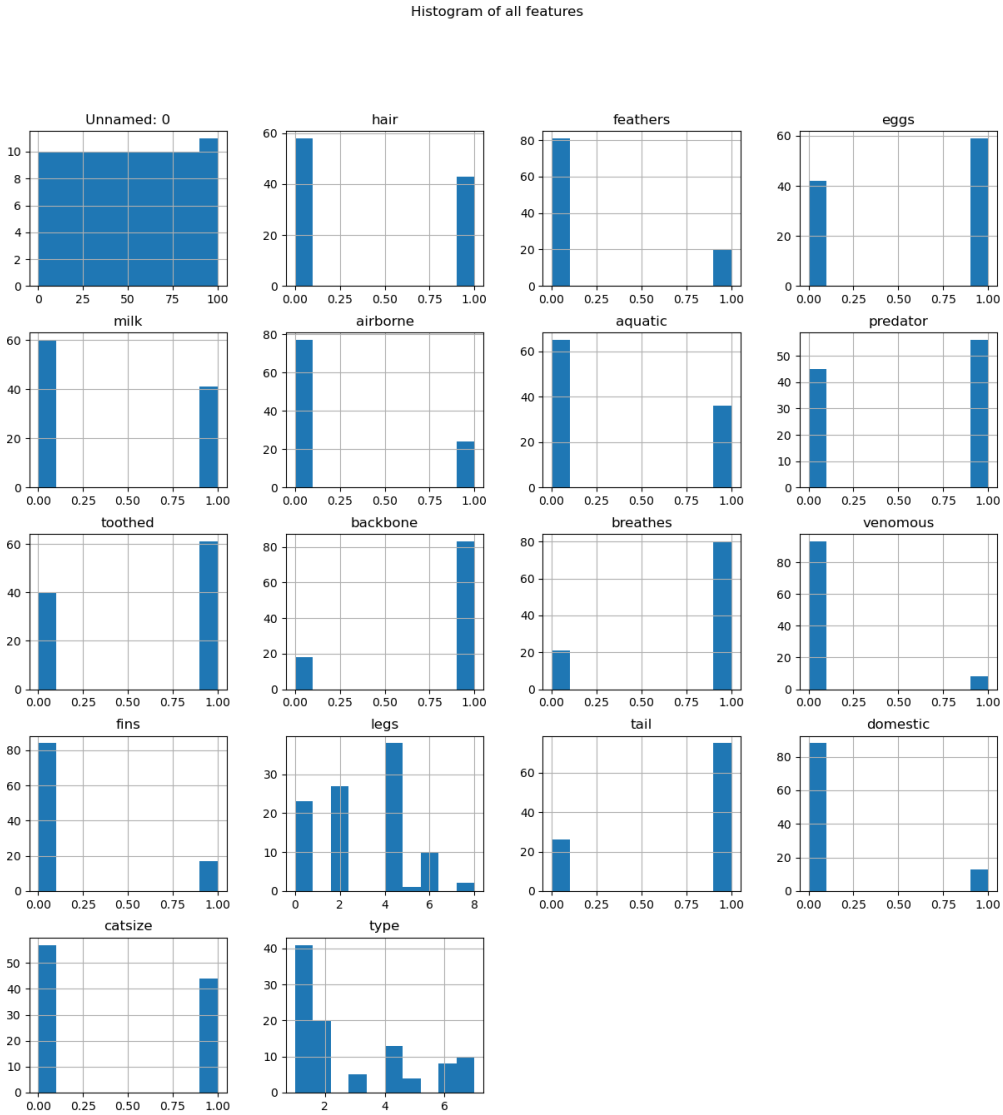

Fig. 2.1: A summary table of the data set

- Logistic Regression

To train and evaluate each model, we split the dataset into training and testing sets. We use 80% of the total data to train the models, and the rest of the data is aimed to test the models.

2.1.1 KNN

KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some basic mathematics we might have learned earlier. Basically in terms of geometry we can always calculate the distance between points on a graph. Similarly, using KNN we can group similar points together and predict the target with our feature variables(x).

First of all, we have to train the model for different set of K values and finding the best K value. Then we want to plot the accuracy for different K values.

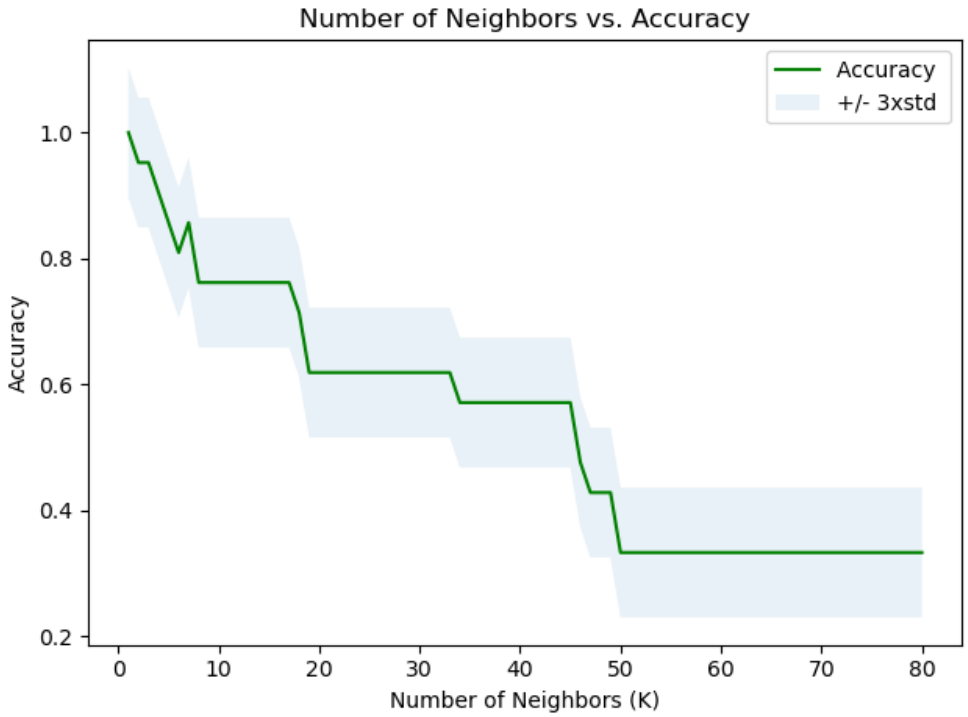

Fig. 2.2: A plot reveals the relationship between K and corresponding accuracy

As shown in *fig.2*, less K values provide higher accuracy. To find the best K value, we tuned the hyperparameter using GridSearch algorithm. After tuning, the best K value is 1.

2.1.2 KNN final model & Evaluation

After fitting the model using `K=1`, we evaluate the KNN model by Cross Validation and calculating the precision, recall, f1-score and support.

KNN Cross Validation Result:

```
      criteria     score
0     fit_time  0.002855
```

(continues on next page)

(continued from previous page)

```
1   score_time  0.003670
2   test_score  0.936847
3  train_score  1.000000
```

KNN Classification Report:

	index	precision	recall	f1-score	support
0	1	1.0	1.0	1.0	7.0
1	2	1.0	1.0	1.0	5.0
2	4	1.0	1.0	1.0	1.0
3	5	1.0	1.0	1.0	1.0
4	6	1.0	1.0	1.0	3.0
5	7	1.0	1.0	1.0	4.0
6	accuracy	1.0	1.0	1.0	1.0
7	macro avg	1.0	1.0	1.0	21.0
8	weighted avg	1.0	1.0	1.0	21.0

2.1.3 Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

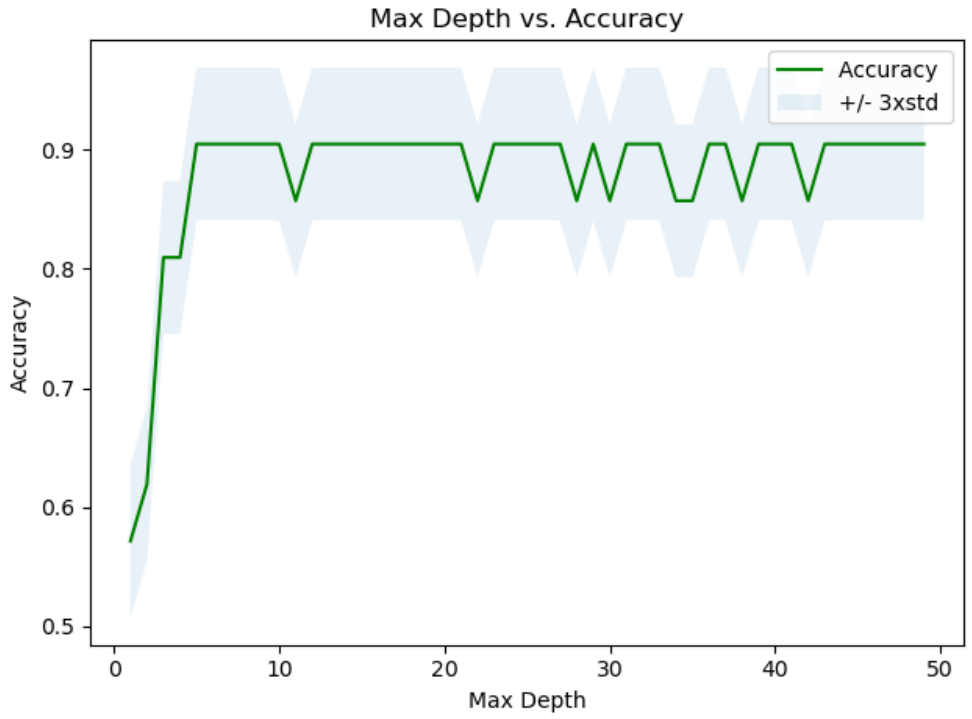

Fig. 2.3: A plot reveals the relationship between deepth and corresponding accuracy

As shown in the *fig.3*, the best depth of the Decision Tree is around small. We can confirm that the best value of the depth is 5 after tuning the hyperparameter and calculating the accuracy.

2.1.4 Decision Tree final model & evaluation

After training the model, we obtain the Cross Validation score, as well as the precision, recall, f1-score and support.

DT Cross Validation Result:

	criteria	score
0	fit_time	0.001509
1	score_time	0.001098
2	test_score	0.950000
3	train_score	0.995833

DT Cross Validation Result:

	index	precision	recall	f1-score	support
0	1	1.000000	1.000000	1.000000	7.000000
1	2	1.000000	1.000000	1.000000	5.000000
2	4	1.000000	1.000000	1.000000	1.000000
3	5	1.000000	1.000000	1.000000	1.000000
4	6	0.750000	1.000000	0.857143	3.000000
5	7	1.000000	0.750000	0.857143	4.000000
6	accuracy	0.952381	0.952381	0.952381	0.952381
7	macro avg	0.958333	0.958333	0.952381	21.000000
8	weighted avg	0.964286	0.952381	0.952381	21.000000

2.1.5 Support Vector Machine

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes[editors, 2022].

Final SVM is here used the splited test part to train again for better training, and better prediction. An svm evaluation as well as the final model is also provided below.

2.1.6 SVM training model Jaccard Score, final model and evaluation

SVM Classification Report:

	index	precision	recall	f1-score	support
0	1	1.000000	1.000000	1.000000	7.000000
1	2	1.000000	1.000000	1.000000	5.000000
2	4	0.333333	1.000000	0.500000	1.000000
3	5	1.000000	1.000000	1.000000	1.000000
4	6	1.000000	1.000000	1.000000	3.000000
5	7	1.000000	0.500000	0.666667	4.000000
6	accuracy	0.904762	0.904762	0.904762	0.904762
7	macro avg	0.888889	0.916667	0.861111	21.000000
8	weighted avg	0.968254	0.904762	0.912698	21.000000

2.1.7 Logistic Regression

Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems[Education, 2020].

2.1.8 Logistic Regression training model Jaccard Score, final model and evaluation

LR Classification Report:

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [1], in <module>
      1 {
      2     "tags": [
      3         "hide-input"
      4     ]
      5 }
----> 6 lr_classification_report= pd.read_csv("../results/csv/lr_classification_
 ↪report.csv")
      7 lr_classification_report.columns.values[0]="index"
      8 lr_classification_report

NameError: name 'pd' is not defined
```

CHAPTER

THREE

DISCUSSION

After analyzing all the different 4 models K Nearest Neighbor(KNN), Decision Tree, Support Vector Machine and Logistic Regression, we found KNN is best to predict the animal type here. As you have seen in the model evaluation tables before, for accuracy KNN is the best, the second-best is decision tree method and following by Support Vector Machine and Logistic Regression.

The result of KNN was expected as KNN is the best in grouping similar data points together and giving the best prediction results. Predicting the correct animal type with the highest accuracy have a huge impact on identifying animal types. These models can be used to identify animal types instantly for example if someone saw/discovered an animal and the type is not identified then they can feed all the characteristics fields to the model. The model can predict the animal type accurately, which is way more accurate than identifying and classifying the animal based on common sense. Thus our model can increase the research potential in many fields but not just limited to Marine Science, Animal Science, Forestry, and etc.

A future question might be led to: how we are going to maintain the accuracy of predictions when working with more diverse groups of animals? Another possible aspect of this can be how some attributes of animals will relate to each other, for instance, relation between animals which has teeth vs predator. Furthermore, how we are going to use those relations to predict behaviors and attributes of newly discovered animals and how we are going to make our perceptions on animals even more detailed. These models and their advancements will not only widen our knowledge in terms of animal biology but will also let us find all other possible relations within the nature in a much more efficient way.

CHAPTER

FOUR

REFERENCES

BIBLIOGRAPHY

[1] BioExplorer.net. Types of animals. *BioExplorer*, 2022. URL: https://www.bioexplorer.net/animals/.

[2] Y. H. Sharath Kumar N. Manohar and G. H. Kumar. Supervised and unsupervised learning in animal classification. *Communications and Informatics (ICACCI)*, 2016. 2016 International Conference on Advances in Computing.

[3] UCI Machine Learning Repository. Zoo. *UCI Machine Learning Repository*, 1990. URL: https://archive.ics.uci.edu/ml/datasets/zoo.

[4] TDS editors. How data can make a difference in the real world. *Towards Data Science*, 2022. URL: https://towardsdatascience.com.

[5] IBM Cloud Education. What is data science. 2020. URL: https://www.ibm.com/cloud/learn/data-science-introduction.