Prediction on Animal Type

Elaine Zhou, Jossie Jiang, Swakhar Poddar & Weihao Sun

Contents

0.1 Summary

The data set we will be using is Zoo (Repository 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, we will use classification as our method to predict a most likely type of a given animal.

0.2 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:

animal name: Unique for each instance hair: Boolean feathers: Boolean eggs: Boolean milk: Boolean airborne: Boolean aquatic: Boolean predator: Boolean toothed: Boolean backbone: Boolean breathes: Boolean venomous: Boolean fins: Boolean legs: Numeric (set of values: {0,2,4,5,6,8}) tail: Boolean domestic: Boolean catsize: Boolean type: Numeric (integer values in range [1,7])

0.3 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.

Table 1: zoo data

...1	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins
0	1	0	0	1	0	0	1	1	1	1	0	0
1	1	0	0	1	0	0	0	1	1	1	0	0
2	0	0	1	0	0	1	1	1	1	0	0	1
3	1	0	0	1	0	0	1	1	1	1	0	0
4	1	0	0	1	0	0	1	1	1	1	0	0

After checking whether there are 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.

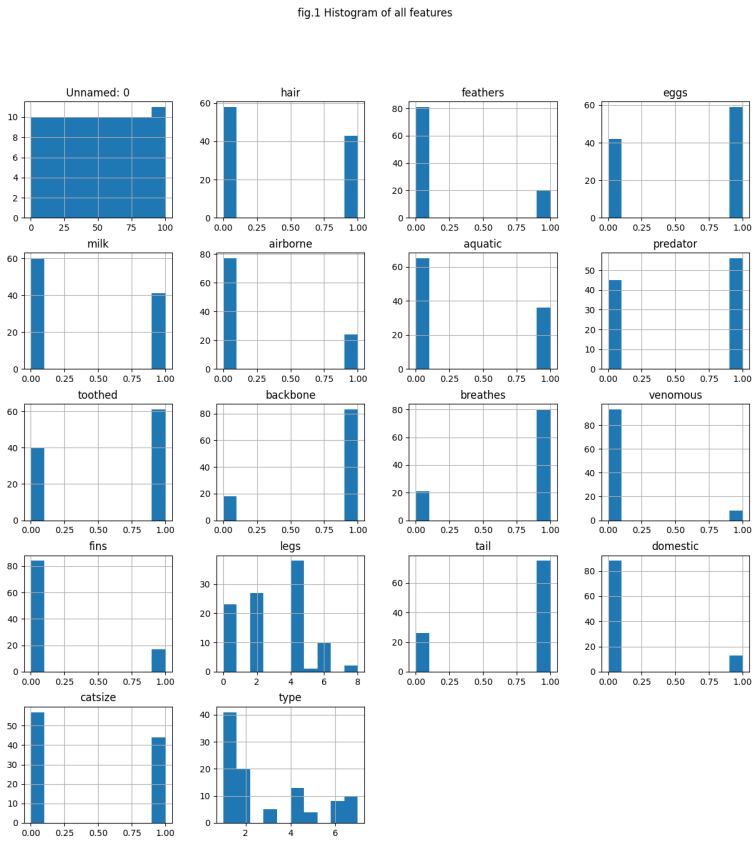

Figure 1: Histogram of all features

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.

0.3.1 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 Logistic Regression

0.3.2 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.

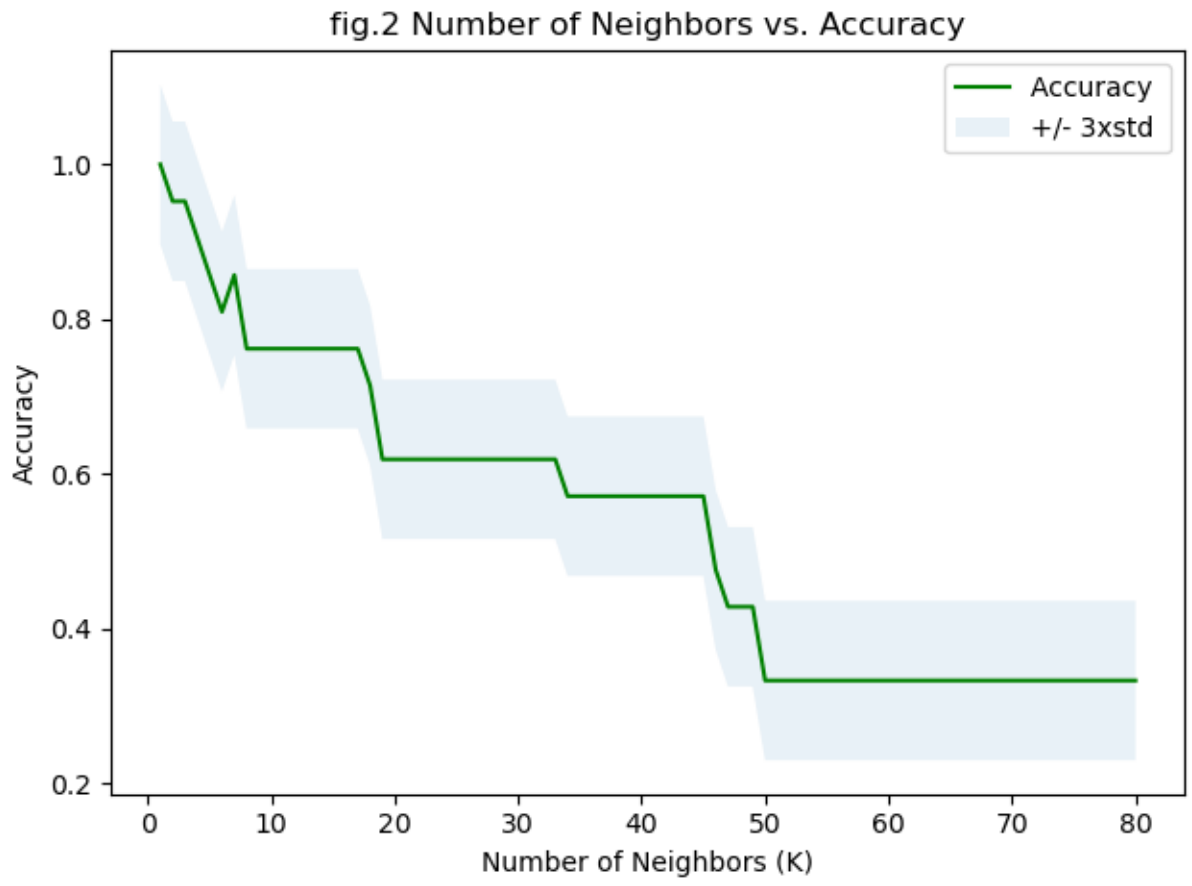

Figure 2: Number of Neighbors vs. Accuracy

The best accuracy was with the values 1.0 when k = 1.

0.3.3 KNN final model & Evaluation

As the best accuracy was with K = 1, by using K = 1 for the final KNN model the final KNN model is built in the following.

Table 2: cross validate result of knn

...1	0
fit_time	0.0003622
score_time	0.0026897
test_score	0.9368471
train_score	1.0000000

Table 3: classification report of knn

...1	precision	recall	f1-score	support
1	1	1	1	7
2	1	1	1	5
4	1	1	1	1
5	1	1	1	1
6	1	1	1	3
7	1	1	1	4
accuracy	1	1	1	1
macro avg	1	1	1	21
weighted avg	1	1	1	21

0.3.4 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).

The best accuracy was with value 0.9047619047619048 with maximum depth = 5.

0.3.5 Decision Tree final model & evaluation

As the best accuracy takes place when the max depth is 5. Using max depth = 5 for the final decision tree we can get the following result.

0.3.6 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.

Table 4: cross validate result of decision tree

...1	0
fit_time	0.0012702
score_time	0.0002733
test_score	0.9500000
train_score	0.9958333

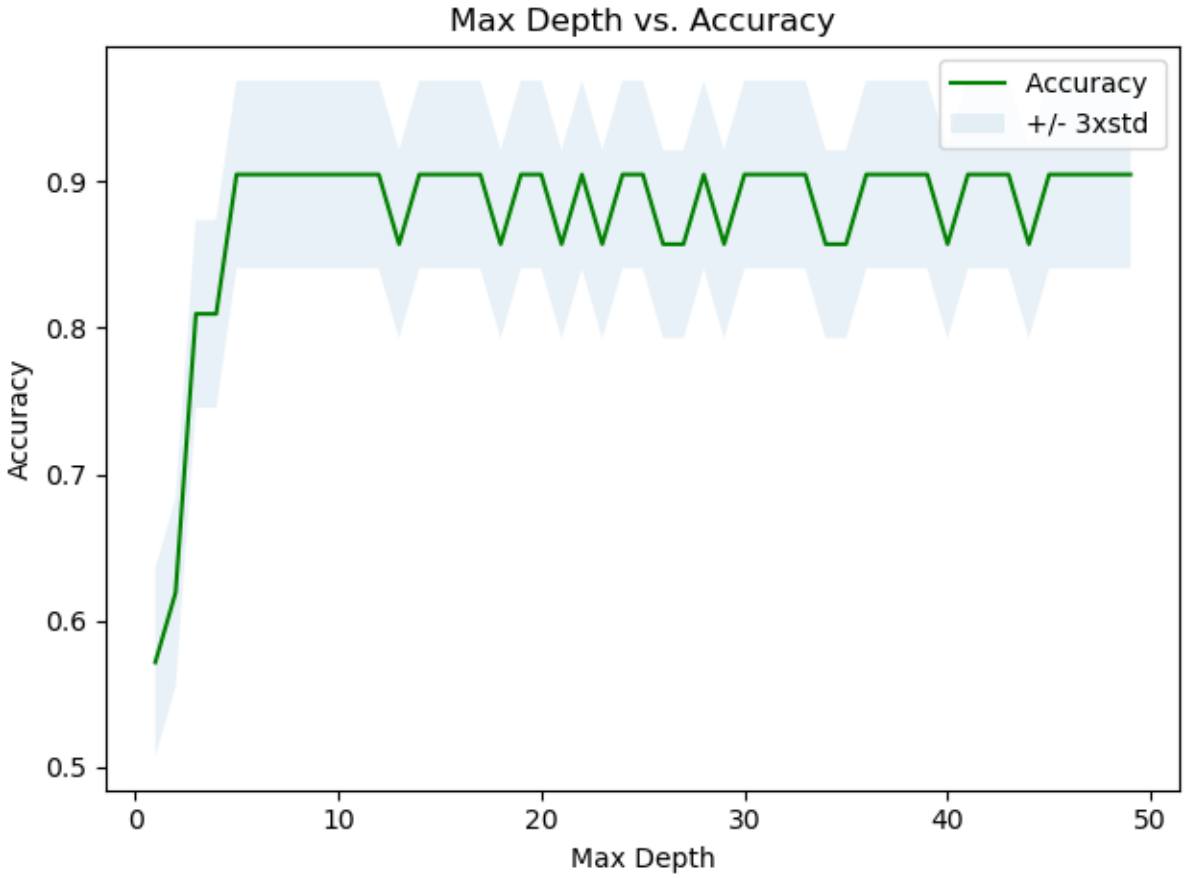

Figure 3: Max Depth vs. Accuracy

Table 5: classification report of decision tree

...1	precision	recall	f1-score	support
1	1.0000000	1.0000000	1.0000000	7.000000
2	1.0000000	1.0000000	1.0000000	5.000000
4	1.0000000	1.0000000	1.0000000	1.000000
5	1.0000000	1.0000000	1.0000000	1.000000
6	0.7500000	1.0000000	0.8571429	3.000000
7	1.0000000	0.7500000	0.8571429	4.000000
accuracy	0.9523810	0.9523810	0.9523810	0.952381
macro avg	0.9583333	0.9583333	0.9523810	21.000000
weighted avg	0.9642857	0.9523810	0.9523810	21.000000

Table 6: classification report of support vector machine

...1	precision	recall	f1-score	support
1	1.0000000	1.0000000	1.0000000	7.0000000
2	1.0000000	1.0000000	1.0000000	5.0000000
4	0.3333333	1.0000000	0.5000000	1.0000000
5	1.0000000	1.0000000	1.0000000	1.0000000
6	1.0000000	1.0000000	1.0000000	3.0000000
7	1.0000000	0.5000000	0.6666667	4.0000000
accuracy	0.9047619	0.9047619	0.9047619	0.9047619
macro avg	0.8888889	0.9166667	0.8611111	21.0000000
weighted avg	0.9682540	0.9047619	0.9126984	21.0000000

Table 7: classification report of logistic regression

...1	precision	recall	f1-score	support
1	0.8750000	1.0000000	0.9333333	7.0000000
2	0.8333333	1.0000000	0.9090909	5.0000000
4	0.5000000	1.0000000	0.6666667	1.0000000
5	0.0000000	0.0000000	0.0000000	1.0000000
6	0.7500000	1.0000000	0.8571429	3.0000000
7	1.0000000	0.2500000	0.4000000	4.0000000
accuracy	0.8095238	0.8095238	0.8095238	0.8095238
macro avg	0.6597222	0.7083333	0.6277056	21.0000000
weighted avg	0.8115079	0.8095238	0.7579468	21.0000000

0.3.7 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).

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

Final LR model is here used the splited test part to train again for better training, and better prediction. The LR evaluation also down here.

0.4 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. This might lead to a future question in which 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.

Reference

BioExplorer.net. 2022. "Types of Animals." *BioExplorer*. https://www.bioexplorer.net/animals/.

editors, TDS. 2022. "Towards Data Science." https://towardsdatascience.com.

Education, IBM Cloud. 2020. "What Is Data Science." IBM. https://www.ibm.com/cloud/learn/data-science-introduction.

N. Manohar, Y. H. Sharath Kumar, and G. H. Kumar. 2016. "Supervised and Unsupervised Learning in Animal Classification." *Communications and Informatics (ICACCI)*.

Repository, UCI Machine Learning. 1990. "Zoo." *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/zoo.