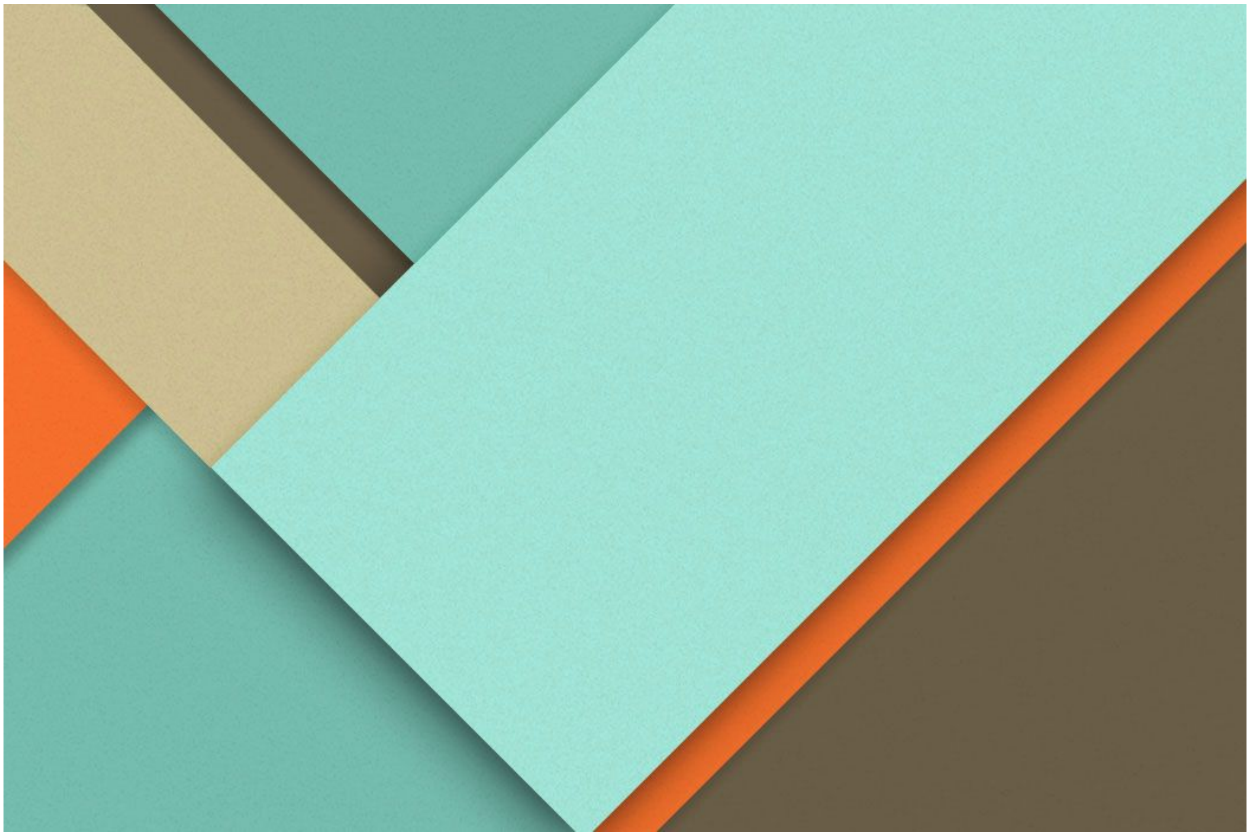


# ML MINI PROJECT

09.11.2018

Akash Cheerla
Computer Science
111601029

## Problem Statement

This dataset consists of 101 animals from a zoo. There are 16 variables with various traits to describe the animals. The 7 class types are: Mammal, Bird, Reptile, Fish, Amphibian, Bug and Invertebrate. The purpose for this dataset is to be able to predict the classification of the animals, based upon the variables. Find  the best machine learning ensembles/methods for classifying these animals based upon the variables given?

## Data Set

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. class_type Numeric (integer values in range [1,7])

## class.csv

This csv describes the dataset

1. Class_Number Numeric (integer values in range [1,7])
2. Number_Of_Animal_Species_In_Class Numeric
3. Class_Type character -- The actual word description of the class
4. Animal_Names character -- The list of the animals that fall in the category of the class

## Finding best parameters

To find best parameters for a model, we will be doing k-split to the Dataset and for each parameter, we will be calculating its performance on each split and we will take the mean of validation scores obtained in each split and we will measure that score with the best score we have observed so far and if this parameter is better than the best parameter, then mark this parameter as the best parameter. Then, after finding the best parameter, we will find the split which gives best Validation Score for the model with this best parameter. And we will use this model to predict the test dataset.

## Approach

I have used Four classification algorithms to perform multi-classification namely

- Logistic Regression
- Random Forest Classifier
- Support Vector Classifier
- Decision Tree

I have chosen Logistic Regression because on analyzing features, i observed that the data may be linearly separable and Linear Regression works really well when data is linearly separable.

I have chosen Support Vector Classifier with 'rbf' kernel and 'linear' kernel as SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

I have chosen Random Forest Classifier because Random Forest can handle missing values, and the Random Forest classifier can be modeled for categorical values.

The difference between Random Forest and the decision tree is that in Random Forest, the process es of finding the root node and splitting the feature nodes will run randomly.

## Model Selection

Here, i am using k-fold technique to split the given Dataset into k-different folds. We have to do this with every fold (all k-folds). In each split we will train our model and estimate the correct model parameters by analyzing the accuracy in Validation Set for each parameter in different splits of Dataset. The parameters which performs the best will be chosen for the model predicting the Test Data.

We want to compare different algorithms, selecting the best-performing one as well as the best performing model from the algorithm's hypothesis space.

## Best Parameter

To find best parameters for a model, we will be doing k-split to the Dataset and for each parameter, we will be calculating its performance on each split and we will take the mean of validation scores obtained in each split and we will measure that score with the best score we have observed so far and if this parameter is better than the best parameter, then mark this parameter as the best parameter. Then, after finding the best parameter, we will find the split which gives best Validation Score for the model with this best parameter. And we will use this model to predict the test dataset.

## OUTPUTS

Let's check the experimental scores obtained

Output of  Random Forest : 0.966486

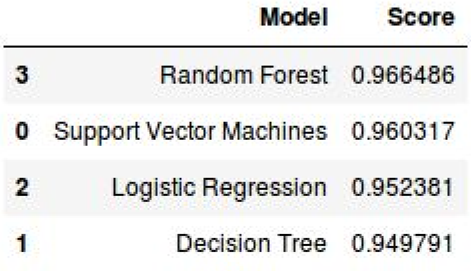
Output of Support Vector Classifier : 0.960317

Output of Logistic : 0.952381

Output of Decision Tree : 0.949791

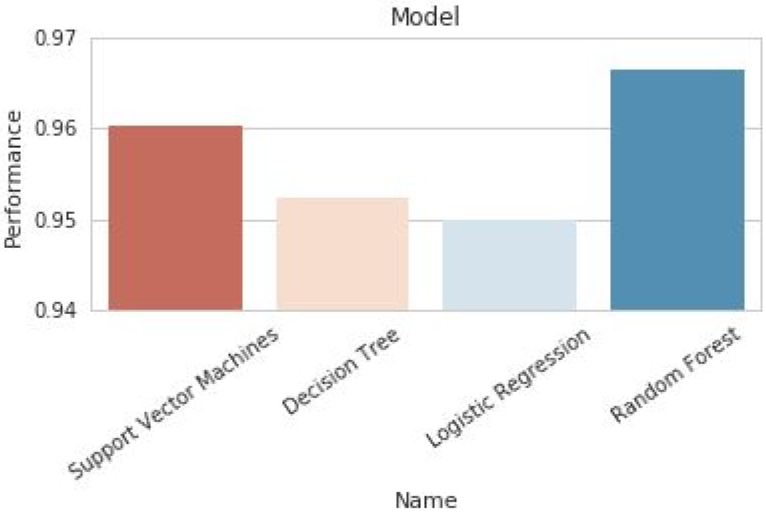
## Performance of the Models

So here the performance of the 4 models used



| | Model | Score |
|---|---|---|
| 3 | Random Forest | 0.966486 |
| 0 | Support Vector Machines | 0.960317 |
| 2 | Logistic Regression | 0.952381 |
| 1 | Decision Tree | 0.949791 |

Here score is the mean score obtained by taking mean of scores in each split during k-fold cross-validation.



**After comparing the score of each model, the Random Forest model seems to be the most accurate.**

## Problems Faced

Model Selection - How to find the best parameters for the model ?

How to split Dataset - Which technique to use for splitting the data set ?

Applying the right ML methods, To adopt right strategies for e.g. cross-Validation.

How to give a Right conclusion.

Internet Facility in the Hostel.

## Conclusion

I have trained Four models for the given  Dataset and observed their performances. As we can see  Random Forest and SVM are performing really well and seeing experimentally RF is performing overall better . i think I can say Random Forest as my model for predicting the Test Dataset.

## Acknowledgements

UCI Machine Learning: https://archive.ics.uci.edu/ml/datasets/Zoo