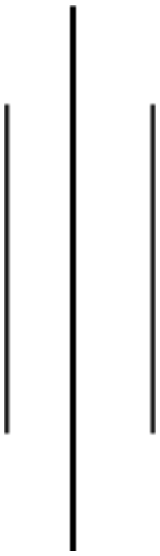
A
Report On

PROJECT - IV

For
Machine Learning Course
Fuse.ai Microdegree in AI



Submitted By:
Anish Bhusal
bhusal.anish12@gmail.com

Date of Submission: 4th May, 2020

ABSTRACT

The bee classification dataset consists of two bee species encoded as “1” and “0” where “1” referred to Bumblebee and “0” to Honey bee respectively. Hog features were extracted from images and used as features for training. The dataset was split into train and test sets with 75-25 split rates. Three models: SVM, Decision Tree and Random Forest were chosen for Grid Search with different params. After evaluating the prediction results on the test dataset, Random Forest algorithm provided better results with F1-Score of 0.46

1. DATA EXPLORATION

The given dataset consists of two kinds of bee species: Bumblebee and Honey Bee which are encoded as “1” and “0” respectively.

Total Image Labels	3969

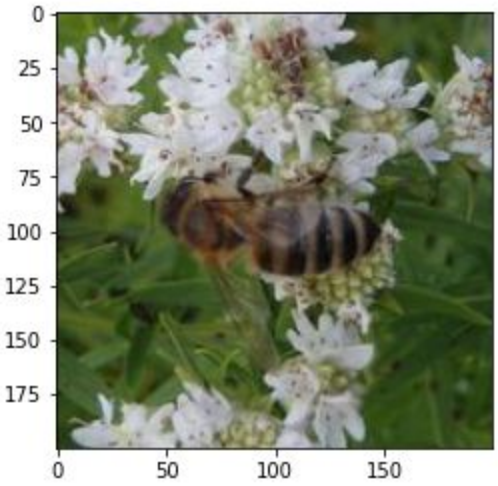


Fig.: Image showing Honey Bee species species

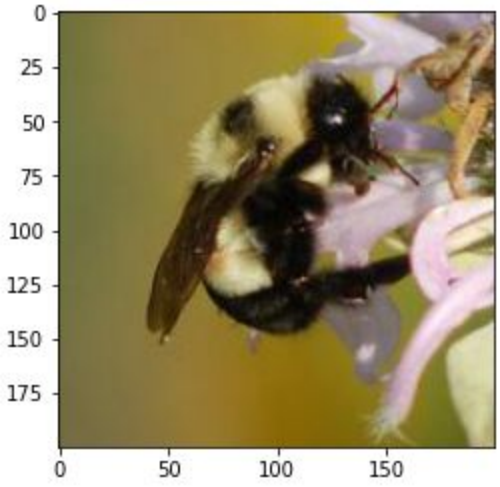


Fig.: Image showing Bumblebee

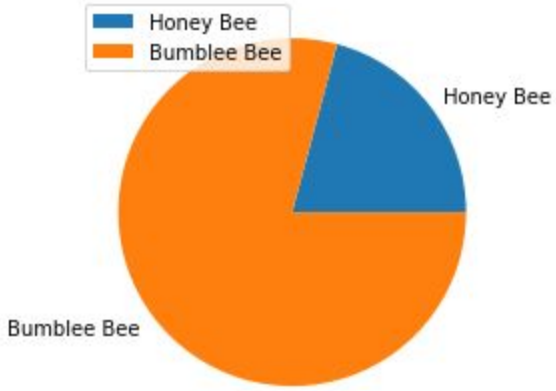


Fig.: Pie-chart showing composition of Honey Bee and Bumblebee in dataset

2. DATA PRE-PROCESSING AND FEATURE SELECTION

For each image in the dataset, its hog features were extracted and a dataset X,y was created containing hog features and image labels respectively.

Then the dataset was split in training and test set with following params:

test_size	0.25
random_state	27
stratify	y

3. GRID SEARCH

Three algorithms were chosen:

- SVM
- Decision Tree
- Random Forest

With following grid params :

Classifier	Params	Chosen Params
SVM	*'kernel':('linear','rbf'),* *'C':(1,20)*	*{* *C: 2,* *Kernel: rbf* *}*
Decision Tree	*'max_depth':(9,11,13),* *'min_samples_split':(4,6,8, 10)*	*{* *'max_depth': 9,* *'min_samples_split': 6* *}*
Random Forest	*'n_estimators':(5,8,11,13),* *'max_depth':(9, 11, 13),* *'min_samples_split': (4, 6, 8, 10),*	*{* *'max_depth': 9,* *'min_samples_split': 10,* *'n_estimators': 13* *}*

4. MODEL EVALUATION

Three models were evaluated with the help of test set and following classification results were obtained:

	Macro			
Model	**Precision**	**Recall**	**Accuracy**	**F1-Score**
SVM	0.9	0.51	0.80	0.47
Decision Tree	0.53	0.52	0.72	0.52
Random Forest	0.73	0.51	0.79	0.46

5. CONCLUSION

The given dataset is imbalanced so we need to use Recall metric while choosing the model. While evaluating the predictions from three models, Random Forest provided better classification results with 0.51 recall , 79% accuracy and f1-score of 0.46

Therefore, Random Forest was chosen to be the best model among three for bee classification.

All the code is available here.