Bee Classification Report

First of all, load labels.csv file into a dataframe called labels, where the index is the image name and the genus columns are the bee types. There are 3969 images of bees where 3142 are bumble bee images and 827 are honey bee images.

Honey bee as 0.0
Bumble bee as 1.0

Image of Honey bee

Image of Bumble bee

HOG Features Extraction

- Create image features into a single row
- loop over all of images
  In the resulting features matrix, rows correspond to images and columns to features.
  Features matrix shape :(3969, 8100)
- Rescale data using StandardScaler
  - Data that has a mean of 0 and unit variance.
- Reduce the number of features using Principal Component Analysis(PCA).
- Initially, 8100 features for each image
- By using PCA(n_components=500)
  Reduce number of features into 500

Feature matrix shape is: (3969, 8100)
PCA matrix shape is: (3969, 500)

Split into train and test sets

convert data into train and test sets.Use 75% of images as training data and remaining 25% of images as testing data

Train Model

- Linear support vector machine
- Decision Tree
- Random Forest

Grid search

Grid-search is the process of scanning the data to configure optimal parameters for a given model.

Model	grid_param	best_param
SVM	C: [1.0, 2.0, 4.0, 8.0]	C: 1.0

Decision Tree	max_depth: (3, 5, 7, 8,9, 11,13) min_samples_split: (2, 4, 6, 8, 10)	max_depth: 3 min_samples_split: 6
Random Forest	max_depth: (8,16) min_samples_split: (2, 4, 6, 8, 10) n_estimators:(10,50,100)	max_depth: 16, min_samples_split: 2 n_estimators: 100

Model Evaluation and comparison

Best score of SVM, Decision Tree, and Random Forest is given below:

Model	best_score
SVM	0.65
Decision Tree	0.78
Random Forest	0.79

Conclusion: Random Forest is best classifier

Final Model

```
import sklearn.metrics as metrics
model =RandomForestClassifier(max_depth=16,min_samples_split=2,n_estimators=100)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

         0.0       1.00      0.02      0.04       207
         1.0       0.79      1.00      0.89       786

    accuracy                           0.80       993
   macro avg       0.90      0.51      0.46       993
weighted avg       0.84      0.80      0.71       993
```

ROC Curve and AUC for Random Forest classifier