V M Yoga Darsa (21137073)
B.Tech
CSE AI and ML

RANDOM FOREST ON PENGUIN DATASET

Problem Statement: Predicting the species of the penguin

ABSTRACT:

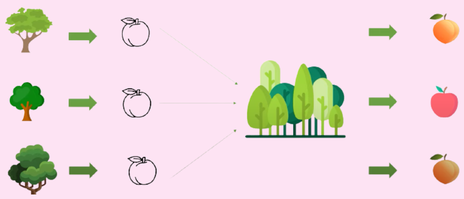
This presents a comprehensive analysis of the application of the Random Forest classification algorithm to the popular penguin dataset. Random Forest, a powerful ensemble learning technique, is employed to predict the species of penguins based on various morphological features such as bill length, bill depth, flipper length, and body mass.

INTRODUCTION TO RANDOM VARIABLE:

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.A random forest algorithm consists of many decision trees. Random forests typically perform better than decision trees because they combine the output of multiple decision trees to come up with a final prediction.

Types of Random Forest Classifier Models
1. Random forest classifier prediction for a classification problem:
f(x) = majority vote of all predicted classes over B trees
2. Random forest classifier prediction for a regression problem:
f(x) = sum of all subtree predictions divided over B trees

Example :
STAGE I : From the every decision tree the unknown fruit must be predicted



Here each tree plays a role of decision tree and from that random forest will classify into classes

STAGE II : Classify the unknow fruit based number of predictions

STAGE III : Predict based on the maximum values

SPLITTING METHOD: Gini Impurity | Gini Impurity | Information Gain | Entropy

FORMULA:

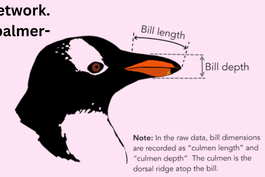
Impurity	Task	Formula
Gini impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$

f_i is the frequency of label i at a node and C is the number of unique labels.

ABOUT DATASET:

Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER , a member of the Long Term Ecological Research Network. Which is available in kaggle (https://www.kaggle.com/datasets/ashkhagan/palmer-penguins-datasetalternative-iris-dataset).

The Attributes are :i)species- Adelie,Chinstrap,Gentoo
ii)bill_length
iii)bill_depth
iv)body_mass
v)sex- male,Female



DATA PREPROCESSING

step I: Removing the null values

```
In [12]: 1 df.dropna(inplace=True)
```

Here we are removing the null values present in the raw dataset of penguin

BEFORE

```
In [9]: 1 df.isnull().sum()
Out[9]: species              0
        island               0
        bill_length_mm       2
        bill_depth_mm        2
        flipper_length_mm    2
        body_mass_g          2
        sex                 11
        dtype: int64
```

AFTER

```
In [13]: 1 df.isnull().sum()
Out[13]: species              0
         island               0
         bill_length_mm       0
         bill_depth_mm        0
         flipper_length_mm    0
         body_mass_g          0
         sex                  0
         dtype: int64
```

step II: Converting categorical data to numerical data

```
In [28]: 1 df['sex']=df['sex'].replace({"Male":1,"Female":0})
```

BEFORE

```
Out[27]: 0       Male
         1     Female
         2     Female
         4     Female
         5       Male
               ...
         338   Female
         340   Female
         341     Male
         342   Female
         343     Male
         Name: sex, Length: 333, dtype: object
```

AFTER

```
[30]: 0      1
      1      0
      2      0
      4      0
      5      1
            ..
      338    0
      340    0
      341    1
      342    0
      343    1
      Name: sex, Length: 333, dtype: int64
```

step III: Converting categorical data to numerical data and fomin a new column

```
In [25]: 1 island=pd.get_dummies(df['island'],drop_first = True)
         2 island.head()
```

Out[25]:

	Dream	Torgersen
0	0	1
1	0	1
2	0	1
4	0	1
5	0	1

step IV: Converting categorical data to numerical data

```
In [32]: 1 y=y.replace({'Adelie':0,'Chinstrap':1,'Gentoo':2})
         2 y.head()
Out[32]: 0    0
         1    0
         2    0
         4    0
         5    0
         Name: species, dtype: int64
```

Note :
These are data preprocessing technique used to form a preprocessed data

PREPROCESSED DATA:

```
In [62]: 1 new_data.head()
```

Out[62]:

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	Dream	Torgersen
0	39.1	18.7	181.0	3750.0	1	0	1
1	39.5	17.4	186.0	3800.0	0	0	1
2	40.3	18.0	195.0	3250.0	0	0	1

SPLITING DATA INTO TRAINING AND TESTING DATA:

```
In [44]: 1 from sklearn.model_selection import train_test_split
         2 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size
```

USING SPLITING METHOD ENTROPY AND GINI:

```
In [46]: 1 from sklearn.ensemble import RandomForestClassifier
         2 classifier=RandomForestClassifier(n_estimators=5,criterion='entropy',random_state=0)
         3 classifier.fit(x_train,y_train)
Out[46]: RandomForestClassifier(criterion='entropy', n_estimators=5, random_state=0)
```

```
In [58]: 1 from sklearn.ensemble import RandomForestClassifier
         2 classifier=RandomForestClassifier(n_estimators=8,criterion='gini',random_state=0)
         3 classifier.fit(x_train,y_train)
Out[58]: RandomForestClassifier(n_estimators=8, random_state=0)
```

ACCURACY:

```
In [60]: 1 accuracy_score(y_test,y_pred)
Out[60]: 0.98
```

CONCLUSION

In conclusion, the Random Forest algorithm is a powerful and versatile tool for predicting the species of penguins based on various input features. In the context of predicting penguin species, the Random Forest algorithm can be considered a reliable choice. By training the model on a dataset with features like bill length, bill depth, flipper length, and body mass, you can make accurate predictions about the species of penguins. However, it's crucial to keep in mind that model performance can vary depending on the quality and quantity of data, as well as the specific dataset used. Regular maintenance and retraining of the model may be necessary as new data becomes available. Additionally, as with any machine learning model, Random Forest's predictions should be used in conjunction with domain knowledge and expert input to ensure the most accurate species classification for penguins.

Reference

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9:2015.
D. Amaratunga, J. Cabrera, and Y.S. Lee. Enriched random forests. Bioinformatics, 24:2010–2014, 2008.
G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. Journal of Multivariate Analysis, 101:2499–2518, 2010.
G. Biau, F. Cerou, and A. Guyader. On the rate of convergence of the bagged nea´ rest neighbor estimate. Journal of Machine Learning Research, 11:687–712, 2010.

https://github.com/Yogadarsa/Randomforest.git