CHAPTER 4

Project -1 Machine Learning

Prediction Of Galactic Astral type Identification

In this project, we will work with train_dataset.csv dataset to develop a machine learning algorithm that predicts the galactic astral type. A model like this would be very valuable to predict one's galactic astral identification using alpha , delta , u ,g ,r ,I ,z ,run_id ,rerun_id ,cam_col ,field_id ,spec_obj_id ,redshift ,plate,MJD ,fiber_id.
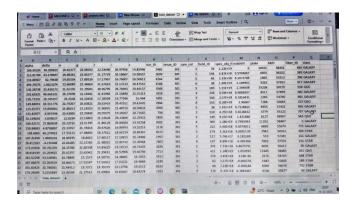
4.1 Problem Statement

Develop a model that has the capacity of predicting galactic astral identification by making use of the information provided in train Dataset

4.2 Dataset

The dataset used in this project consists of 16 variables.The main variable we are interested is 'class'. This variable predicts the galactic astral identification based on the inputs given in dataset

1. **Alpha**   Enter the alpha
2. **Delta**   Enter the delta
3. **u**   Enter the u
4. **g**   Enter the g
5. **r**   Enter the r
6. **i**   Enter the i
7. **z**   Enter the z
    .   .
    .   .
    .   .
16. **fiber_id**   Enter the fiber_id

The overview of the original dataset is shown in figure with its original features:

4.3 Algorithm –Random forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset".

4.4 Programming Steps

•This project requires us to predict the galactic astral identification based on the given input dataset.

•First, we read the given dataset using pandas function.

•Then we print the inputs and output from csv file.

•We initialize the model i.e., Random forest Algorithm.

•We further implement this using Django in order for better representation

Code:

```python
import pandas as pd
path="C:\\Users\\shara\\OneDrive\\Desktop\\Sanjana\\train_dataset.csv"
data=pd.read_csv(path)
print(data)
print(data.info())
print(data.shape)

inputs=data.drop('class',axis=1)
output=data['class']
print(inputs)
print(output)
```

```python
import sklearn
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(inputs,output,test_size=0.2)
print(x_train)
print(x_test)
print(y_train)
print(y_test)

import sklearn
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier(n_estimators=100)
model.fit(x_train,y_train)

y_pred=model.predict(x_test)
print(y_pred)
print(y_test)

import math
from sklearn.metrics import accuracy_score
acc=accuracy_score(y_test,y_pred)*100
print("acc:",math.ceil(acc))

result=model.predict([[240.0523131,34.25582134,24.01377,23.98592,22.53496,20.9
7958,19.85996,3965,301,3,70,1.23E+19,0,10921,58251,465]])
print(result)

result=model.predict([[149.8805331,45.51177587,20.79207,20.63821,20.61416,20.4
1011,20.43491,2964,301,6,163,8.20E+18,1.96667,7284,56683,223]])
print(result)
```

OUTPUT:

```
55005      STAR
21968     GALAXY
           ...
15731      STAR
12386     GALAXY
34412     GALAXY
12965      STAR
2665        QSO
Name: class, Length: 14000, dtype: object
acc: 98
['GALAXY']
['QSO']

C:\Users\shara\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but RandomF
orestClassifier was fitted with feature names
  warnings.warn(
C:\Users\shara\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but RandomF
orestClassifier was fitted with feature names
  warnings.warn(
```