CSE 6363 001 Machine Learning

Fall 2022

Project 1 Report

Instructor's Name: Dr. Dajiang Zhu

Student Name: Siva Srinivasa Sameer Miriyala

Student ID: 1002024871

Given Problem Statement:

The project's goal is to create a linear regression model for a given iris data for the prediction of species and the classification type of flowers. We need to train the data and to evaluate the efficiency of the model we perform cross validation with good accuracy.

Dataset:

This data has been extracted from (http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data)

The Iris Data set consists of 150 rows and 5 columns.

The 5 columns are - Sepal Length, Sepal Width, Petal Length, Petal Width and Species. These can be used as features and the last column can be our label. With these details we can form a matrix.

The sample data is as follows.

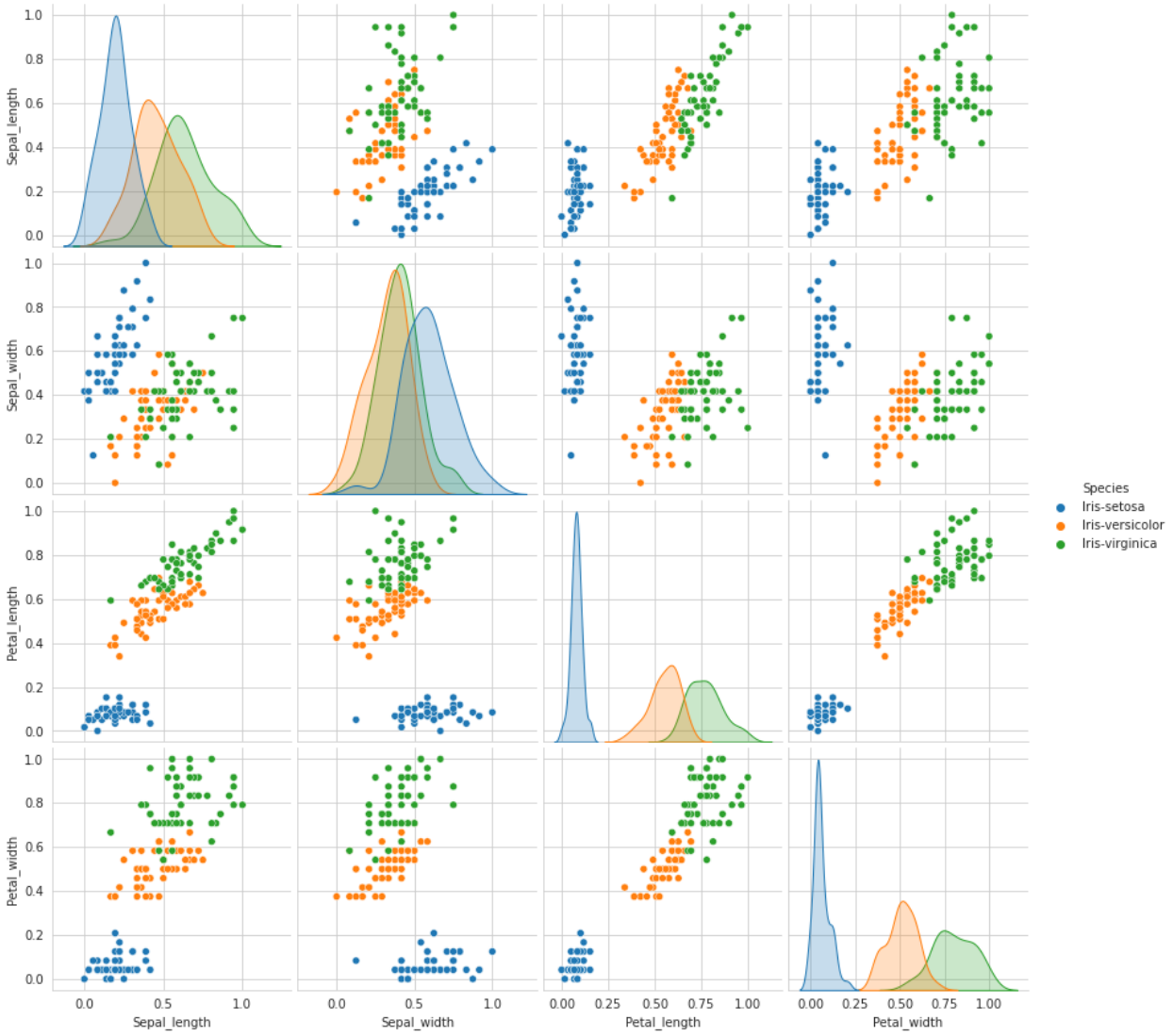
```
#Print any 5 sample items in the dataframe
print(iris_dataframe.sample(5))
```

```
      Sepal Length (cm)    Sepal Width (cm)    Petal Length (cm)   \
85                  6.0                 3.4                 4.5
31                  5.4                 3.4                 1.5
80                  5.5                 2.4                 3.8
149                 5.9                 3.0                 5.1
38                  4.4                 3.0                 1.3

      Petal Width (cm)     Flower Class
85                 1.6  Iris-versicolor
31                 0.4     Iris-setosa
80                 1.1  Iris-versicolor
149                1.8   Iris-virginica
38                 0.2     Iris-setosa
```

Data Visualization:

Scatterplot the different class values in the split that are previously given.



Implementation:

In order to obtain the desirable solution we use k-folds by creating the certain number of bin in cross validation. The beta mean usually generated by each of the beta value to gradually minimize overfitting.

Formula:

We use the below formula to obtain all the beta values.

$$\hat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$$

A is the matrix of all the values

A^T is the A Transpose

Y is the mapped values of flower classes

Approach:

The predict method creates expected Y values using the B values and the A values from the input matrix. The resulting float numbers are being rounded up to the nearest integer and any negative values are being converted to positive ones. Label Encoding is used to convert the Y values into a category number value that can be predicted. Since we are employing mathematical functions to anticipate new values, we need the species categories to be represented by a numerical number. I use my code to convert the Species column in my data into the category and use it to generate a new column named Species cat in order to store the label-encoded information.

```
       Sepal_length  Sepal_width  Petal_length  Petal_width  Species_cat
count    150.000000   150.000000    150.000000   150.000000   150.000000
mean       0.428704     0.439167      0.467571     0.457778     1.000000
std        0.230018     0.180664      0.299054     0.317984     0.819232
min        0.000000     0.000000      0.000000     0.000000     0.000000
25%        0.222222     0.333333      0.101695     0.083333     0.000000
50%        0.416667     0.416667      0.567797     0.500000     1.000000
75%        0.583333     0.541667      0.694915     0.708333     2.000000
max        1.000000     1.000000      1.000000     1.000000     2.000000
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Sepal_length  150 non-null    float64
 1   Sepal_width   150 non-null    float64
 2   Petal_length  150 non-null    float64
 3   Petal_width   150 non-null    float64
 4   Species       150 non-null    category
 5   Species_cat   150 non-null    int8
dtypes: category(1), float64(4), int8(1)
memory usage: 5.2 KB
None
/usr/local/lib/python3.7/dist-packages/seaborn/axisgrid.py:2076: UserWarning
  warnings.warn(msg, UserWarning)
```

The accuracy of the algorithm if the test-split is 0.3 and the k-fold bins are 6

```
For each bin, the generated values are
[[-1.6340657818113702 0.024553360465899288 3.2181377066356394
  0.3348527962805896]
 [0.27732004205383787 -0.38010157835923847 0.8067198735285787
  1.5078581739621235]
 [0.052332488710205016 -0.3138282438163908 2.023454930269722
  0.24610843631209967]
 [-1.4931786999517132 0.04058172614745828 1.6169021139366837
  1.9882757187229292]
 [0.20176394058299096 -0.32085662062636666 1.9959150942726955
  0.2429986060790289]
 [-0.910582307801447 -0.33476765542353626 0.5375201353309679
  2.5096753041003046]]
The Mean of values
[-0.5844017197029161 -0.21406983526869575 1.6997749756623814
 1.1382948392428458]
The dataset when performed Linear Regression has an accuracy of:
98.0%
```