Plotting for Exploratory data analysis (EDA)

Submitted for the partial fulfilment of the Degree

of

Bachelor of Technology

(Information Technology)

Submitted By:

Arshdeep Singh Ahuja

D3 IT - A1

University Roll:1507902

Submitted to:

Department of Information Technology

Guru Nanak Dev Engineering College

Ludhiana 141006

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely fortunate to have got this all along the completion of my project work. Whatever I have done is only due to such guidance and assistance and I would not forget to thank them.

I **ARSHDEEP SINGH AHUJA** the students of **GURU NANAK DEV ENGINEERING COLLEGE (**Information Technology), am extremely grateful to "**Piford**" for the confidence bestowed in us and entrusting our project entitled **"Exploratory data analysis".** .We express our gratitude to College Director **Dr. Sehijpal Singh** for arranging the summer training in good schedule.

We would also like to thank all the faculty members of **GNDEC** for their critical advice and guidance without which this project would not have been possible. We are thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of Department of Information Technology which helped us in successfully completing our project work. Also, we would like to extend our sincere regards to all the non-teaching staff of department of Information Technology for their timely support.

Table Of Contents

List of Figures

CHAPTER – 1

INTRODUCTION TO COMPANY – Piford Technologies

Piford Technologies is a "Software Development Company" with its development and training centre at IT Park, Mohali .The portfolio of services includes legacy application maintenance, large application development, e-strategy consulting and solutions. Piford Technologies offers service in social networking, e-commerce, real estate, e-learning and learning management system, daily deals and group buying applications, SAAS, CRM, ERP, Smartphone applications, desktop applications, migration, search engine creation. We have expertise across business domains and a long list of satisfied customers. Piford Technologies uses its strengths in technology, software, data mining, research & services, to create new revenue-generating opportunities for its customers & at the same time reducing the overheads, while enabling them to quickly deploy and better manage and direct their businesses.

Seal

Authorized Signatory

CHAPTER – 2

INTRODUCTION TO PROJECT

1. OVERVIEW

About our EDA, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics . EDA is done on The Iris flower data set. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species ". This project defines the data analysis of species of flower on the basis of traits like sepal Length, speal width and some of the common analysis like PDF's and CDF's are used for building Toy model.

1.1) Key Features

Detailed Information: This Project aim is to analysis of species of flowers and distinguish each other with one or the other feature. Along with this , every analysis technique also explains the toy prediction model which can be further deployed on a programming platform.

Ease to access: This demonstrate eye catching and crispy analysis which enables any user to understand it up to fullest. It contains simple snippest of code of each techniques along with its advantes and disadvantages.

User Reviews: This feature allows users to classify the different species of given flowers on the basis of their features. The user are required to have knowledge of graph interpretation and statistics to interpret the result and deploy as the final predicting model

1.2) Technologies Used

Python (ver-3.06)

Python is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, Python has a design philosophy that emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer lines of code than might be used in languages such as C++ or Java. It provides constructs that enable clear programming on both small and large scales.

i-python Notebook

1IPython Notebook is a web-based interactive computational environment for creating IPython notebooks. An IPython notebook is a JSON document containing an ordered list of input/output cells which can contain code, text, mathematics, plots and rich media. IPython notebooks can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through 'Download As' in the web interface and 'ipython nbconvert' in a shell. IPython notebooks frequently draw from SciPy stack libraries like NumPy and SciPy, often installed along with IPython from one of many Scientific Python distributions.

2) EXISTING SYSTEM

Current condition assessment methods are predominantly manual and time consuming Existing applications for cooking recipes are somewhat not eye appealing in User Interphase and most of them are incapable of saving the reviews of users. Most of the applications available do not have all the three information's regarding the cuisine, i.e., ingredients required, directions and notes. Earlier the apps don't had images of the dishes

3) OBJECTIVES OF PROJECT

The main objective of this research is to share the analysis of different species of iris flowers on the basis of Petal length, width, sepal width and length.

CHAPTER 3

DEVELOPMENT AND IMPLEMENTATION

1) **Importing The Data and Libraries**

To start exploring your data, you'll need to start by actually loading in your data**(in csv formt)**. You'll probably know this already, but thanks to the Pandas library, this becomes an easy task: you **import the package as pd**, following the convention, and you use **the read_csv()** function, to which you pass the URL in which the data can be found and a header argument. This last argument is one that you can use to make sure that your data is read in correctly: the first row of your data won't be interpreted as the column names of your DataFrame.

The following Libraries are imported From the anaconda package.

- Pandas
- Seaborn
- Matplotlib
- Numpy

2) **Describing The Data**

For example, you can use the **print()** function to get various summary statistics that exclude NaN values. Consider this example in which you describe the famous Iris dataset.

```
In [9]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        import numpy as np


        '''downlaod iris.csv from https://raw.githubusercontent.com/uiuc-cse/data-fa14/gh-pages/data/iris.csv'''
        #Load Iris.csv into a pandas dataFrame.
        iris = pd.read_csv("iris.csv")

In [10]: # (Q) how many data-points and features?
         print (iris.shape)

         (150, 5)
```

Figure 1

3) **Frame Column and Rows**

In every data set there are no of rows and columns. Each column can be called as variable or vector or class label . each tuple is called Data sets. From the given data set we can get the number of variables and data set and number of species of each flower by using **column() and value_count()** function.

```
In [11]: #(Q) What are the column names in our dataset?
         print (iris.columns)

         Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
                'species'],
               dtype='object')

In [12]: #(Q) How many data points for each class are present?
         #(or) How many flowers for each species are present?

         iris["species"].value_counts()
         # balanced-dataset vs imbalanced datasets
         #Iris is a balanced dataset as the number of data points for every class is 50.

Out[12]: setosa        50
         versicolor    50
         virginica     50
         Name: species, dtype: int64
```

Figure 2

We can analyze that of each species there exsist a equal number of flowers. So they are called BALANCED DATA SET. In the Given data set the classification problem is balanced in which class distribution is uniform among the classes.

4) 2-D plotting

In 2-D scatter plotting we build the plot of any of the two features.in this report variable and features are two interchangeable words. In the done analysis plotting is done within **Sepal Length** and **Sepal Width.**

Syntax of building this graph Includes simple **iris.plot()** function and **plt.show()** to print the graph on screen.

Disadvantage: We are not able to distinguish the 3 different species of flower from this graphs as legend are not present in this graph.

Conclusion: We can conclude that if sepal length of any flower is X then its corresponding sepal width is Y according to the given data set.

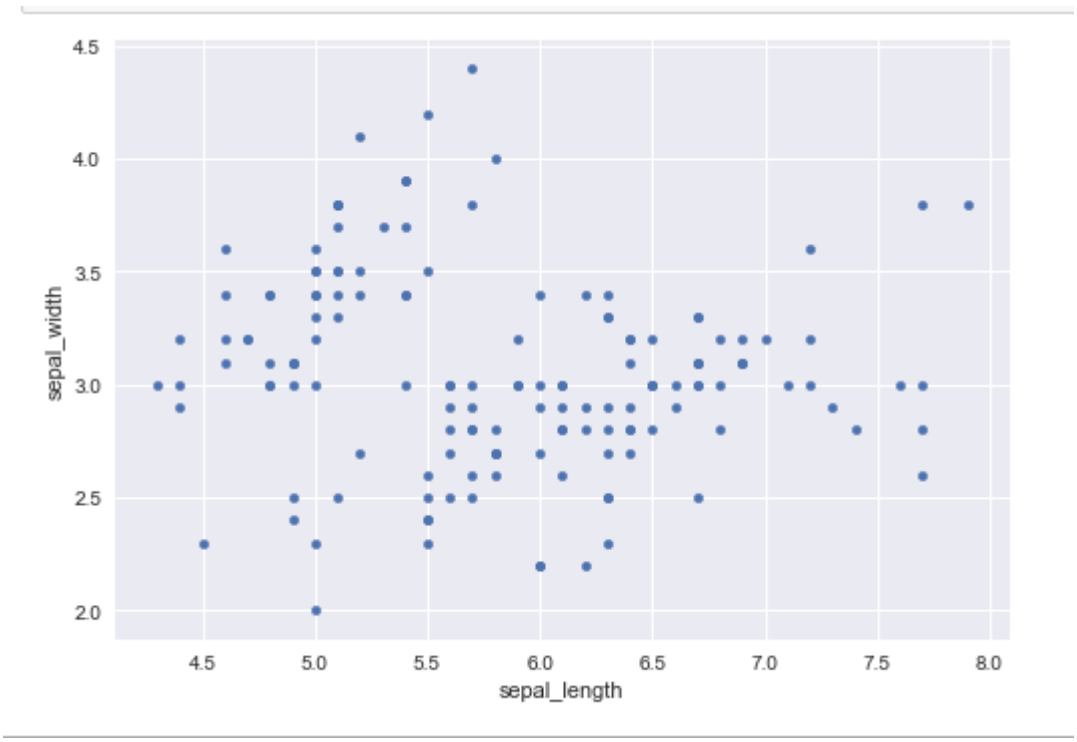

Figure 3

5) Using Legends

the legend is most often located on the right-hand side of the chart or graph and can sometimes be surrounded by a border. The legend is linked to the data being graphically displayed in the plot area of the chart. The legend is also known as a **Chart's Key**. We will be importing **Seaborn** library for Legends and **add_league()** function .

2-D scatter plot is shown below:

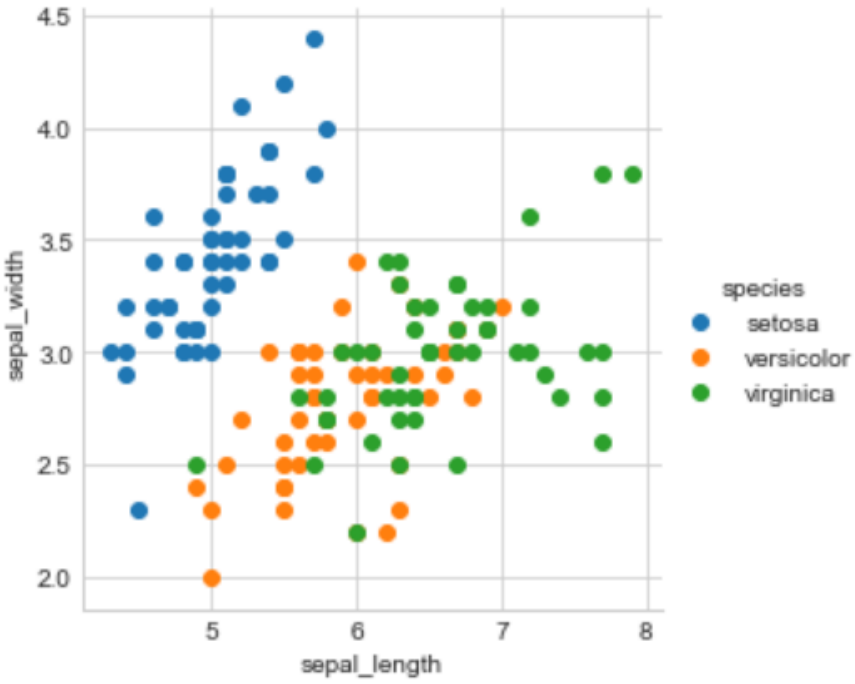

Figure 4

Conclusion:

- Notice that the blue points can be easily separated.
- From red and green by drawing a line.
- But red and green data points cannot be easily separated.
- Using **sepal_length** and **sepal_width** features, we can distinguish **Setosa flowers** from others.
- **Seperating Versicolor from Viginica** is much **harder** as they have considerable overlap.

6) **3-D Scatter Plot**

We can 3-D plot , but its main problem is that it needs lot of mouse interactions and to interpret the data

Another major problem is that it is unable to print 3-D plot on 2-D surface so we will be using PLOTLY tool . It is one of the best tool for ploting and analyzing 3 –D graphs. The link of graph is given : https://plot.ly/pandas/3d-scatter-plots/

Different view of 3-D scatter plot:

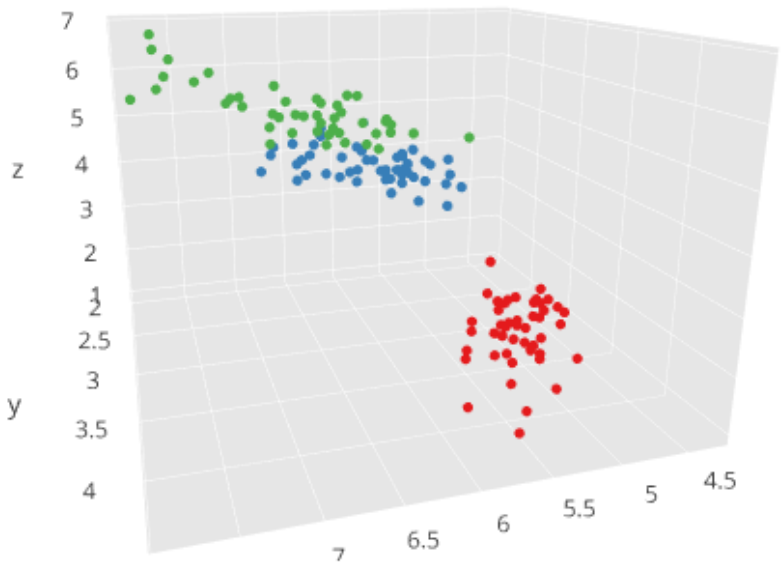

Figure 5

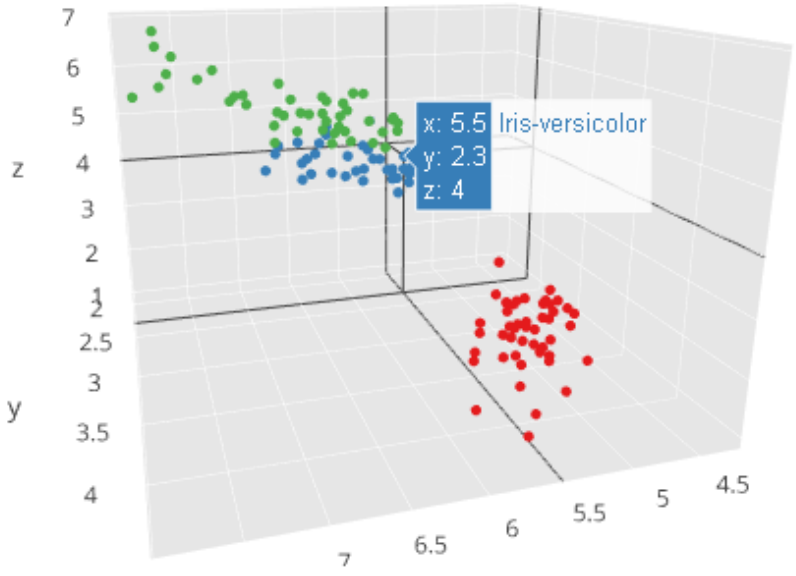

Figure 6

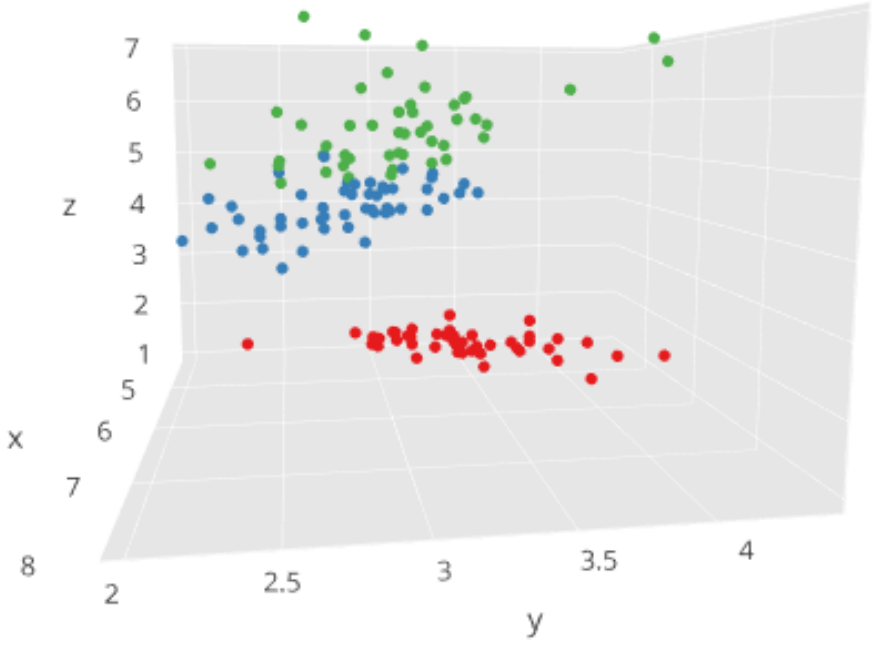

Figure 7

Conclusion:

- Using **sepal_length** and **sepal_width** features, we can distinguish **Setosa flowers** from others.
- Seperation can be done by iintroducing the **Plane**

Now question here arises that can we able to plot n-D plot where n>3? To support this answer we can say that humans are evolved in 3-D world not having abilities to analyse any dimension greater then 3-D.

7) Pair Plot

A “pairs plot” is also known as a scatterplot, in which one variable in the same data row is matched with another variable's value, like this: Pairs plots are just elaborations on this, showing all variables paired with all the other variables.

In this EDA we have 4 features to deal with i.e Sepal length(sl), Sepal width(sw), Petal length(pl), and Petal width(pw). So pair ploting is done in following pairs: {sw,pl} {sw,pw} {sl,sw} {pl,pw} {sl,pl} {sl,pw}. In geneal form we can say that having X variables in data set , number of pair plots can be formulated as $\mathbf{X_c 2.}$ Syntax used is **Sns.pairplot()** by importing **seaborn** library.

Limitation of Pair plot: Difficult to study pair plot of data set having large number of features.

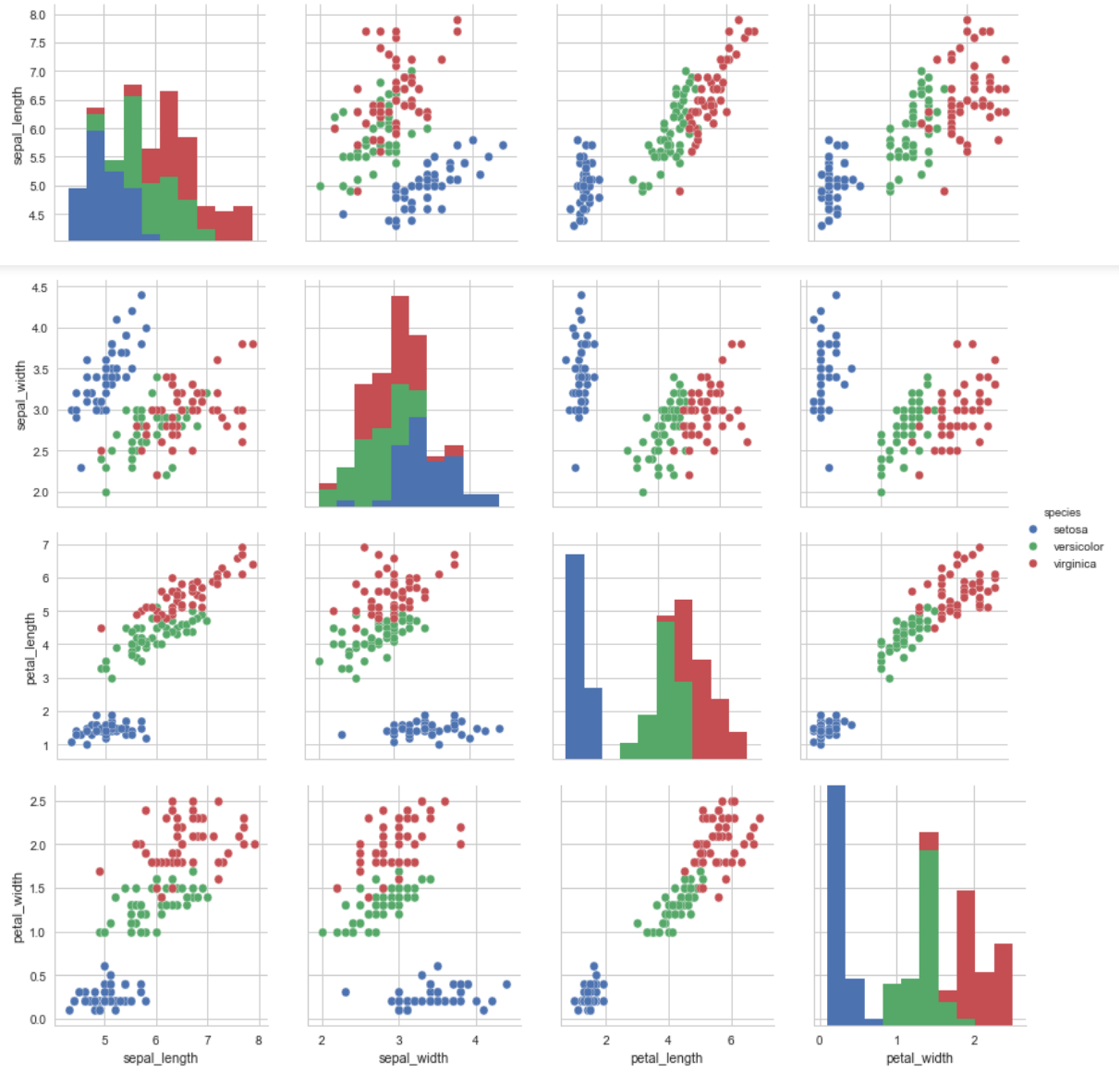

Figure 8

Conclusion:

- petal_length and petal_width are the most useful features to identify various flower types.
- While Setosa can be easily identified (linearly seperable), Virnica and Versicolor have some overlap (almost linearly seperable).
- We can find "lines" and "if-else" conditions to build a simple model to classify the flower types.

Predicted Model: *if* ***PL<=2 && Pw=>1*** *then flower type is* ***SETOSA.***

8) **1-D scatter Plot**

We can also use 1-D scatter plot in which only one feature is used . In my project I have ploted 1-D plot using feature **Petal length.**

Disadvantage:

Very hard to make sense as points are overlapping a lot. So we are not able to get the number of points we get on of any species.

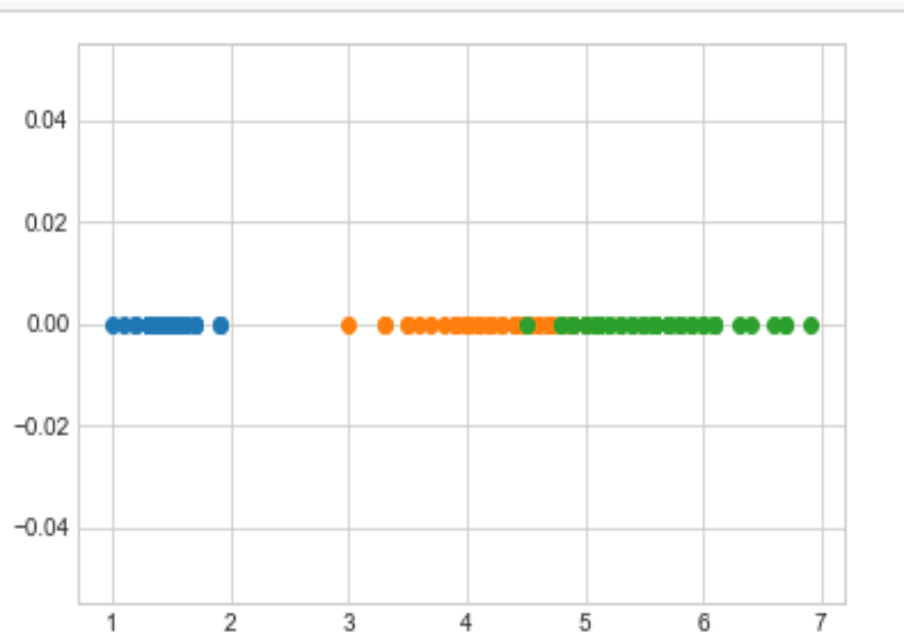

Figure 9

9) **PDF and Histograms**

In probability theory, a probability density function (PDF), or density of a continuous random variable, is a function, whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.[citation needed] In other words, while the absolute likelihood for a continuous random variable to take on any particular value is 0 (since there are an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample.

In a more precise sense, the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variable's PDF over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one.

The plot is shown below. The **Y-AXIS** of the graph shows **number of count of flowerS.**

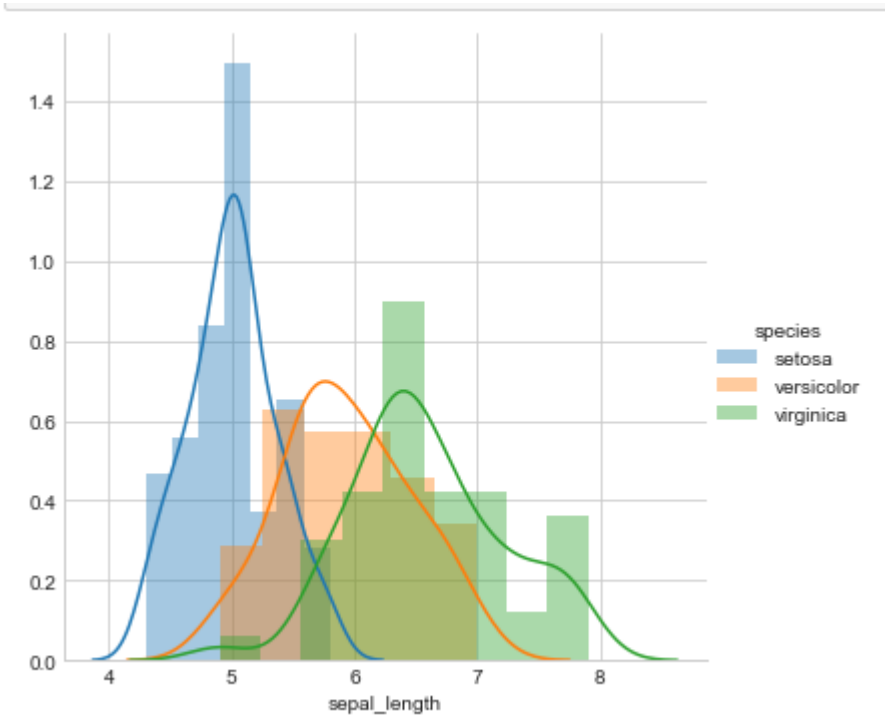

Figure 10

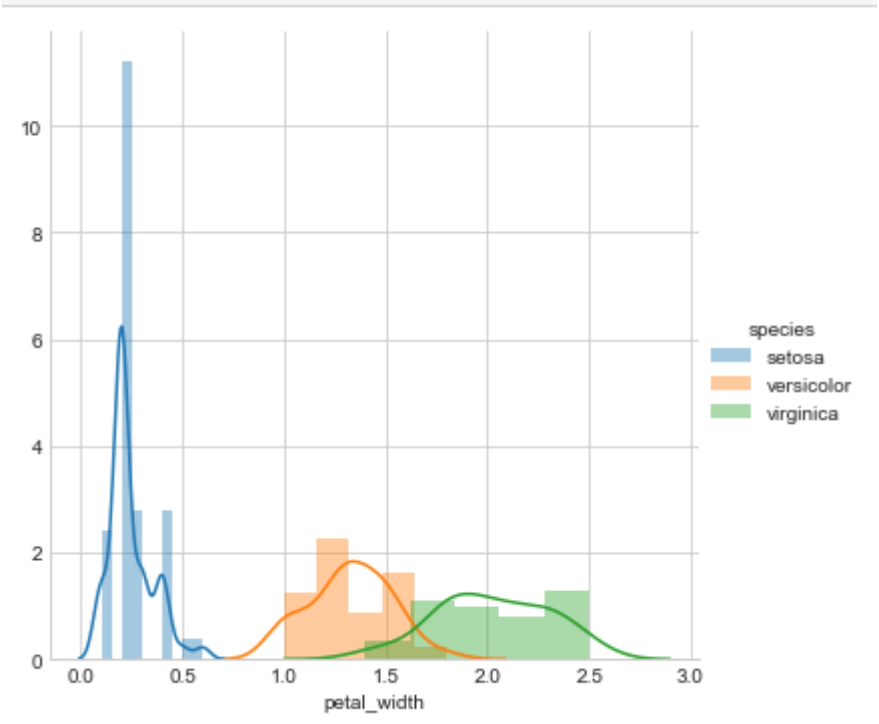

Figure 11

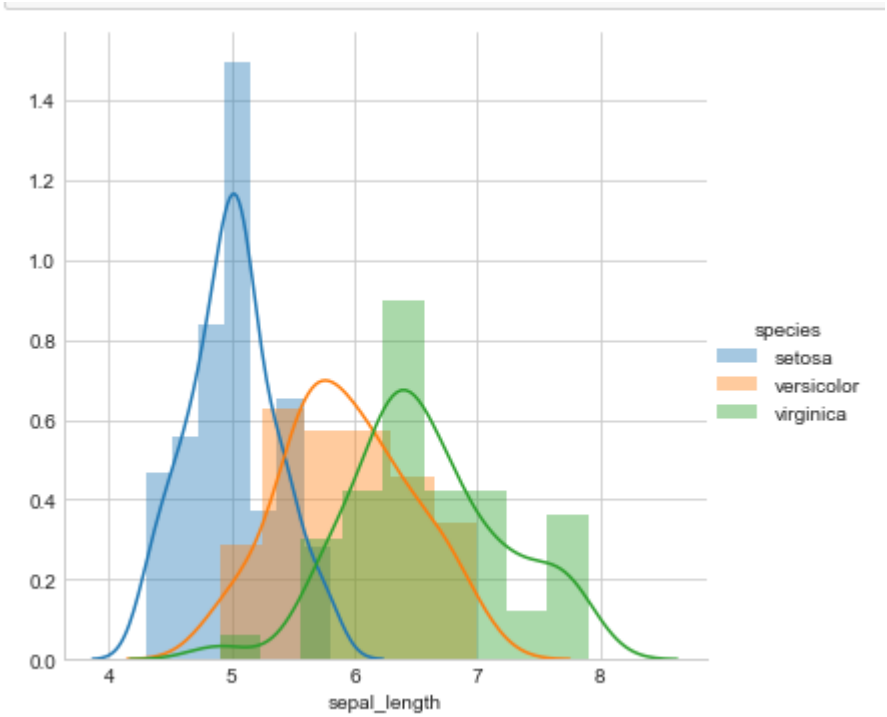

Figure 12

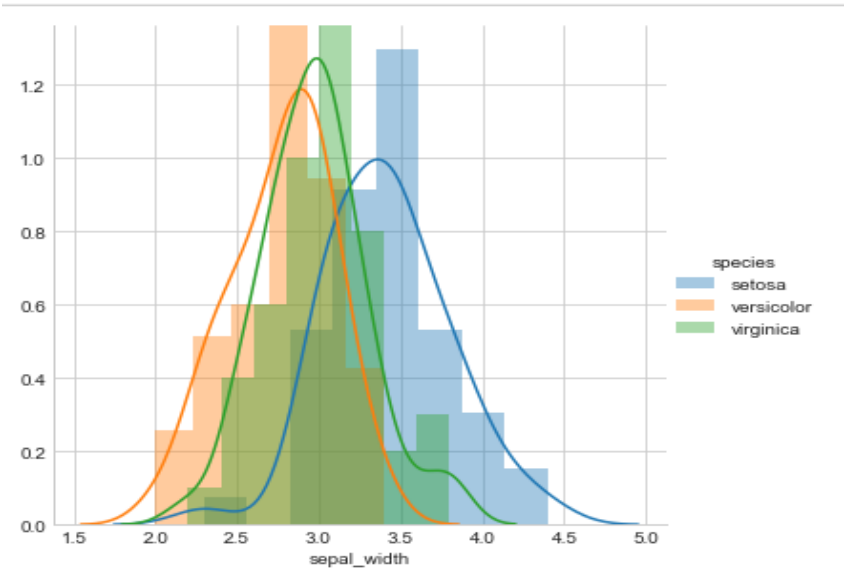

Figure 13

Predicted Model: if **PL <=2** , **SETOSA** else if **PL<4.7**, **VERSICOLOR** else **VIRGINICA**

Disadvantage of PDF: We Cannot say about percentage of versicolor points have a **petal_length** of less than X(numeric count).

10) Cumilative Density Function (CDF)

In probability theory and statistics, the cumulative distribution function (CDF, also cumulative density function) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x.

Syntax of the code include **cumsum()** of **Numpy** library : cdf=np.cumsum(pdf)

Advantage: It measures the accuracy of the model .

In our EDA we have plot the CDF of each species and graph is shown below

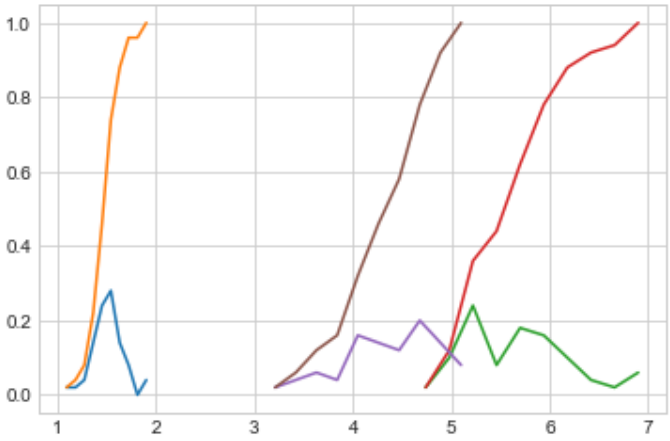

Figure 14

Predicted Model: If **PL>2** && **PL<=5** – it is **versicolor** . If **PL>2** && **PL>5 –** it is **verginica.**

Accuracy of this model is predicted as 95%

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

1) CONCLUSION

As compared to the older models , the new technique used in improving the dataset prediction in real time.

2) FUTURE SCOPE

Future scope of this can be made very useful if this analysis is in further state of the art – Machine Learning. This model can be used to build various ML models predicting the speciesof various flowers. Not only bounded to one application but this state of the art EDA can be used to explore number of huge data sets for creating crispy and beautiful results and conclusion from that.