Nexus

Project – 1

Submitted by – Abhinav Adarsh Email – adarsh061997@gmail.com

Approach, methodologies, and any challenges faced during the data science task and EDA and clear explanations for the choices made.

- The project involves exploratory data analysis (EDA) on the Iris dataset, which includes features such as Sepal Length Cm, Sepal Width Cm, Petal Length Cm, Petal Width Cm, and the target variable Species. Data science task involves making of ML model by deploying suitable ML algorithm.
- For gaining insights about the dataset I performed various tasks to observe the data set precisely and understood the relationships between the input and output data.
    - ➢ Uploaded the dataset using pandas lib of python into Data variable.
    - ➢ Observe the input parameters and output data which is basically a classification problem of Iris flower.
    - ➢ Further, I checked missing data and found that there is no missing data in dataset.
    - ➢ For observing the variation of input features vs count, I plotted Histogram of all the input parameters and observed that for sepal length the distribution is somehow like equal distribution, for sepal width its looks like similar to normal distribution, for petal length and petal width its distributed in three segment.
    - ➢ For observing how the input features varies with each type of Iris flower , I plotted Scatter plot between input features and the types of Iris flower and observed that for sepal length increases with Iris-setosa , Iris-versicolor, Iris-virginica respectively. Similar observation were also made with other features which can be observed in the plot.
    - ➢ For observing the how the data points in particular type of Iris flower spread in each input features, I plotted box plot for each features with the types of Iris flower. For also finding the outliers in the dataset I used the box plot. Although there is very less number of outliers and also they do not affect that much for learning of model so I ignore the outliers data.
    - ➢ Further I plotted the correlation matrix with species and also with the different type of Iris Flower individual with the input features. With the correlation matrix I concluded about the most important feature or that mostly affect the learning of model. Petal width with 0.96 correlation factor with species then petal length then sepal length and finally sepal width with opposite nature. This all can be visualised in the plotted graph in the code.

Based on the above points or insights that I find out from the data set I considered Decision tree algorithm would be better choice over the logistic regression as we can classify easily them based on splitting which can be easily observed in the scatter plot where petal width and petal length can able to classify that the flower belongs to which species, although when I trained the model on both algorithm I found both the model stand out with accuracy of 100% , precision = 1 and Recall = 1. I have only show the one model in the code that I have uploaded on GitHub. Further I also check the performance of model on test dataset and found that it classify with 100% accuracy.