Dataset Classification Project

Introduction: I had done this project for 'Artificial Intelligence' course offered by MyCaptain. The project contains brief introduction and code of 5 machine learning algorithm implanted on a single dataset, to check for the most accurate algorithm for classification of that particular dataset.

Machine Learning Algorithms : This project contains 5 machine learning algorithms given below:

1. **Linear Regression :** Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a sloped straight line representing the relationship between the variables.
2. **Linear Discriminant Analysis :** Linear Discriminant Analysis or LDA is a dimensionality reduction technique. It is used as a pre-processing step in machine learning and applications of pattern classification. The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs.
3. **K-Nearest Neighbours :** K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
4. **Naive Bayesian Classification:** Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
5. **Support Vector Machine :** A support vector machine is a selective classifier formally defined by dividing the hyperplane. Given labeled training data the algorithm outputs best hyperplane which classified new examples. In two-dimensional space, this hyperplane is a line splitting a plane into two parts where each class lies on either side. The intention of the support vector machine algorithm is to find a hyperplane in an N-dimensional space that separately classifies the data points.

Dataset : I have used the iris flower datasets for this project. This is also known as hello world dataset for machine learning beginners.

Source : https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv

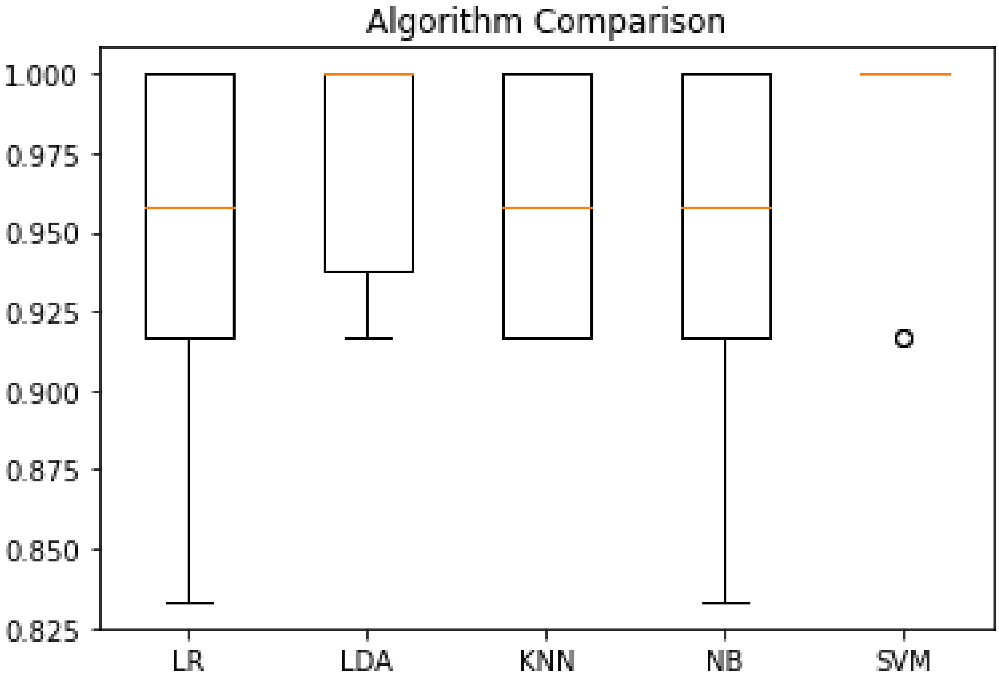
Shape : 150 rows and 5 columns

Accuracy :

1. **Linear Regression :** 0.950000

2. Linear Discriminant Analysis : 0.975000 (0.038188)
3. K-Nearest Neighbours : 0.958333 (0.041667)
4. Naive Bayesian Classification: 0.950000 (0.055277)
5. Support Vector Machine : 0.983333 (0.033333)

Algorithm Comparison :



Conclusion : It turns out that the support vector machine does a really great job at classifying iris flowers with accuracy of 98.33 % (test data) and 96.66% (validation data).