ISSN No:-2456-2165

Analysis of Different Classification Algorithms Used for the Classification on Three Species of *Iris* Limniris (Tausch) Spach Dataset

Prannov Jamadagni
School of Computing and Information Technology, Manipal University Jaipur

Abstract:- Iris is a flowering plant having 5-6 sepals which is a characteristic feature of classification of plant species. To determine the species of this genus the number of sepals are one of the factors and this work provides in an easier way to classify the *Iris* species. Hence, this paper explores the analysis of commonly used machine learning supervised classification algorithms on classifying and predicting three *Iris* flowering plant species i.e., *Iris setosa, Iris virginica, Iris versicolor* from the iris flower dataset present in UCI Machine Learning Repository which was compiled by Ronald Fisher. The dataset contains data of sepal length, sepal width, petal length, and petal width which is used for predicting the required species. Classification algorithms are a subset of supervised learning. Support Vector Machines, Decision Tree Classifier, and Logisitic Regression are the algorithms used in this paper for the purpose. The dataset is analyzed and preprocessed before fitting the algorithms for the prediction using scikit learn library and the data is analyzed using Python language. Machine Learning libraries for python, which include, pandas, numpy, matplotlib, and seaborn are used. The environment used for this project is Google Collaboratory. The parameters are adjusted for the dataset requirements. Finally, the three algorithms are evaluated based on their accuracy score and confusion matrix, where the SVM showed higher accuracy compared to that of the other two algorithms under these parameters. The significance of the prediction helps in predicting the species as well as eliminating the source of human error in separating different species. This work may contribute to prediction of more species of different genera. This paper comes under the theme: Life Sciences, Biomedical Sciences and Biotechnological aspects.

Keywords:- *Iris, Classification Algorithm, SVM, Logistic Regression, Decision Tree Classifier, Machine Learning.*

I. INTRODUCTION

Iris, the biggest variety of the family Iridaceae is almost 300 species of flowering plants with violet flowers (ref 1). The inflorescences on the peduncle are in the shape of a fan and contain one or more six-lobbed flowers. The three sepals spread downwards and expand from the narrow base. They are referred to as “falls”. On the other hand, the petals stand upright, behind the sepal bases. These are referred to as “standards” (ref 2). Indeed, the plants are classified depending on the flower features like the number of petals, sepals, and stamens, and so forth angiosperms are restricted with 5-6 sepals (ref 4). In this paper, three species have been taken for the examination of the classification machine learning algorithms which helps in a simpler method to the group. The iris dataset is a specific dataset compiled by Ronald Fisher, a biologist in the 1930s (ref 11). Henceforth, these calculations are utilized to classify the species (ref 3). It is, in this way, executed with the goal of performing classification analysis on the Iris dataset with three supervised algorithms and compare the accuracy.

II. METHODOLOGY

The iris dataset used for the analysis of classification algorithms is from the Kaggle and it contains 150 data points of three species of the iris flower (50 for each species). The Iris flower classifies with the help of features, namely sepal length, sepal width, petal length, and petal width (ref 3). The dataset is divided into a training set and a testing set. With the training set, the machine distinguishes the collection into three class values. The machine is then trained with the provided labels and features from the training set. Machine Learning supervised algorithms help the computer to learn the data. These algorithms then check on the testing set and predict the label (species of the iris) (ref 5).

Additionally, the co-relation between sepal length and sepal width; petal length, and petal width also plays an important role in predicting the species (figure 1).

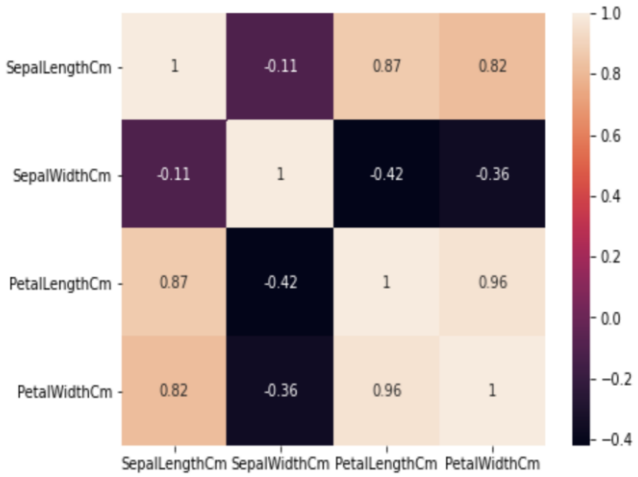

Fig 1. Correlation graph

It is observed that the co-relation between sepal length and sepal width is high compared to the latter. The Scikit Learn tool is used for implementation purposes. The three following algorithms used in this analysis are:

- SVM
- Decision tree classifier
- Logistic regression

The algorithms are analyzed based on the accuracy score and the f1 score (ref 9). SVM is a set of supervised learning algorithms which includes classification and regression (ref 6). Logistic regression is another approach for classification, which uses a sigmoid function. (ref 7) SVM and Logistic Regression are mainly used for binary classification. Whereas, Decision Tree classifier is used both for binary and multi-variable classification (ref 8). There are different parameters inside these algorithms that help in increasing the accuracy score on the testing set. The accuracy score can be interpreted in different ways with distinctive parameters (ref 9). which include:

- random state - used in splitting the data to both training and testing set.
- parameters specific to different algorithms.

Firstly, the dataset is computed to a CSV file as an input. Secondly, the data is split into a training set and a testing set. The test set size for this analysis is kept as 0.3 as it yields a better accuracy score over other test set sizes (ref 10). The random state is set to 5. Since the data set is of 150 sizes, a random state of 5 is a better estimate than other numbers. Then, the algorithms are passed to the dataset and predicted using the provided functions. Finally, the accuracy score of each algorithm is analyzed.

III. RESULTS AND DISCUSSION

After performing three classification algorithms on the testing set, few inferences can be made to understand the best algorithms to predict the species of the iris plant.

- SVM predicts the species with 99% accuracy. The true positives and the true negatives have been predicted accurately by this algorithm. Though this algorithm is mainly used for binary classification, the accuracy score is good.
- The decision tree classifier predicts the species with 97% accuracy. Since the difference between sepal length and sepal width is minimum, the accuracy score is less than that of SVM.
- The logistic regression predicts the species with 96.7% accuracy. This is a regression algorithm used for classification purposes. Therefore, the accuracy score is the least compared to the other classification algorithms (table 1).

Table 1. Training dataset with five different features and label

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

The relation of iris setosa compared to sepal and petal dimensions can be easily classified as observed from the graph. But the petal and sepal dimensions are mixed for iris Versicolor and iris virginica. The real accuracy is tested between these two species. The parameters of the algorithms help in classifying these mixed-species (figure 2)(figure 3).

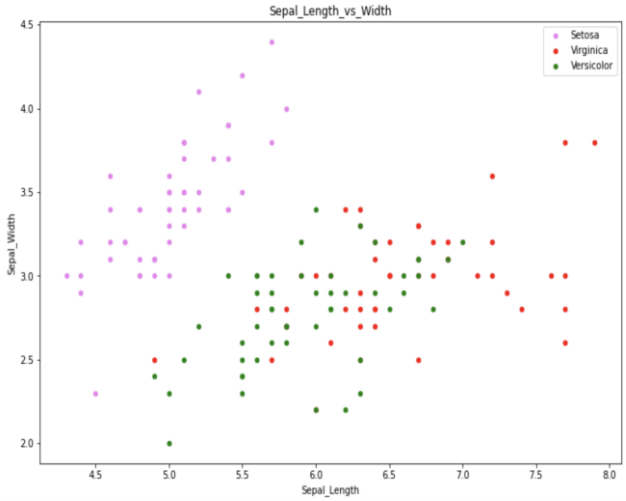

Fig 2. The relation between sepal length and sepal width

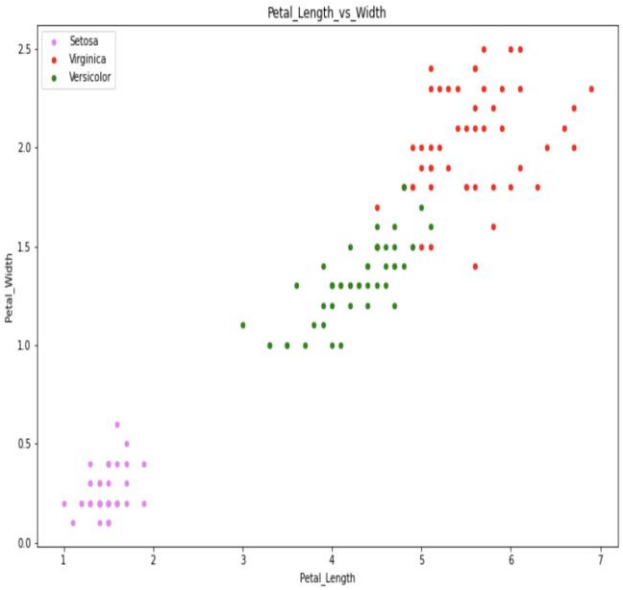

Fig 3. The relation between petal length and petal width

IV. CONCLUSION

Thus, using these common supervised learning classification algorithms: SVM, Decision tree classifier, and logistic regression, Support Vector Machines provide accurate results compared to the other algorithms. There are other factors that can change the accuracy score and can lead to poor scores for several algorithms. The parameters used here are ideal for the size of the dataset and can be changed, but the accuracy score changes with the change in parameters. Therefore, it can be concluded that the SVM is a better classifier than the other two under these parameters.

REFERENCES

[1]. "WCSP: Iris". *World Checklist of Selected Plant Families*. Retrieved 2 June 2014.
[2]. ^ "Iris". Pacific Bulb Society. 2011-11-26. Retrieved 2012-03-03.
[3]. www.kaggle.com
[4]. R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems" (PDF). *Annals of Eugenics*. **7** (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.
[5]. https://scikit-learn.org/stable/
[6]. https://scikit-learn.org/stable/modules/svm.html
[7]. https://scikit-learn.org/stable/modules/tree.html
[8]. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
[9]. https://scikit-learn.org/stable/modules/cross_validation.html
[10]. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
[11]. Iris dataset UCI Machine Learning Repository