Performance Evaluation of Supervised Learning for Iris Flower Species

Shivam Vatshayan
[1]Department of Computer Science and Technology, Galgotias University, Greater Noida, U.P. - 201310, India

***Abstract*- Machine Learning is a field of computer science pro-vides the ability to learn without programming and the program explicitly. Supervised Machine Learning (SML) is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future in-stances. Classification and Regression is a supervised learning in which the response is categorical that is its values are in finite Discrete and continuous set. To sim-ply the problem of classification, scikit learn tools has been used. This paper focuses on IRIS flower classification using Machine Learning with scikit tools. In this paper we will train the machine learning model with the given Iris Dataset and Analysis the performance and accuracy of Iris with Supervised Learning Algorithms**

1.INTRODUCTION

The Machine Learning is the sub field of computer science, ac-cording to Arthur Samuel in 1959 told “computers are having the ability to learn without being explicitly programmed”. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicitly algorithms with good performance is difficult or unfeasible; example applications include email filtering, detection of network intruders, learning to rank and computer vision. Machine learning focuses on the development of computer programs that can teach them-selves to grow and change when exposed to new data. It is a research field at the intersection of statistics, artificial intelli-gence and computer science and is also known as predictive analytics or statistical learning. There are two main categories of Machine learning. They are Supervised and Unsupervised learning and here in this, the paper focuses on supervised learning. Supervised learning is a task of conclude a function from labeled training data. The training data consists of set of training examples. In supervised learning, each example is a pair of an input object and desired output value. A supervised learning algorithm analyze the training data and produces an inferred function, which can be used for mapping new examples. Super-vised learning problems can be further grouped into regression and classification problems. Classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. Regression problem is when the output variable is a real value, such as “dollars” or “weight”. In this paper a novel method for Identification of Iris flower species is presented. It works in two phases, namely training and testing phase. During training the training dataset are loaded into Machine Learning Model and Labels are assigned. Further the predictive model, predicts to which species the Iris flower be-longs to. Hence, the expected Iris species is labeled. This paper focuses on IRIS flower classification using Machine Learning with scikit tools. The problem statement concerns the identification of IRIS flower species on the basic of flower attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS flower and how the prediction was made from analyzing the pattern to form the class of IRIS flower. In this paper we train the Machine Learning Model with Iris Flower data and when unseen data is discovered the predictive model predicts the accuracy of species using what it has learn from trained data.[1]

2. RELATED WORK

Many methods have been presented for Identification of Iris Flower Species. Every method employs different strategy. Re-view of some prominent solutions is presented.

The methodology for Iris Flower Species System is described [2]. In this work, IRIS flower classification using Neural Network. The problem concerns the identification of IRIS flower species on the basis of flower attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS flower and how the prediction was made from analyzing the pattern to form the class of IRIS flower. By using this pattern and classification, in future upcoming years the unknown data can be predicted more precisely. Artificial neural networks have been successfully applied to problems in pattern classification, function approximations, optimization, and associative memories. In this work, Multilayer feed-forward networks are trained using back propagation learning algorithm.

Letter Optics Letters 2

The model for Iris Flower Species System is described [3]. Exist-ing iris flower dataset is preloaded in MATLAB and is used for clustering into three different species. The dataset is clustered using the k-means algorithm and neural network clustering tool in MATLAB. Neural network clustering tool is mainly used for clustering large data set without any supervision. It is also used for pattern recognition, feature extraction, vector quantization, image segmentation, function approximation, and data mining. Results/Findings: The results include the clustered iris dataset into three species without any supervision.

The model for Iris Flower Species System is described [4]. The proposed method is applied on Iris data sets and classifies the dataset into four classes. In this case, the network could select the good features and extract a small but adequate set of rules for the classification task. For Class one data set we obtained zero misclassification on test sets and for all other data sets the results obtained are comparable to the results reported in the literature.

3. IRIS FLOWER

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris Flower of three related species. Two of the three species were collected in Gaspe Peninsula all from the same pasture, and picked on the same day and measured at the same time by the same person with same apparatus. The data set consists of 50 samples from each of three species of Iris that is 1) Iris Setosa 2) Iris Virginica

3)Iris Versicolor. Four features were measured from each sample. They are 1) Sepal Length 2) Sepal Width 3) Petal Length

4)Petal Width. All these four parameters are measured in Centimeters. Based on the combination of these four features, the species among three can be predicted 3.3 Problem Definition To design and implement the Identification of Iris Flower species using machine learning using Python and the tool ScikitLearn. 3.4 Work Carried Out Data collection: Various datasets of Iris Flower are collected. There are totally 150 datasets belonging to three different species of Iris Flower that is Setosa, Versicolor and Virginica. Literature survey: Studied various papers related to proposed work. Algorithms developed

1. A K-Nearest Neighbor Algorithm to predict the species of Iris Flower.
2. A Logistic Regression Algorithm to predict the species of Iris Flower.
3. Support vector machine Algorithm to predict the species of Iris Flower.
4. Metric Algorithm to predict the species of Iris Flower.
5. Decision tree Classifier to predict the species.[1]

Figures and Tables should be labelled and referenced in the standard way using the \label{} and \ref{} commands.

4. IMPLEMENTATION OF SUPERVISED LEARNING AL-GORITHM

Scikit tools and libraries are used for classification of dataset and Implementation of Machine learning Algorithms by using

Iris species

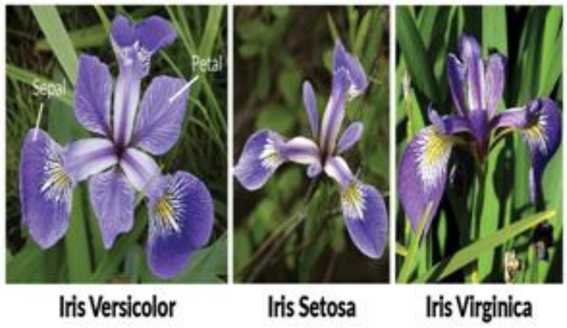


Fig. 1. Iris species image, where each image is assigned to one of three species of Iris flower

python programming language. You can download iris dataset from https://www.kaggle.com/arshid/iris-flower-dataset and https://archive.ics.uci.edu/ml/datasets/iris website.

A. Algorithm Selection

The selection of algorithm for achieving good results is an impor-tant step. The algorithm evaluation is mostly judge by prediction accuracy. The classifier's (Algorithm) evaluation is most often based on prediction accuracy and it can be measured by given below formula

There are number of methods which are being used by different

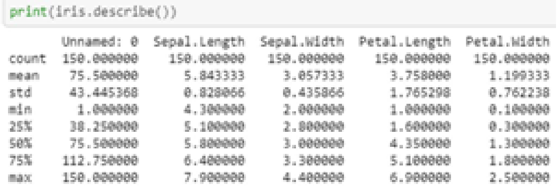
$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

Researchers to calculate classifier's accuracy. Some researcher's splits the training set in such a way that, two-thirds retain for training and the other third for estimating performance. Cross-Validation (CV) or Rotation Estimation is another approach. CV provides a way to make a better use of the available sample.

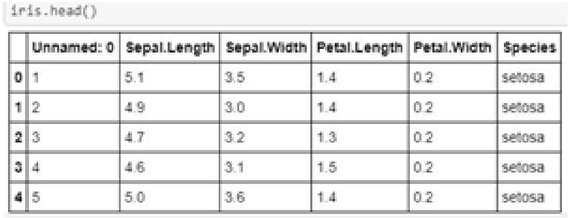
The comparison between supervised ML methods can be done through to perform statistical comparisons of the accuracies of trained classifiers on specific datasets.

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
iris=pd.readcsv(”iris.csv”)

print(iris.describe())
articlegraphicx



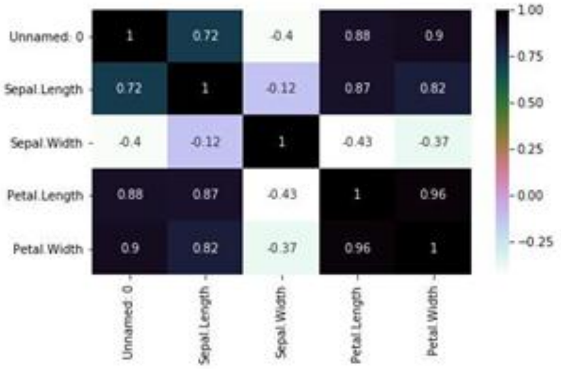
```
print(iris.describe())

       Unnamed: 0  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
count  150.000000    150.000000   150.000000    150.000000   150.000000
mean    75.500000      5.843333     3.057333      3.758000     1.199333
std     43.445368      0.828066     0.435866      1.765298     0.762238
min      1.000000      4.300000     2.000000      1.000000     0.100000
25%     38.250000      5.100000     2.800000      1.600000     0.300000
50%     75.500000      5.800000     3.000000      4.350000     1.300000
75%    112.750000      6.400000     3.300000      5.100000     1.800000
max    150.000000      7.900000     4.400000      6.900000     2.500000
```



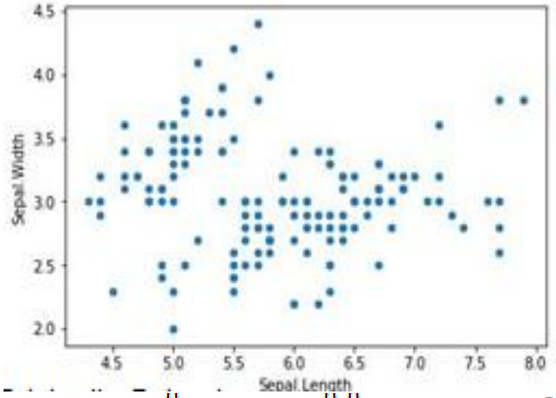
```
iris.head()
```

	Unnamed: 0	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	1	5.1	3.5	1.4	0.2	setosa
1	2	4.9	3.0	1.4	0.2	setosa
2	3	4.7	3.2	1.3	0.2	setosa
3	4	4.6	3.1	1.5	0.2	setosa
4	5	5.0	3.6	1.4	0.2	setosa

plt.figure(figsize=(6,4)) sns.heatmap(iris.corr(), annot=True, cmap='cubehelixr0)plt.show()

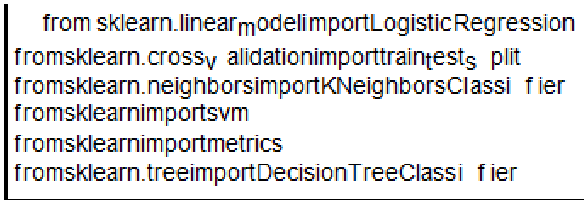


iris.plot(kind="scatter", x="Sepal.Length", y="Sepal.Width")



$$train_X = train[[{}^{U}Sepal.Length^{U}, {}^{U}\ Sepal.Width^{0}, {}^{0}Petal.Length^{0}, {}^{0}Petal.Width^{0}]]\ train_y = train.Species$$

$$test_X = test[[{}^{0}\ Sepal.Length^{0}, {}^{0}\ Sepal.Width^{0}, {}^{0}Petal.Length^{0}, {}^{0}Petal.Width^{0}]]\ test_y = test.Species$$



```
   from sklearn.linearmodelimportLogisticRegression
fromsklearn.crossv alidationimporttraintests plit
fromsklearn.neighborsimportKNeighborsClassi f ier
fromsklearnimportsvm
fromsklearnimportmetrics
fromsklearn.treeimportDecisionTreeClassi f ier
```

B. support vector machine (SVM)

A support vector machine (SVM) is a set of related supervised learning methods used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, the SVM training algorithm builds a model that predicts whether a new example falls into which specific category. Intuitively, SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [7] [8] [9]. More correctly, SVM constructs a hyper plane or a set of hyper planes in a high dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [7] [8] [9].

Currently, SVM is widely used in object detection and recogni-tion, content-based image retrieval, text recognition, biometrics, speech recognition, speaker identification, benchmarking time-series prediction tests. Using SVM in text classification is proposed by [10], and subsequently used in [11] [12].

SVM classifier gives around 97dataset with this classifier.

SVCmodel = svm.SVC()

model.fit(train$_X$, train$_y$)

prediction = model.predict(test$_X$)

print(^{0} Accuracy :^{0} , metrics.accuracy$_s$core(prediction, test$_y$))

Accuracy : 0.9736842105263158

C. Logistic regression classifier

Logistic regression classifier is very popular and widely used classification technique. This is simple, easy to implement, and provide good performance on a wide variety of problems. Logistic regression is a discriminative probabilistic classification model that operates over real-valued vector inputs. The dimensions of the input vectors being classified are called "features" and there is no restriction against them being correlated. Logistic regression is one of the best probabilistic classifiers, measured in both log loss and first-best classification accuracy across a number of tasks [15].

Logistic regression gives around 94 Logistic Regression model = Logistic Regression ()

model.fit(train$_X$, train$_y$)

prediction = model.predict(test$_X$)

print(^{0} Accuracy :^{0} , metrics.accuracy$_s$core(prediction, test$_y$))

Accuracy: 0.9473684210526315

D. Decision tree

Decision tree is one of the most widely used classifiers in statistics and machine learning. Decision tree is a hierarchical design that implements the divide-and-conquer approach. It is a nonparametric technique used for both classification and regression. It can be directly converted to a set of simple if-then rules. It’s straightforward representation makes the reader able to interpret the result and easy to understand. This section presents the basic features of the decision tree method for classification (Alpaydin, 2014) (Mitchell,1997) (Myles, Feudale, Liu, Woody, Brown, 2004).

Dicision tree gives around 92 DecisionTreeClassifier model = DecisionTreeClassifier() model.fit(train$_X$, train$_y$)

prediction = model.predict(test$_X$)

print(^{0} accuracy :^{0} , metrics.accuracy$_s$core(prediction, test$_y$))

accuracy : 0.9210526315789473

E. K-Nearest Neighbour (K-NN) Classifier

K-Nearest Neighbour (K-NN) algorithm is one of the supervised learning algorithms that have been used in many applications in the area of data mining, statistical pattern recognition and many others. It follows a method for classifying objects based on closest training examples in the feature space. An object is classified by a majority of its neighbours. K is always a positive integer. The neighbours are selected from a set of objects for which the correct classification is known [13]. K-NN works well even when there are some missing data. K-NN is good at specified which predictions have low confidence. It has some strong consistent results. As the amount of data approaches infinity, the algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data) [14].

K-NN classifier gives around 89KNeighborsClassifier model = KNeighborsClassifier ($n_n$eighbors = 3) model. f it(train$_X$, train$_y$)
prediction = model.predict(test$_X$)
print(^{0} Accuracy :^{0} , metrics.accuracy$_s$core(prediction, test$_y$))
Accuracy: 0.8947368421052632

5. CONCLUSION

The primary goal was to see performance analysis of Supervised Learning for Iris Flower species and present different technique for supervised learning methods like classification Regression technique with practically Implemented. This paper makes it a clear vision that every algorithm of supervised machine learning with practically implemented code which help us how to solve real life problem with best accuracy. The Selection of an algorithm should be made depending on the type of problem whose data available to you. The accuracy can be increased by using two or more algorithm together in suitable conditions.

REFRENCES

[1] Shashidhar T Halakatti, " Identification Of IRIS Flower Species Using Machine Learning " , IPASJ INTERNATIONAL JOURNAL OF COMPUTER SCIENCE(IIJCS), Volume 5, Issue 8, August 2017 , pp. 059-069 , ISSN 2321-5992.

[2] Diptam Dutta, Argha Roy, Kaustav Choudhury, “Training Aritificial Neural Network Using Particle Swarm Optimization Algorithm”, International Journal on Computer Science And Engineering (IJCSE). Volume 3, Issue 3, March 2013.

[3] poojitha V, Shilpi Jain, “A Collecation of IRIS Flower Using Neural Network CLusterimg tool in MATLAB”, Interna-tional Journal on Computer Science And Engineering(IJCSE).

[4] Vaishali Arya, R K Rathy, “An Efficient Neura-Fuzzy Approach For Classification of Dataset”, International Confer-ence on Reliability, Optimization and Information Technology, Feb 2014

[5] Mohamed, Amr. (2017). Comparative Study of Four Supervised Machine Learning Techniques for Classification.

[6] Badresiya, Ashok Saifee Vohra, Prof Jay Teraiya, Prof. (2014). International Journal of Advanced Networking Applica-tions (IJANA) Performance Analysis of Supervised Techniquesfor Review Spam Detection. 10.13140/2.1.1784.0962.

[7] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition. Morgan Kaufmann Publishers, San Francisco, USA. (ISBN-55860-901-6).

[8] R. Duda, P. Hart, and D. stork, Pattern Classification. 2nd Edition, Wiley Interscience, 2001.

[9] E. Frank, and I. Witten, Data Mining: Practical Ma-chine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[10] T. Joachims,Text Categorization with support Vector Machines: Learning with Many Relevant Features. In Pro-ceedings of ECML-98, 10th European Conference on Machine Learning, 1998.

[11] S. Dumais, D. Heckerman, J. Platt, and M. Sahami,” Inductive Learning Algorithms and Representations for Text Categorization,” In Proceedings of ACM-CIKM98, Pages 148-155, 1998.

[12] M. Haruno, and H. Taira, “Feature Selection in SVM Text Categorization,” In Proceedings of the 16th National Conference on Artifical Intelligence, Pages 480-486, 1999.

[13] M. Bramer, Principles of Data Mining, SpringerVer-lag, London, 2007. (ISBN 184628-765-0).

[14] O. David, and M. Francesco, "Research Challenge on Opinion Mining and Sentiment

Analysis," The CROSSROAD Roadmap on ICT for Governance and Policy Modeling, 2010.

[15] L. Liu, and Y. Wang, "a Method for Sorting Out the Spam from Chinese Product Reviews," In Proceeding of the Conference on Consumer Electronics, Communications and Networks.(CECNet), 2012

AUTHOR BIOGRAPHIES

ShivamVatshayanpursuing his B.tech degree in ComputerScience Engineering from Galgotias University, Greater Noida, Uttar pradesh, . His research interests include Machine Learn-ing, Artificial Intelligence, Image processing, Cloud computing, Nanotechnology, Big Data, Internet of things and Data science.