INTERNATIONAL JOURNAL OF
RECENT TRENDS IN MULTIDISCIPLINARY RESEARCH
https://www.doi.org/10.59256/ijrtmr.20240402015
March-April 2024, Vol 4(02), 82-87. ISSN No: 2583-0368

Comparative study on classification techniques through IRIS Data Analysis

Susmita Mondal[1], Aryapriya Roy[2], Sk Wasim Akram[3], Sk Md Zakir[4], Ankur Biswas[5], Kaustuv Bhattacharjee[6], Anirban Das[7]

[1,2,3,4,5,6,7] *Department Of Computer Application, University of Engineering & Management, Kolkata, West Bengal, India*.

OPEN ACCESS

Article Citation:
Susmita Mondal[1], Aryapriya Roy[2], Sk Wasim Akram[3], Sk Md Zakir[4], Ankur Biswas[5], Kaustuv Bhattacharjee[6], Anirban Das[7]" 'Comparative study on classification techniques through IRIS Data Analysis", International Journal of Recent Trends In Multidisciplinary Research, March-April 2024, Vol 4(02), 82-87.

Publishedby5th Dimension Research Publication

Abstract: Many classification techniques are implemented on different datasets in Machine Learning. This paper gives an idea on different classification techniques implementation and comparison with example an Iris Species Dataset. First dataset is preprocessed and categorized into two parts training set and test set then techniques like Decision Tree, Gaussian NB, Logistic Regression and Random Forest are used. Finally, accuracy of different techniques is compared.

Key words: Decision Tree, Gaussian NB, Logistic Regression, Random Forest

1. Introduction

In contemporary technologies, Machine Learning operates as a subset of Artificial Intelligence, endowing systems with the capability to autonomously learn and enhance performance based on experience without explicit programming. It primarily focuses on the creation of computer programs adept at accessing data and autonomously learning from it. Machine Learning is broadly categorized into Supervised Learning and Unsupervised Learning. In Supervised Learning, the machine utilizes pre-tagged data with correct answers, subsequently applying this knowledge to a new set of data. This category further divides problems into regression and classification, employing diverse algorithms such as Linear Regression, Logistic Regression, Decision Tree, Gaussian NB, and more. This paper explores the development of various machine learning models using an iris dataset, comprising measurements of irises previously identified by an expert botanist as Setosa, Versicolor, or Virginica. Through these measurements, different modeling algorithms predict the iris species, transforming the task into a three-class classification problem within the realm of supervised learning. Various tools and libraries, including Scikit, Numpy, Pandas, Matplotlib, Seaborn, and Jupyter Notebook with Python programming, are employed to achieve accurate predictions for iris species classification.

2. Background Study

Fisher's Iris dataset [1] is introduced by Ronald Fisher with multivariate characteristics in his 1936 paper. He developed a linear discriminant model with the mission to recognize the species from each other. Asmita et.al [2] implemented their method can automatically recognize the class of flowers with three approaches are segmentation, feature extraction and classification. Using Neural network, Logistic Regression, SVM and K-Nearest Neighbors. K.Thirunavukkarasu et.al [3] author discussed various methods and used different tools like Scikit and libraries like Numpy, Pandas etc. using all this tools they tested iris dataset flowers. There are growing interest in "ensemble learning" - methods that generate many classifiers and aggregate their

Results [4]. Random Forests algorithm is known as "the method representing the technical level of integrated learning". It is the representative algorithm based on Bagging developed on the basis of the decision tree [6]. Due to a large amount of computation in Neural Networks algorithm, regression tree and classification were brought up [7] in Random Forests, which reduce the processing time significantly by using binary method repeatedly.

3. Methodology

To deploy diverse machine learning models for determining the accuracy in identifying the iris species, we primarily employ four machine algorithms: Decision Tree, Gaussian Naive Bayes, Logistic Regression, and Random Forest classifier. However, all these four supervised learning algorithms are implemented using the scikit-learn toolkit based on Python.

Iris Versicolor

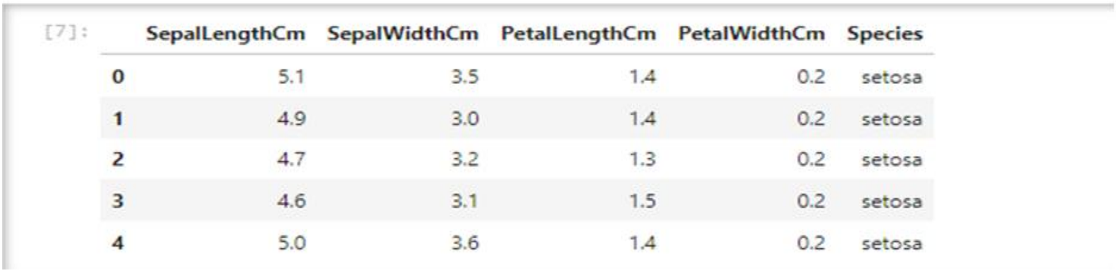
Iris Setosa

Iris Virginica

Loading Packages

At first, we import the packages required for analyzing the given dataset like pandas, NumPy, seaborn etc. Post that we also import several machine learning packages from scikit-learn module.

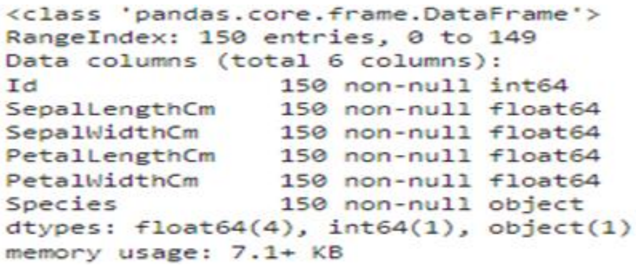
Reading data - R.A. Fisher's seminal 1936 work, "The Use of Multiple Measurements in Taxonomic Problems," introduced the Iris dataset, which is also accessible on the UCI Machine Learning Repository. Comprising three distinct iris species, each with 50 samples, the dataset provides various attributes for each flower. While one species exhibits linear separability from the other two, the remaining two species are not linearly separable from each other. The dataset columns encompass SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, and Species, offering comprehensive insights into the characteristics of these iris flowers.



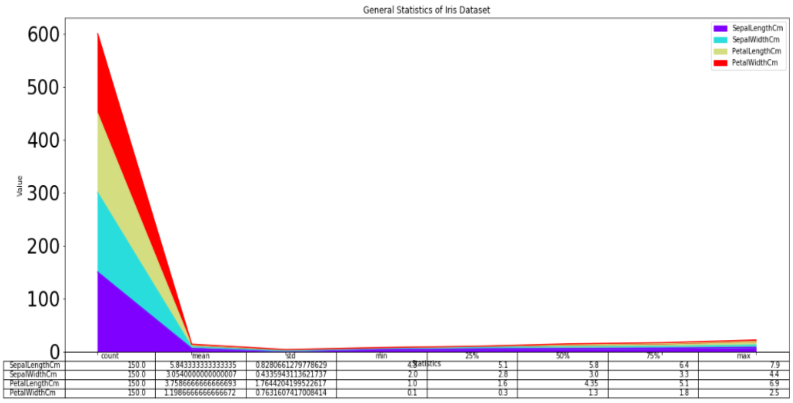
[7]:

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

Understanding the data-Let's first understand each column present in the given dataset. Here we get to know about the data types of each columns and if any missing value present. We also observe statistics about the numerical columns.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
Id                150 non-null int64
SepalLengthCm     150 non-null float64
SepalWidthCm      150 non-null float64
PetalLengthCm     150 non-null float64
PetalWidthCm      150 non-null float64
Species           150 non-null object
dtypes: float64(4), int64(1), object(1)
memory usage: 7.1+ KB
```

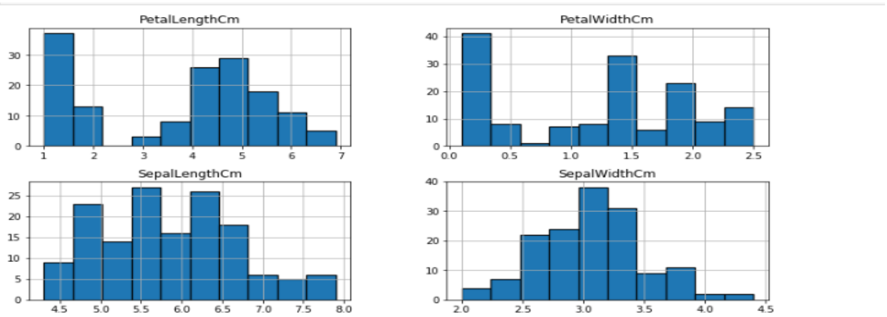


	count	mean	std	min	25%	50%	75%	max
SepalLengthCm	150.0	5.843333333333335	0.8280661279778629	4.3	5.1	5.8	6.4	7.9
SepalWidthCm	150.0	3.0540000000000007	0.4335943113621737	2.0	2.8	3.0	3.3	4.4
PetalLengthCm	150.0	3.7586666666666693	1.7644204199522617	1.0	1.6	4.35	5.1	6.9
PetalWidthCm	150.0	1.1986666666666672	0.7631607417008414	0.1	0.3	1.3	1.8	2.5

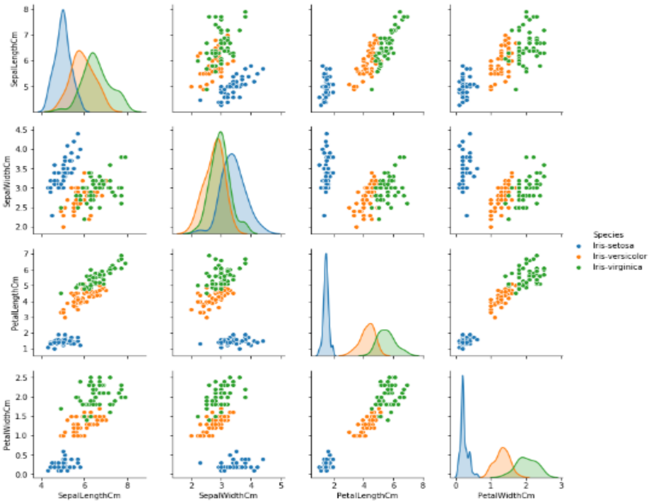
Exploratory Data Analysis

Now, let us visually analyze our dataset using some beautiful charts by matplotlib and seaborn module.

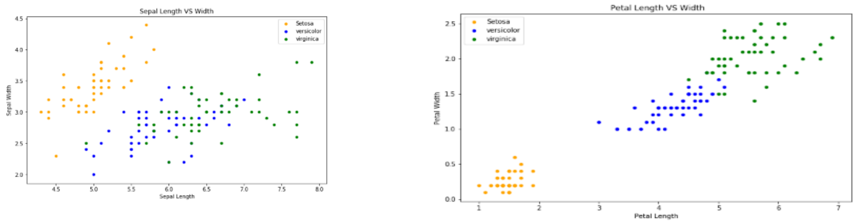
Histogram -



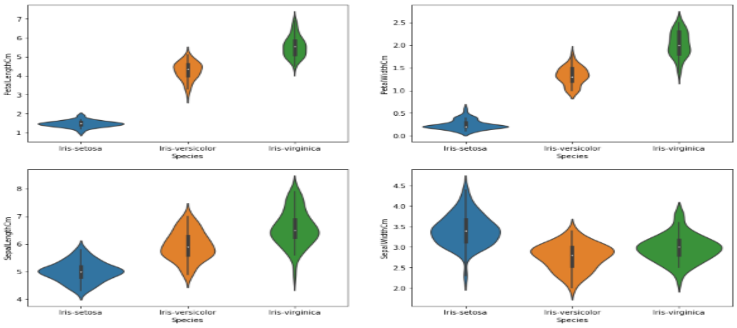
Pair Plot – Another useful seaborn plot is a hybrid plot called pair plot, which shows the bivariate relation between each pair of features. Let us see the same



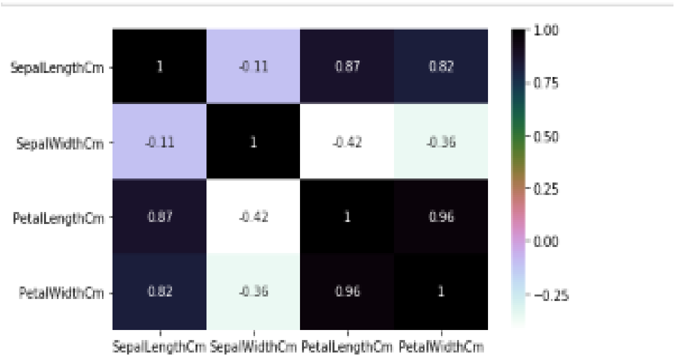
Scater Plot – At the outset, let us look at a simple scatter plot, to get a visual feel of the data. (We are going to view a host of them)



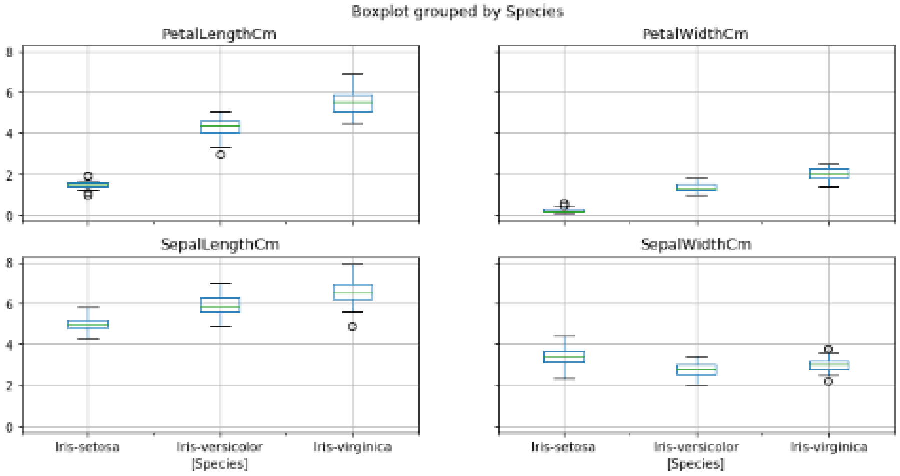
Violin Plot – violin plot is a statistical representation of numerical data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side



Heatmap – It is defined as a graphical representation of data using colors to visualize the value of the matrix. In this representation higher value or darker color means higher co-relation between the columns and negative values or brighter color means less co-relation.



Box Plot – Examine the box plot of the dataset, providing a visual depiction of the data distribution across the plane. A box plot, a graph reliant on percentiles, partitions the data into four quartiles, each representing 25%. Employed in statistical analysis, this approach facilitates the comprehension of diverse metrics such as mean, median, and deviation.

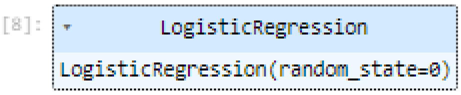


Implementing Machine Learning Algorithms - After gaining an understanding of the dataset, the next step involves training a model based on various algorithms. In this phase, we will apply some frequently used machine learning algorithms. Commencing with the training of our model using sample data, we will employ the 'train_test_split' function from an integrated library, which partitions our dataset into a 70:30 ratio.

Model Training
Utilizing commonly employed algorithms, we will train our model to evaluate the accuracy of each algorithm. The algorithms to be implemented for comparison include:
1] Logistic Regression
2] Decision Trees
3] Random Forest
4] Naive Bayes Classifier

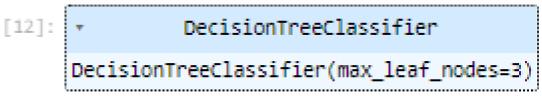
We will initiate the construction of our model and assess the accuracy of each algorithm to determine which one yields the most favorable results.

Logistic Regression Model Training



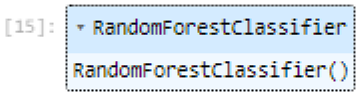
```
[8]: ▾ LogisticRegression
LogisticRegression(random_state=0)
```

Decision Tree Model Training



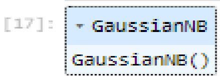
```
[12]: ▾ DecisionTreeClassifier
DecisionTreeClassifier(max_leaf_nodes=3)
```

Random Forest Model Training



```
[15]:  ▾ RandomForestClassifier
       RandomForestClassifier()
```

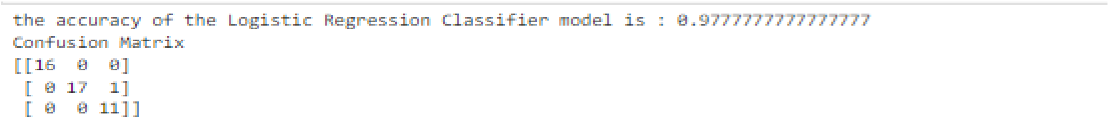
GaussianNB Model Training



```
[17]:  ▾ GaussianNB
       GaussianNB()
```

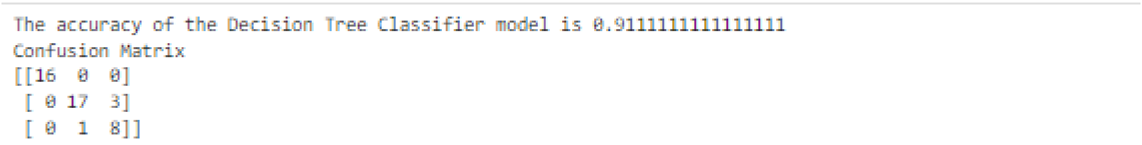
4. Results and Discussion

Logistic Regression Model Prediction



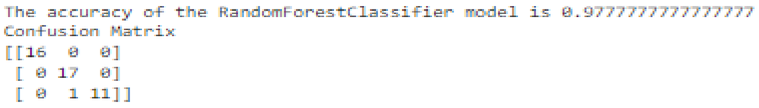
```
the accuracy of the Logistic Regression Classifier model is : 0.9777777777777777
Confusion Matrix
[[16  0  0]
 [ 0 17  1]
 [ 0  0 11]]
```

Decision Tree Model Prediction



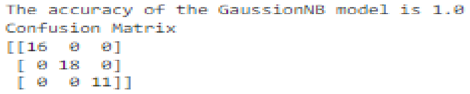
```
The accuracy of the Decision Tree Classifier model is 0.9111111111111111
Confusion Matrix
[[16  0  0]
 [ 0 17  3]
 [ 0  1  8]]
```

Random Forest Model Prediction



```
The accuracy of the RandomForestClassifier model is 0.9777777777777777
Confusion Matrix
[[16  0  0]
 [ 0 17  0]
 [ 0  1 11]]
```

GaussianNB Model Prediction



```
The accuracy of the GaussionNB model is 1.0
Confusion Matrix
[[16  0  0]
 [ 0 18  0]
 [ 0  0 11]]
```

Following the execution of classification, training, and testing on the iris species dataset using different models across various percentages, it becomes evident that the Gaussian Naive Bayes (NB) model consistently yields the most accurate results.

SNo	Model	Accuracy
1	Logistic Regression	0.97
2	Decision Tree	0.91
3	Random Forest	0.97
4	Gaussian NB	1

5. Conclusion

We have experimented with various machine learning models to forecast outcomes using the Iris Dataset, utilizing a 30% test dataset. The comprehensive test results presented in the table above indicate that the Gaussian Naive Bayes algorithm outperforms other algorithms in terms of accuracy. It is essential to recognize that the selection of the most suitable algorithm depends on several factors, including the data's nature, dataset size, and specific problem requirements. While simpler models such as Naive Bayes may excel in certain cases, more intricate models like decision trees, random forests, or support vector machines may be better suited for other scenarios.

References

1. *EN.WIKIPEDIA.ORG/WIKI/IRIS_FLOWER_DATA_SET*
2. *Asmita Shukla, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra, "Flower Classification using Supervised Learning,"Int. J. Eng. Res., vol. V9, no. 05, pp. 757–762, 2020.*

3. *K. Thirunavukkarasu, A. S. Singh, P. Rai, and S. Gupta, "Classification of IRIS dataset using classification based KNN Algorithm in supervised learning," 2018 4th Int. Conf. Comput. Commun. Autom. ICCCA 2018, pp. 1–4, 2018.*
4. *Roe, B. P., Yang, H. J., Zhu, J., Liu, Y., Stancu, I., McGregor, G.: Boosted decision trees as an alternative to artificial neural networks for particle identification. Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 543(2-3), 577-584 (2005)*
5. *Iverson, L. R., Prasad, A. M., Matthews, S. N., Peters, M.: Estimating potential habitat for 134 eastern us tree species under six climate scenarios. Forest Ecology and Management 254(3), 390-406 (2008)*
6. *Liaw, A., Wiener, M.: Classification and regression by randomForest. R news 2(3), 18-22 (2002)*
7. *Breiman, L.: Random forests. Machine learning 45(1), 5-32 (2001)*