IDENTIFICATION OF DIFFERENT SPECIES OF IRIS FLOWER USING MACHINE LEARNING ALGORITHMS

GUNJAN AHUJA[1], MUSKAN AGGARWAL[2], JASHN TYAGI[3], ONKAR MEHRA[4]

[1,2,3,4]B.Tech, Department of Computer Science, Delhi Technical Campus, Greater Noida, Uttar Pradesh, India
Professor -Ms. Nidhi, Associate Professor, Department of Computer Science, Delhi Technical Campus, GreaterNoida, Uttar Pradesh, India

---

Abstract- *The diversity of life on earth is incredibly rich.It is very challenging to pinpoint any species due to the fact that some flower species share the same shape, size, and colour on a physical level. Similar to this, there are three subspecies of the iris flower: Versicolor, Setosa and Virginica. The Iris dataset is what we are using. There are three classes with a total of 50 occurrences in the dataset for iris flowers. Machine learning is used to distinguish between Iris flower subclasses in the Iris dataset. The study concentrates on how Machine Learning algorithms can quickly and accurately identify the class of flower rather than relying just on approximations.*

Key Words: Iris dataset, Machine Learning Algorithms, Iris flower species, three classes, 50 occurrences

1.INTRODUCTION

Arthur Samuel claimed that machine learning is a branch of computer science. In 1959, he asserted that computers were capable of learning without explicit programming. Learning-based algorithms that can forecast based on data are studied and developed through machine learning. The study of computational learning and pattern detection theory in AI developed it. By creating a model from sample inputs, these algorithms avoid following entirely static programmer instructions and instead make data-based predictions or decisions.

When it is difficult or impossible to directly design and develop high-performance algorithms, machine learning is utilised in a range of computer processing related tasks. Examples include email filtering, network intruder detection, concept of rank learning and computer vision. The focus of machine learning is for computer programs to be made that can educate themselves to develop and adapt in response to fresh input. Computer science, artificial intelligence, and statistics are integrated by this branch of research. It is frequently called as statistic learning/predictive analytics.

This study's primary focus is the classification of IRIS flowers using machine learning and scikit technologies. The problem statement asks whether measurements of floral attributes can be used to identify IRIS flower species. IRIS flowers' measurable attributes such as sepal length, sepal width, petal length and petal width are used for pattern recognition. These attributes are used to recognize the pattern and identify the class of Iris flower. The predictive model makes a guess about the species based on what it has learned from the trained data whenever new data is discovered in this study. This strategy is carried out in three stages: segmentation, feature extraction, and classification using neural networks, logistic regression, k-Nearest Neighbors, decision trees, Linear Support Vector Classifiers (Linear SVC) random forest classifiers and Gaussian Naive Bayes.

2. OBJECTIVE

The study uses a dataset created in advance by qualified biologists to analyse the different flower kinds using data mining techniques and neural network classifiers in an effort to identify the type of iris blooms.

3. OVERVIEW

Classification is a supervised ML-based method that determines the group to which data instances belong. The introduction of neural networks has made the categorization problem easier to understand. This approach focuses on classifying iris flowers using neural networks. We will make use of the Scikit Learn Tool Kitto simplify classification.

This project mostly uses Scikit Learn to classify datasets. In order to distinguish between the three iris flower types-Setosa, Versicolor, and Virginica, the difficulty lies in measuring the length and width of the flower's sepal and petal. We can select the classification model with the highest accuracy to more accurately predict the species of iris flower after using the its dataset to train a variety of machine learning algorithms.

4. METHODOLOGY

Data sets- The dataset comprises of three classes of each species having 50 samples-versicolor, setosa, and virginica. This data is based on the well-known Fisher's model and has become an essential dataset for many classification applications in machine learning. The scikit-learn package includes this dataset. The rows are examples, while the

columns represent iris flower characteristics. The prediction model receives the data set. Each sample was measured using four characteristics: length and width of sepal and length and width of petal. These four measurements are incentimetres (cm).

Fig -1. Iris flower species

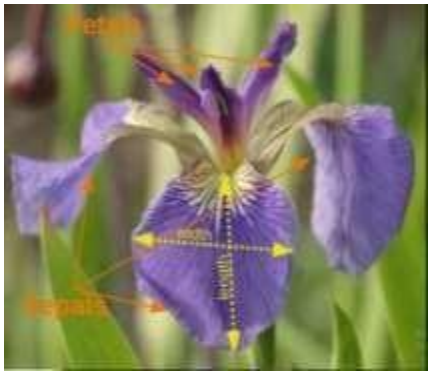

Fig2. Iris flower (petal length, petal width, sepal widthand sepal length)

4.1 Training- We utilise the dataset to train our model to predict output appropriately. We concentrate on categorising the iris flower class by extracting data from this dataset. The data supplied is processed in such a manner that each parameter is examined. Data preparation is essential in the machine learning process so that the data may be turned to a format that the computer can interpret. The algorithm can now readily comprehend the data characteristics. The aim is to categorise the flowers depending on their characteristics.

4.2 Testing- To categorise the testing data, we utilise the Random Forest Classifier in our code. We discovered that the K-means algorithm has a very high accuracy after utilising it. To determine the colour codes of the flowers, we employ the Random Forest Classifier.

5. PROPOSED WORK

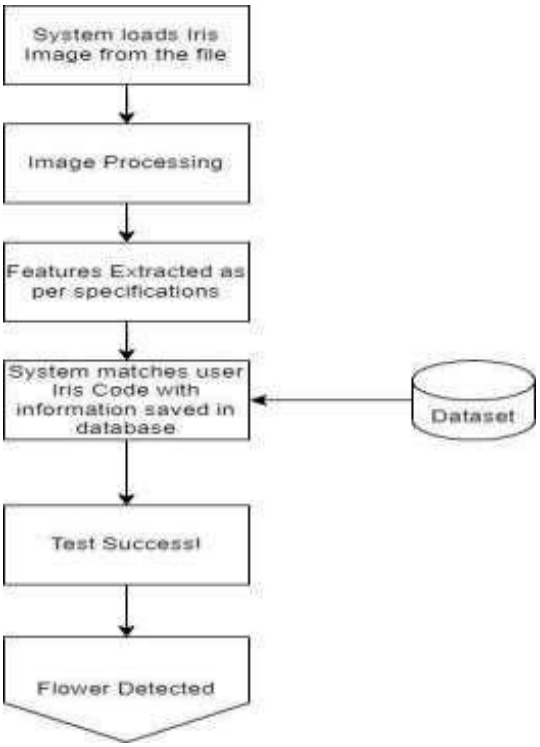

Fig.3 Work flow of the code

6. PROPOSED ALOGRITHMS

The proposed algorithms employ Neural Networks, segmentation, and feature extraction to identify the flower type as well as its measurements. Segmentation is utilised to eliminate the irrelevant backdrop and focus on the highlighted item, which is the flower. The major goal is to simplify the flower's depiction and deliver something more substantial and easy to examine. Feature Extraction extracts properties or information from flowers in the form of real values such as float, integer, or binary. The following methods were used: Logistic Regression, Decision Tree, k-Nearest Neighbor, Random Forest Classifier, Gaussian Nave Bayes, and Linear SVC.

6.1 Decision tree

The most effective and widely used tool for AI and ML- based predictions is the Decision Tree. A class label is held by each leaf node of a decision tree, an outflow is represented by each branch, and a test on an attribute is represented by each internal node. Decision trees are tree-based structures that resemble flow charts. Regression and classification techniques use it. It is a particular supervised learning strategy that does have parameters. A model that can forecast the value of the specified variable using simple decision rules is created using decision trees.

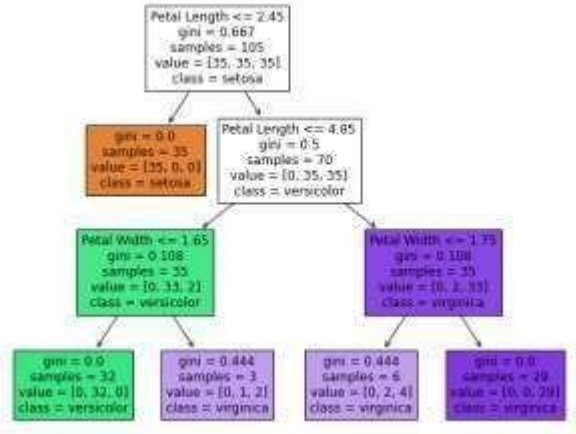

Fig 3. Decision tree

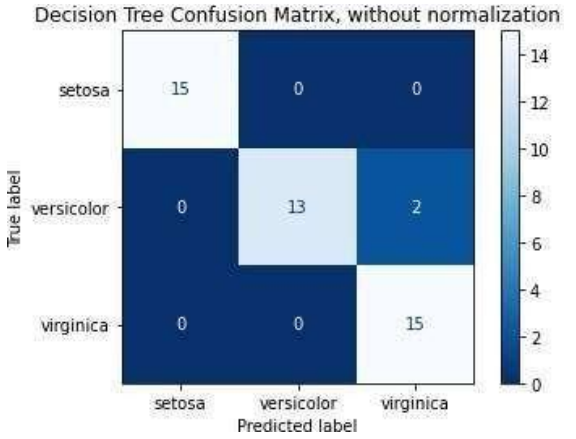

Fig 4. Decision tree confusion matrix

6.2 Random forest classifier

During the training phase, the ensemble learning technique known as random forest constructs decision tree models. Several models' outputs are combined to make the final decision by a group of models. Machine learning methods for classification include random forest. Random forest is basically an ensemble of trees. The random forest is constructed by only using a piece of the dataset and a restricted number of attributes, in comparison to decision trees, which use the complete dataset and take into account all characteristics.

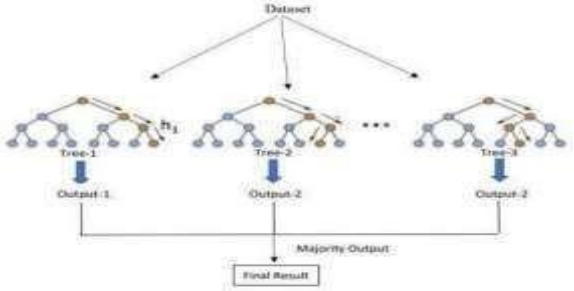

Fig 5. Random forest classifier

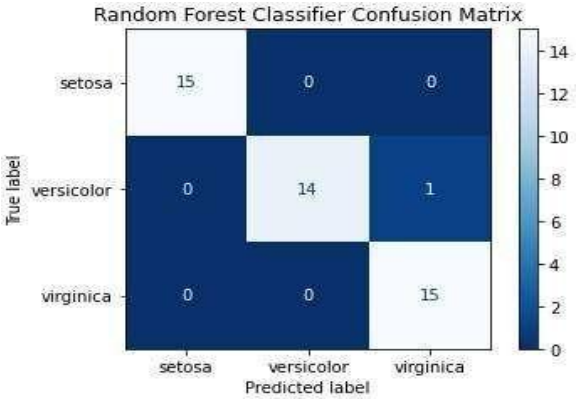

Fig 6. Random forest classifier confusion matrix

6.3 Gaussian Naïve Bayes

Gaussian Naive Bayes is a statistics-based classification approach that is based on Bayes Theorem. It uses stringent independence conditions. Here independence conditions in classification technique means that the change in one value does not impact another value. The performance of naive Bayes classifiers quickly deteriorate as the training set grows, but they are known to be highly expressive, scalable, and somewhat accurate in the domain of machine learning. The efficiency of naive Bayes classifiers is influenced by a number of factors.

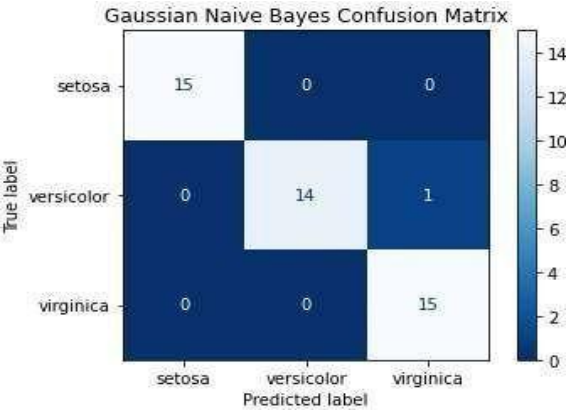

Fig7. Gaussion Naïve Bayes confusion matrix

6.4 Logistic Regression

A logistic function is used for the estimation of the odds of the happening of an event, with relevant or cleaned data, by a specific type of regression called logistic regression. It uses a lot of predictor variables, which may be numerical or categorical, just like other kinds of regression analysis. Binary data are used in logistic regression, where an event either occurs (true or 1) or does not occur (False or 0). As a result, it makes an effort to ascertain whether or not an event y takes place by offering some feature x. Consequently, y can either be 0 or 1. y is set to 1 if the event takes place. The value of y is set to 0 if the event does not take place.

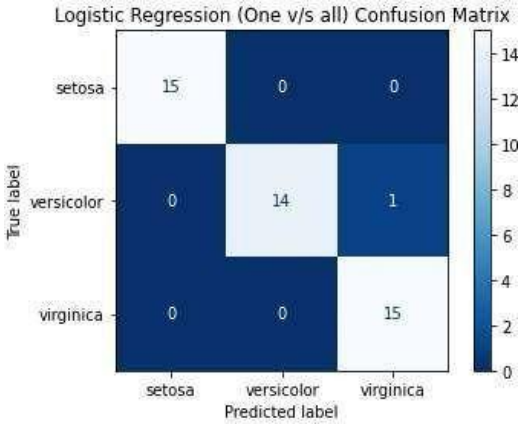

Fig 8. Logistic regression confusion matrix

6.5 K-Nearest Neighbour (KNN)

K-Nearest Neighbour is a supervised machine learning approach that is commonly used for classification process and employs a full dataset in extreme phase. When predicting unknown data, it explores the complete training dataset for k more similar cases and returns the data with the indistinguishable instance as the forecast. The K- Nearest Neighbour method maintains every available example and defines new cases based on the component similarity measure.

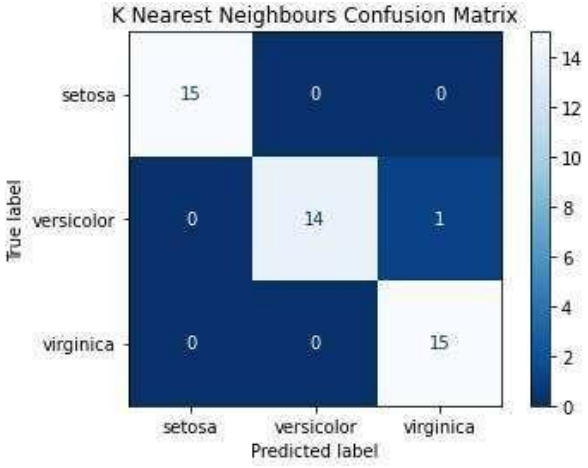

Fig 9. K-Nearest Neighbour confusion matrix

6.6 Linear SVC

Classification using a linear kernel function is conducted by the Linear Support Vector Classifier (SVC) approach and it works effectively with a large number of data. The Linear SVC contains extra parameters such as penalty normalisation ('L1' or 'L2') and loss function. A Linear SVC (Support Vector Classifier)'s goal is to adjust to the given data and provide a "best fit" hyperplane that categorises your data. The "predicted" class can be obtained by having some characteristics inputted to your classifier after the hyperplane is obtained.

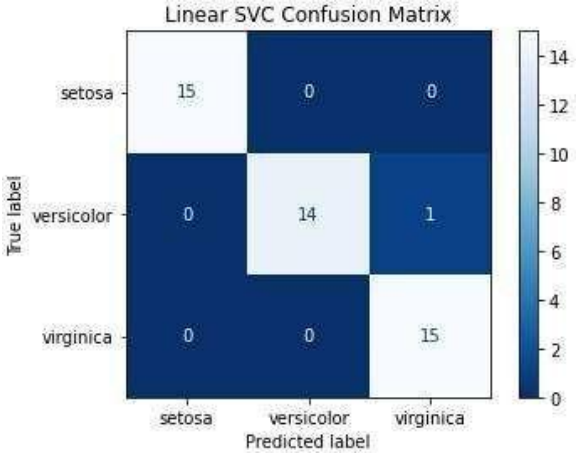

Fig10. Linear SVC confusion matrix

7 CONCLUSION

Technology is rapidly evolving. Artificial intelligence has been employed in a variety of sectors. Machine learningis the most fundamental approach for achieving artificial intelligence. Loading the iris dataset we learnt to generate our own supervised machine learning model utilising Iris Flower Classification. We learnt about data visualisation, machine learning, data analysis, model construction, and other topics as a result of this research. Our study discusses the various techniques used for data set analysis. To achieve high accuracy, k-Nearest Neighbors, Logistic Regression Decision Tree, Linear SVC methods, Gaussian Nave Bayes, and Random Forest classifier are utilised. We may infer that Linear SVC provides the highest level of accuracy.

Table -1: Accuracy and CVS of algorithms

S.no	Algorithms used	Accuracy %	CVS
1	Decision tree	95.556	94.67
2	Random forest classifier	97.778	95.33
3	Gaussian Naïve bayes	97.778	96.00
4	Logistic regression	97.778	96.00
5	K-nearest neighbor	97.778	97.33
6	Linear SVC	97.778	98.67

REFERENCES

[1] B. Keşkekçi, H. Bayrakçı and R. Arslan, "Classification of Iris Flower by Random Forest Algorithm", Advances in Artificial Intelligence Research, vol. 2, no. 1, pp. 7-14, Feb. 2022,

[2] Singh, R. Akash and G. R. V, "Flower Classifier Web App Using Ml & Flask WebFramework,"

2022 2nd International Conference on Advance Computing andInnovative Technologies in Engineering (ICACITE), 2022, pp. 974-977 doi: 10.1109/ICACITE53722.2022.9823577.

[3] Mishra P., Shukla A., Agarwal A. & Pant H. (2020). Flower classification using supervised learning. Vol. 9, 757-762.

[4] J.P., Pinto, Shetty, J. & Kelur, S., (2018). Iris Flower Species Identification Using Machine Learning Approach. 2018 4th International Conference for Convergence in Technology (I2CT).

[5] Shilpi Jain, Poojitha V, "A Collation of IRIS Flower Using Neural Network Clustering tool in MATLAB", International Journal on Computer Science and Engineering (IJCSE).

[6] Shashidar Halakatti, Shambulinga Halakatti, "Identification of Iris Flower Species Using Machine Learning", IIJCS Aug 2017.

[7] Anchal Garg, Shilpi Jain, Poojitha V and Madhulika Bhadauria, " A Collection of IRIS Flower Using Neural Network Clustering Tool In MATLAB", IEEE 2016.

[8] Asmita Shukla, Hemlata Pant, Ankita Agarwal & Priyanka Mishra (2020). "Flower Classification Using Supervised Learning", IJERT Vol-9 Issue-05