Identification of Iris Flower Varieties Through Machine Learning Techniques

[1] Sabiha Alam Chowdhury
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
sabiha.alam.chowdhury@g.bracu.ac.bd

[2] Fairuz Tassnim Prapty
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
fairuz.tassnim.prapty@g.bracu.ac.bd

[3] Ashakuzzaman Odree
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
ashakuzzaman.odree@g.bracu.ac.bd

[4] Syed Ashik Mahamud
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
syed.ashik.mahamud@g.bracu.ac.bd

[5] Annajiat Alim Rasel
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
annajiat@bracu.ac.bd

[6] Ehsanur Rahman Rhythm
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
ehsanur.rahman.rhythm@g.bracu.ac.bd

[7] Mehnaz Ara Fazal
Dept. of Computer Science and Engineering
BRAC UNIVERSITY
Dhaka, Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

***Abstract*—Classification stands as a crucial aspect of machine learning, whose core function revolves around analyzing data. A diverse range of algorithms are employed for classification, such a, multi-class classification, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), among others. In this research, we have elaborated on three specific methods. The implementation was carried out using the widely-known iris dataset, with the Scikit-learn toolkit facilitating this process. The focus of this paper is primarily on employing both classification and regression algorithms on the IRIS dataset. This involves the identification and examination of patterns based on the sizes of sepals and petals of the iris flower. Our findings indicate that the SVM classifier yields higher accuracy in comparison to the KNN and logistic regression models.**

I. Introduction

Machine learning, a subset of computer science, focuses on developing computer programs capable of self-improvement and adaptation when encountering new, previously unseen data. It's a research area intersecting predictive analytics and statistical analysis. Machine learning is broadly categorized into supervised and unsupervised learning. This paper emphasizes supervised learning, where a function is derived from labeled training data, comprising pairs of input and corresponding desired output values. Supervised learning is split into two types: classification and regression, with classification dealing with categorical outputs and regression handling continuous outputs.

This study presents techniques for identifying species of the flower's iris. This dataset is also regarded as Fisher's Iris dataset alternatively Anderson's Iris dataset, was introduced by Ronald Fisher and is a prominent multivariate dataset in the UCI Machine Learning Repository. The dataset, collected in part from the Gaspe Peninsula, includes three breeds of Iris that are setosa, vesicolor, and virginica, with 50 samples from each. It details four features: calyx length, calyx breath, bloom width, bloom width. However, main target for this whole procedure is achieving high precision in predicting unknown data. Moreover, classification involves training machine learning models with a training dataset and testing them. The Iris dataset's four measured features are used in training. Fisher's linear discriminant model was an early approach to distinguishing between species using these features. In our paper, we use the Scikit-learn toolkit to apply various machine-learning algorithms for Iris species classification, focusing particularly on SVM, Logistic Regression, and KNN.

II. Literature Review

In the field of machine learning and data analysis, the IRIS dataset has emerged as a key focus for numerous research studies, with diverse strategies being employed for classifying and identifying the different species of iris flowers. This comprehensive literature review highlights several notable methodologies in this area.

Deeptam Dutta's et. al.[1] team's trailblazing endeavor has showcased the immense potential of Artificial Neural Networks in accurately classifying iris flowers. Their groundbreaking approach, rooted in pattern recognition and neural network training, has ushered in a new era of predictive analysis and classification. As the world continues to unveil the mysteries of nature, innovative techniques like these will undoubtedly play a pivotal role in expanding our understanding and shaping the future of various scientific disciplines.

Pujitha A et. al.[2] and his team conducted a review that focused on collection and classification of iris flowers using neural networks. Recognizing that machine learning is a sub-discipline of computer science, they used an existing iris flower dataset in MATLAB to cluster into three distinct species.Their approach involves the use of k-means algorithms and neural network clustering tools within MATLAB. Neural network clustering tools are mainly used for supervised classification of large datasets.Their research results include successful clustering of the iris dataset into three species without supervision.

Research conducted by Vaishali Arya et. al.[3] and colleagues presents an efficient neural fuzzy approach for classification. In this study, the proposed technique is applied to iris data sets, classifying them into four categories.The neural fuzzy system they developed was able to select relevant features and generate a compact but adequate rule for the classification task, thereby increasing the efficiency of the classification process.

Sashidhar T et. al.[4] and his team proposed a unique detection method for iris flowers using classification techniques. They have made significant progress in predicting the outcome of unseen data that is not part of the model's training set.Their work involved training machine learning models with datasets to accurately predict iris species characteristics. Moreover, they also developed a predictive model that utilizes these species characteristics for effective forecasting.

Patrick S. et. al.[5] and his team dedicated their research to the statistical analysis of the IRIS flower dataset. Their study is divided into two distinct methodologies. The first method involves graphically representing the dataset to identify patterns in the classification of iris species. The second method entails the development of a Java application designed for extracting and analyzing statistical information from the IRIS dataset.

Each of these studies uniquely contributes to the body of knowledge surrounding iris species classification, using different techniques and methods from neural networks and fuzzy logic to statistical analysis and pattern recognition. These methods not only emphasize the versatility of machine learning applications but also the ongoing progress in botanical data analysis.

III. Methodology

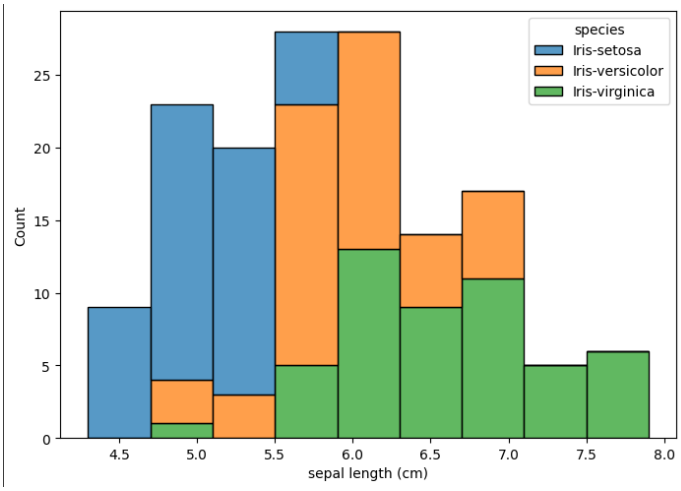
Our study aims to find the algorithms for classifying different species, within the Iris genus. To achieve this we used triads and developed a model. We employed three approaches here : support vector machine (SVM), Logistic regression ,K nearest neighbors (KNN) classification. These models were used in this paper. Evaluated based on four features of the Iris data. Additionally we utilized cross validation, a method that divides the dataset into subsets, for training and testing purposes. This helps improve the accuracy of our models and ensures the reliability of their outcomes.

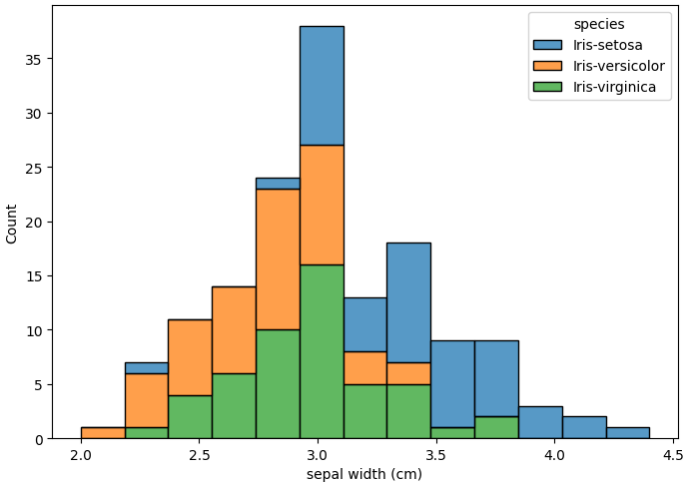
A. Dataset

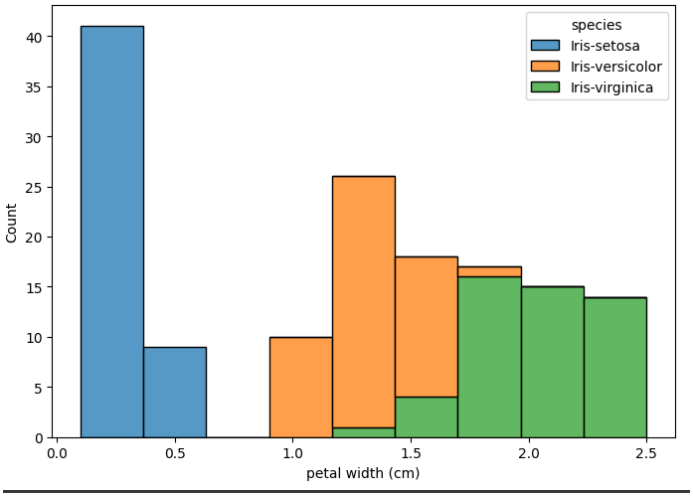
In our research we utilized the Iris dataset, which's well known for its documentation of the size and shape variations, among three distinct species of iris flowers. Our main objective was to create a classification model that could accurately classify these flowers into three types; Iris setósa, Iris versicólór and Iris virginicá. This dátáset consists of fifty samples, from each species resulting in a total of 150 samples.

Fig. 1. Iris Flower Species

Figure 2 illustrates how different characteristics of the iris are distributed among the three species. By analyzing histograms representing each feature we observed that 'petal width' serves as an distinguishing characteristic compared to sepal length, sepal width and petal length.







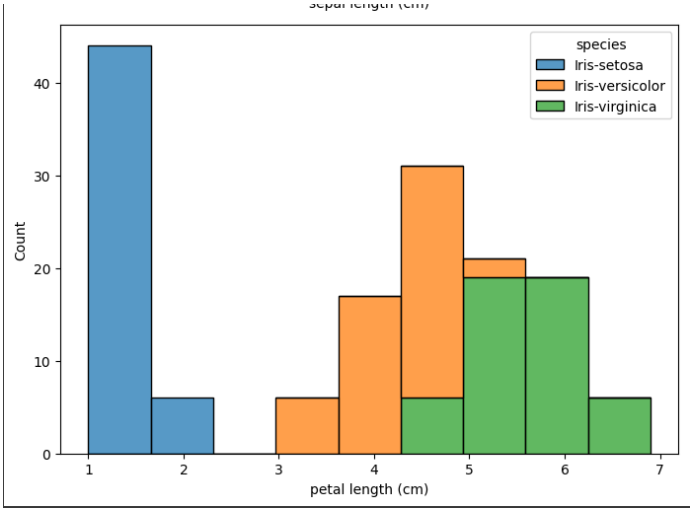


Fig. 2. Iris Feature Distribution

B. Loading

Initially, we sourced the Iris flower dataset from the UCI Machine Learning Repository. This dataset encompasses 150 samples distributed across three distinct Iris flower species: setosa, versicolor, and virginica. Following the collection of the dataset, it was imported into our machine learning model. For this import process, we utilized the scikit-learn toolkit, specifically employing the `load_iris()` function from scikit-learn's datasets module. This function was executed to load the dataset and store the returned value in a variable named "Iris".

Subsequently, we assigned various attributes within the Iris dataset. These attributes include the data itself, the names of the features, the target, and so on. The target names in the Iris dataset correspond to the classes, namely setosa, virginica, and versicolor. The feature names represent the dimensions of the flowers, such as sepal length, petal length, sepal width, and petal width.

Next, the dataset was partitioned into training and testing subsets. We allocated 40% of the dataset for testing, while the remaining 60% was reserved for training. Additionally, we set the `random_state` parameter to 0. It's important to fix the `random_state` to a specific number to ensure consistent results across multiple runs of the model. This consistency is crucial for accurately analyzing the model's performance. Without setting the `random_state`, each execution of the model might yield varying results, complicating the process of accuracy assessment.

C. Choosing the Model

In this phase, we selected the appropriate model to carry out species prediction. Utilizing the scikit-learn toolkit, we imported these 3 classifiers that are given below:

1) Support Vector Machine (SVM): SVM stands as an effective technique for classifying both linear and non-linear datasets. It relies on a non-linear transformation mapping for projecting the original training data into a multi-dimensional space.Here, SVM searches for a best-dividing subspace, which is constructed using a set of vectors. For the IRIS dataset, each data has been plotted in an n-dimensional space (here n is signifies feature number) as well as conducted classification.

For implementing the learning model, we used the SVM algorithm. The SVC() function from sklearn's module was employed to extract the needed model. Moreover, the accuracy of the SVM model was determined to be 98%.

2) Logistic Regression: It is an arithmetical process with statistical analysis used for examining datasets with one or more independent variables that influence the outcome. Its primary purpose is to categorize data accurately based on current information. In the case of the Iris flower dataset, logistic regression is utilized to segment the data according to the length and width of the Iris flowers. It is a renowned algorithm for binary and categorical data analysis, using a sigmoid function as its hypothesis, denoted as $p=1/(1+e^{-y})$. It functions effectively with larger datasets.

The Logistic_Regression() action was utilized for importing the exact necessary dummy. Upon evaluating the dataset using

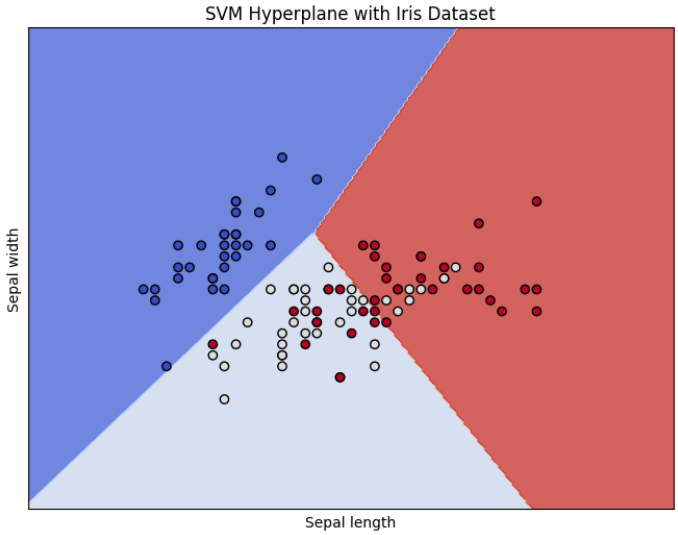


Fig. 3. SVM Hyperplane

logistic regression, the model's accuracy was found to be 97.33%.

3) K-Nearest Neighbor Classifier(KNN): The KNN algorithm is uncomplicated and depends solely on the data it was trained on. It identifies the k closest examples in the training dataset to predict new data points. KNN assigns a class to the new data by measuring how closely it resembles the existing data points. In our implementation of the KNN classifier, we employed the KneighborsClassifier from sklearn.neighbors with n_neighbors set to 3. The model achieved a prediction accuracy rate of 96.67%. Every classifier was chosen and assessed for its precision in distinguishing species of iris, taking into account the characteristics of the flowers, the dimensions of the features, and the methods used for measuring similarity.

D. Implementation of Cross-Validation

Cross validation is a technique used to reserve a portion of a dataset, which is not utilized during the model training process. Next this reserved portion of test data is used to ássess the performance of the módel. If the outcómes, with this validation data are satisfactory it indicates that the model is effective. The implementation of validation enhances both efficiency and effectiveness in improving the performance of the model. In Figure 4 we can observe how accurately SVM, KNN and Logistic Regression classification methods perform when applied to the iris dataset. By comparing accuracy results with and, without using cross validation our analysis revealed that incorporating cross validation significantly improves model accuracy compared to scenarios where it is not employed.

Steps in Cross-Validation:

(I) Initially, set aside a portion of the dataset as sample data.

(II) Use the remaining part of the dataset to train the model.

TABLE I
ILLUSTRATION OF SAMPLE IRIS DATASET

sepal_length	sepal_width	petal_length	petal_width	species
5.6	2.8	4.9	2	Iris-virginica
6.4	3.1	4.5	1.8	Iris-virginica
6.8	2.8	4.8	1.4	Iris-versicolor
7.1	3	5.9	2.1	Iris-Virginica
5.7	2.6	3.5	1	Iris-Virgicolor

(III) Employ the reserved sample from the test set to assess the model's performance. A model that yields favorable results with this validation data is considered effective.

One of the primary benefits of cross-validation is obtaining a more precise result estimate from the sample accuracy. It enhances the efficiency and effectiveness of the model's performance. Figure 4 illustrates the accuracy levels of SVM, KNN, and Logistic Regression classification methods when applied to the iris dataset. We compared the accuracy results with and without the cross-validation process. Our analysis revealed that the accuracy of the models improves when cross-validation is implemented, as opposed to scenarios where it is not used.

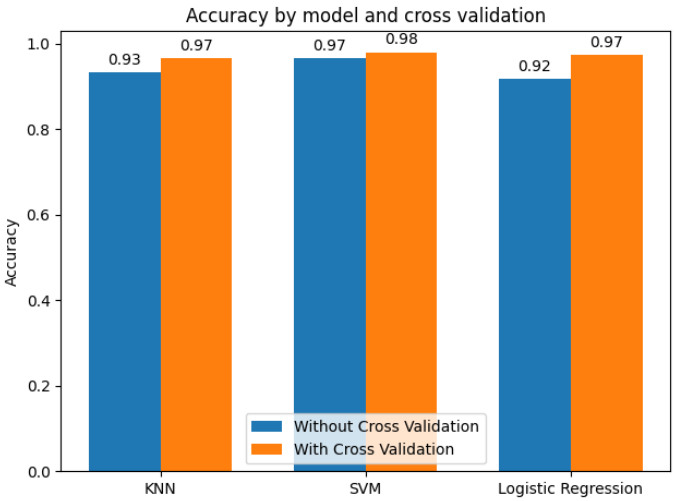


Fig. 4. Model Accuracy Score

E. Utilizing the Confusion Matrix

The confusion matrix serves as a tabular representation to evaluate the efficacy of a classification model. This matrix utilizes test data where the expected output labels are already established. It provides a straightforward indication of whether the model's predictions are accurate or not, and aids in pinpointing any errors made by the model. The model's accuracy score was determined using the confusion matrix.

In the context of the Iris dataset, the first four columns represent the dataset's attributes, while the fifth column denotes the target, indicating the classification label for the given sample data.

IV. Conclusion

The study presents a variety of techniques and algorithms for analyzing the iris dataset, focusing specifically on the effectiveness of Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression in achieving precise classification results. It also mentions cross-validation as a pivotal enhancement technique for model accuracy. Comparative analysis reveals that SVM is the most efficient among the examined techniques.

References

[1] Dutta, D., Roy, A., Choudhury, K., "Training Artificial Neural Network Using Particle Swarm Optimization Algorithm", International Journal on Computer Science and Engineering (IJCSE), Volume 3, Issue 3, March 2013.

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

[3] Arya, V., Rathy, R. K., "An Efficient Neural-Fuzzy Approach For Classification of Dataset", International Conference on Reliability, Optimization and Information Technology Feb 2014.

[4] Cho, S., Dehuri, S., "A comprehensive survey on functional link neural network and an adaptive PSOBP learning for CFLNN", Neural Comput & Application DOI 10.1007/s00521-009-02885

[5] Hoye, P., "Statistical analysis of iris flower dataset", University of Massachusetts at Lowell.

[6] Fisher, R. A., "UCI Machine Learning Repository: Iris Data Set", Available at: http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013.