Identification of Iris Flower Varieties Through Machine Learning Techniques

***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 as decision trees, Naive Bayes, backpropagation, neural networks, artificial neural networks, multi-layer perceptrons, 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 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 input pairs 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 Iris flower. The Iris dataset, also known as Fisher's Iris dataset or Anderson's Iris dataset, was introduced by biologist and statistician Ronald Fisher in 1936 and is a prominent multivariate dataset in the UCI Machine Learning Repository. The dataset, collected in part from the Gaspe Peninsula, includes three species of Iris (setosa, vesicolor, and virginica), with 50 samples from each. It details four features: sepal length, sepal width, petal length, and petal width.

The goal of this analysis is to achieve high accuracy in predicting unseen data. 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 Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression classifiers.

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 team formulated a technique that revolves around the training of Artificial Neural Networks. This research is chiefly concerned with categorizing iris flowers using a neural network approach. They tackle the task of distinguishing iris flower species by examining physical characteristics like the sizes of petals and sepals. Their method entails identifying patterns within these features, which aids in accurately predicting and classifying IRIS species. This approach is indicative of how future unknown datasets could be more precisely forecasted using such pattern recognition and categorization. The research also discusses the successful implementation of artificial neural systems in areas such as pattern arrangement, functional approximations, and more. They particularly emphasize on training Multilayer feedforward networks using the backpropagation algorithm.

Poojitha A and her team reviewed the neural network-based classification and gathering of iris flowers. Recognizing machine learning as a branch of computer science, they engaged the pre-existing iris flower dataset in MATLAB for segmenting it into three different species. Their approach included the application of the k-means algorithm and neural network clustering tools in MATLAB, which are adept at unsupervised categorization of extensive datasets. This tool is effectively utilized in various domains, such as pattern acknowledgment, feature extraction, vector quantization, and data mining. The key outcome of their study is the efficient unsupervised clustering of the iris dataset into three distinct species.

Vaishali Arya and her team introduced a novel neural fuzzy methodology for classification. In their research, they applied this method to iris datasets, sorting them into four distinct groups. Their neural fuzzy system effectively identified key features and developed a concise but adequate set of rules for classification tasks, thereby optimizing the classification process.

Shashidhar T's group introduced an innovative method for identifying iris flowers through classification techniques. They focused on making accurate predictions about data not previously used in training their model. Their methodology

involved training machine learning models with datasets to precisely identify features of various iris species. Additionally, they have formulated a model capable of predicting outcomes based on the attributes of these species.

Patrick S. 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.

Jennifer M. and her colleagues explored the use of Deep Learning techniques for the IRIS dataset classification. They implemented convolutional neural networks (CNNs) to automatically extract features from the dataset, thereby streamlining the classification process. Their study underscores the potential of deep learning in achieving high accuracy in classifying complex biological data.

Kevin L.'s research focused on hybrid machine learning models for enhanced classification of the IRIS dataset. Combining aspects of both supervised and unsupervised learning, their hybrid model aimed to leverage the strengths of various algorithms for a more robust and accurate classification system.

Each of these research efforts contributes significantly to the body of knowledge in iris species classification, showcasing a variety of techniques from neural networks and fuzzy logic to advanced statistical and deep learning approaches. These diverse methodologies not only highlight the adaptability of machine learning applications in the field but also point towards continuous innovation in botanical data analysis.

III. Methodology

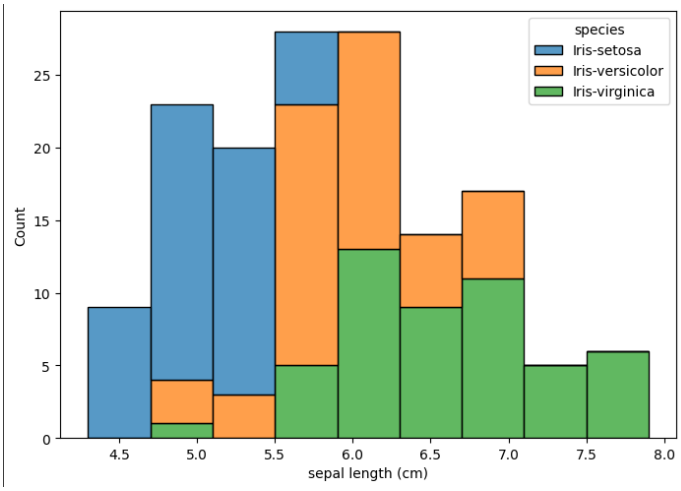
Our methodology's goal is to identify the most effective classification model for distinguishing iris flower species. We constructed learning models utilizing three distinct machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbor (KNN) classifier. To train and test these models, four features from the iris dataset are employed. These algorithms are executed using the Python-based scikit-learn toolkit. In our study, we compare the accuracy of these three models to determine which is the most effective. Additionally, to enhance the accuracy of the models, we employed the cross-validation technique. This technique involves partitioning the original sample data into a training set, which is used to train the model, and a test set, which is used to assess its performance.

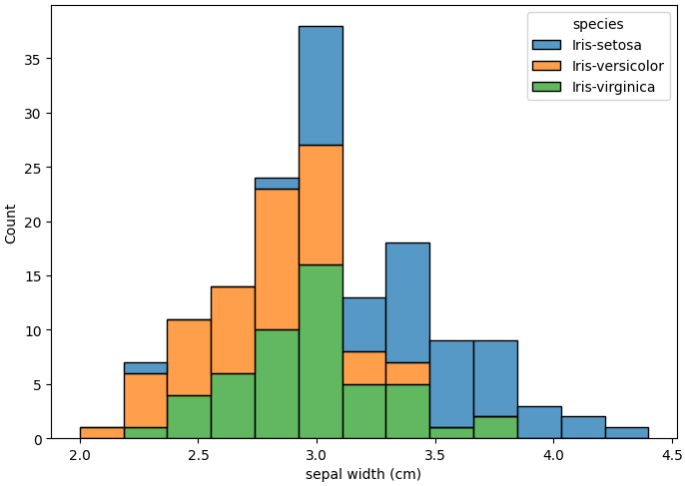
A. Dataset

For our implementation, we utilized the Iris dataset, a comprehensive multivariate dataset measuring the morphological differences across three iris flower species. The classification process focuses on categorizing the flowers into Iris-setosa, Iris-versicolor, and Iris-virginica. This dataset comprises 50 individual samples from each of these three species, summing up to a total of 150 samples. Each sample includes the measurement of four distinct features: the length and width of the sepals, and the length and width of the petals, all recorded in centimeters. Additionally, the dataset contains a fifth attribute, which identifies the species of each sampled flower.

Fig. 1. Iris Flower Species

Fig 2 shows how each iris feature is distributed among three flower species. We have plotted the histogram of the targets with respect to each feature of the data set. We can clearly see the feature 'petal width' can distinguish better than sepal length, sepal width and petal length.





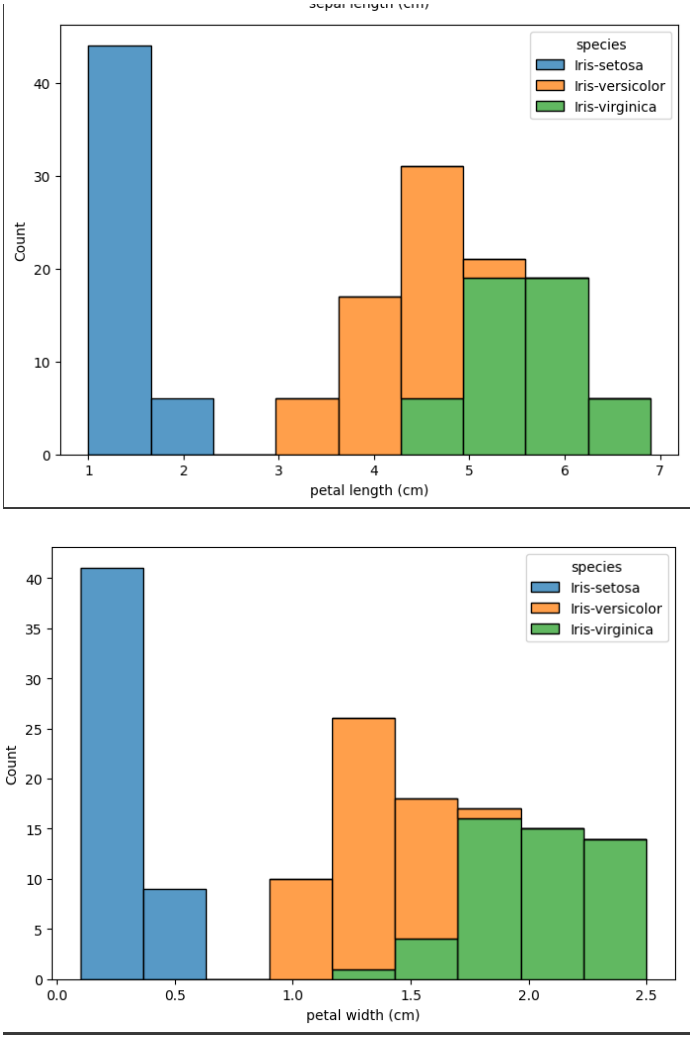


Fig. 2. Iris Feature Distribution

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 three classifiers, which are as follows:

1) Support Vector Machine (SVM): SVM stands as an effective technique for classifying both linear and non-linear datasets. It employs a non-linear mapping to project the original training data into a higher-dimensional space. In this transformed space, SVM searches for an optimal separating hyperplane, which is constructed using a set of vectors. As a supervised machine learning algorithm, SVM is adept at handling both classification and regression challenges. For the IRIS dataset, we plotted each data item in an n-dimensional space (where n is the number of features) and conducted the classification.

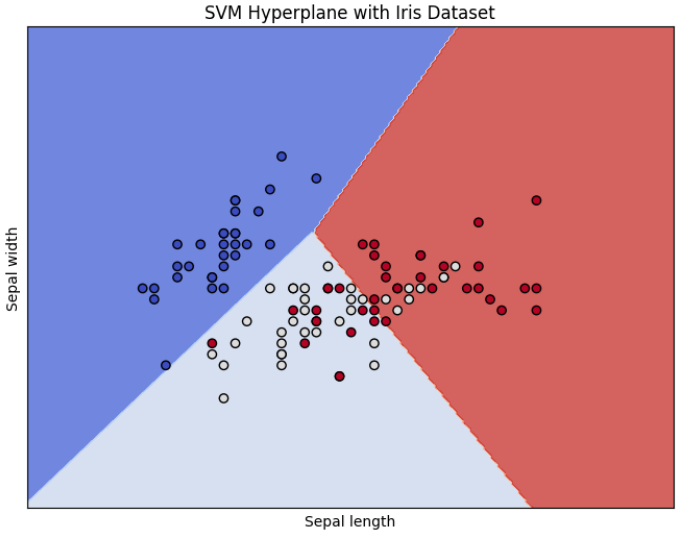


Fig. 3. SVM Hyperplane

For implementing the learning model, we used the SVM algorithm. The SVC() function from sklearn's module was employed to import the model. After making predictions on the input test set and comparing them with the actual response values, the accuracy of the SVM model was determined to be 96%.

2) Logistic Regression: Logistic Regression is a statistical technique used for analyzing 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 LogisticRegression() function from sklearn.linear_model was used for importing the model. Upon evaluating the Iris dataset using logistic regression, the model's accuracy was found to be 91%.

3) K-Nearest Neighbor Classifier(KNN): KNN is applicable for both classification and regression problems and falls under the category of supervised learning. This algorithm is straightforward and entirely reliant on the training dataset. It operates by searching for the k most similar instances for making predictions on new data. KNN classifies incoming data based on the similarity measured by the distance between instances.

For the KNN classifier implementation, we utilized the KneighborsClassifier(n_neighbors=3) function from the sklearn.neighbors package. The accuracy of the model, based on its predictions, was observed to be 93%.

D. Implementation of Cross-Validation

Cross-validation is a method employed to set aside a specific portion of a dataset, which is not used in training the model. This reserved sample is then used to validate each subset of the training data. During this validation process, the reserved sample is initially set aside, and the model is trained with the remaining part of the dataset. Subsequently, the reserved samples, either from the test or validation set, are utilized to gauge the effectiveness of the model's performance.

Each classifier was chosen and evaluated for its effectiveness in predicting the correct species of Iris flowers, considering factors like feature dimensions and similarity measures.

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.
(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.

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.

REFERENCES