IRIS FLOWER PREDICTION

A Project Report in partial fulfillment of the degree

## Bachelor of Technology

in

## Computer Science &Engineering

#### By

**2103A51280** K CHAITANYA JYOTHI

**2103A51412** N CHANDANA AZITH

**Under the Guidance of**

## S. THARUN REDDY

### Assistant Professor, Department of CSE.

**Submitted to**



## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING S R UNIVERSITY, ANANTHASAGAR, WARANGAL





**DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING**

**CERTIFICATE**

This is to certify that the Project Report entitled “Iris Flower Prediction” is a record of Bonafide work carried out by K.CHAITANYA JYOTHI, N.CHANDANA AZITH bearing Roll No(s) **2103A51280, 2103A51481** during the academic year 2022-2023 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science Engineering** by the SR UNIVERSITY, WARANGAL.

## Supervisor Head of the Department

Mr. S. Tharun Reddy Dr. M. Sheshikala

Asst. Professor, Assoc. Prof .& HOD (CSE)

SR University SR University

## ACKNOWLEDGEMENT

We express our thanks to course coordinator Mr. **S. Tharun Reddy**, Asst. prof. for guiding us from the beginning through the end of the course project. We express our gratitude to head of the departmentCS&AI, **Dr. M. Sheshikala,** Associate Professor for encouragement, support and insightful suggestions.We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

By,

**N. Chandana Azith (2103A51481)**

**K. Chaitanya Jyothi (2103A51280)**

## ABSTRACT

Iris flower classification is one the major problem while classifying the iris plant . This project can be used to classify the iris flower into one of the three classes Iris Setosa, iris Virginica, iris versicolour on the basis of sepal length sepal width, petal length and petal width, etc. Machine learning is used to solve this problem because it provides us with promising results. There are several algorithms that can be used to predict the iris plant using its class. The machine learning model used in the project is logistic regression. We have accessed the precision, accuracy and recall after predicting on test data.

**Keywords** - machine learning, iris flower, iris classes, logistic regression, accuracy.

## Table of Contents

#### Chapter No. Title Page No.

1. Introduction
   1. [Overview 1](#_TOC_250015)
   2. [Problem Statement 1](#_TOC_250014)
   3. Existing system 1
   4. Proposed system 1
   5. [Objectives 2](#_TOC_250013)
   6. Architecture 2
2. [Literature survey 3](#_TOC_250012)
3. [Data pre-processing](#_TOC_250011)
   1. Dataset description 5
   2. [Data set 6](#_TOC_250010)
   3. [Data augmentation 6](#_TOC_250009)
   4. [Data Visualization 7](#_TOC_250008)
4. [Methodology](#_TOC_250007)
   1. [Logistic regression algorithm 10](#_TOC_250006)
   2. [K-Nearest Neighbour algorithm 10](#_TOC_250005)
   3. Decision Tree 10
   4. Support Vector System 11
   5. [Software description 11](#_TOC_250004)
5. [Results 12](#_TOC_250003)
6. Conclusion 17
7. Future Scope 18
8. References 19

## INTRODUCTION

### OVERVIEW

Flower classification is a very popular machine learning project. The iris dataset contains three classes of flowers, Versicolor, Setosa, Virginica, and each class contains 4 features, ‘Sepal length’, ‘Sepal width’, ‘Petal length’, ‘Petal width’. The aim of the iris flower classification is to predict flowers based on their specific features. There are various applications where iris class predicition can be useful.

Some of these include:

* + - This data set is often used in data mining classification and clustering.

### PROBLEM STATEMENT

But the problem was to find a suitable classification algorithm that could do the job with maximum accuracy for our data set. Different algorithms have different accuracy depending on the type of problem it has to solve and the data set it has to work with. So, we decided to select four algorithms, namely KNN, SVM and Logistic Regression and to compare the accuracy levels of each of these algorithms, with respect to our problem and data set. The result of this test would help us in determining which algorithm to use while implementing our iris flower prediction.

### 1.2EXISTING SYSTEM

In 2009, V. Borovinskiy connected three diverse neural systems procedures (Multilayer neural system (MLP), Probabilistic neural system and Radial base model is 98%. V. Kumar and N. Rathee introduced a coordinated bunching and characterization show (J48 with Kimplies) to acknowledgment of Iris dataset which gives the 98.66 % testing precision. In 2013, D. Dutta et al. Proposed the adjustment of system loads utilizing PSO proposed as a component to propel the execution of ANN in characterization of IRIS dataset which produces 97.3 % approval exactness[1]. In 2017, K.H. Wandra and L.P. Gagnani utilized the WEKA datamining instruments with various AI calculations (Multi Layer Perceptron, RBF, Naïve Bayes, J48) on IRIS dataset. MLP gives the better precision results 97.33 %[1]. In 2018, Mohan P. M. et al. proposed bolster vector machine methods with various variety of SVM on Iris dataset which given the 96.7 % most astounding precision for Q- SVM.

### PROPOSED SOLUTION

Different datasets of Iris Flower are assembled. There are absolutely 150 datasets having a place with three unique types of Iris Flower that is Setosa, Versicolor and Virginica. The gathered Iris Datasets are stacked into the Machine Learning Model. Scikit-learn accompanies a couple of standard datasets, for example the Iris dataset for order. The load\_irirs work is imported from Scikit- learn. The load\_iris work is run and

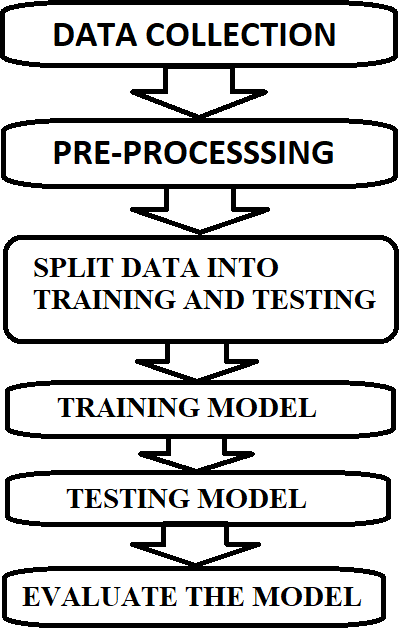
spare the arrival esteem in an item called "Iris". With the assist of dataset obtained we created 5 different machine learning algorithms and they are KNN,SVM,LOGISTIC REGRESSION and DECISION TREE.

### OBJECTIVES

The objectives of our project are as follows:

* + - We are going to compare each and every accuracy.
    - We are going establish the machine learning algorithm which gives the best accuracy.
    - Present different type of results using different models.
    - Select an accurate model and use it for the prediction.
    - Deliver a well-presented outcome from the ML model.

## OVERALL ARCHITECTURE



***Fig: 1.6.1 Diagram to represent architecture of Iris flower prediction***

## LITERATURE SURVEY

The Iris flower dataset is a widely used benchmark dataset in the field of machine learning and pattern recognition. It consists of 150 samples of iris flowers, each with four features (sepal length,sepal width, petal length, and petal width) and a corresponding class label indicating the species of the iris flower (setosa, versicolor, or virginica). Predicting the species of an iris flower based onits features is a classic problem in machine learning and has been extensively studied in the literature. Here is a literature survey of some key works related to iris flower prediction.

* Fisher, R.A. (1936). "The use of multiple measurements in taxonomic problems." - Ronald A. Fisher's original paper on the Iris dataset, published in 1936, is one of the earliest works on this topic. Fisher proposed the use of linear discriminant analysis (LDA) to classify the iris flowers based on their features, which laid the foundation for many subsequent studies.
* Duda, R.O., & Hart, P.E. (1973). "Pattern Classification and Scene Analysis." - This classic textbook introduced the concepts of linear discriminant analysis (LDA) and the k-nearest neighbors (KNN) algorithm, which are commonly used for iris flower prediction. It provides a comprehensive overview of pattern recognition techniques and their application to various classification problems, including the iris flower prediction problem.
* Cover, T.M., & Hart, P.E. (1967). "Nearest neighbor pattern classification." - This seminal paper introduced the k-nearest neighbors (KNN) algorithm, which is a popular non-parametric method for classification tasks. KNN has been widely used for iris flower prediction due to its simplicity and effectiveness in handling multi-class classification problems.
* Cortes, C., & Vapnik, V. (1995). "Support-vector networks." - This influential paper introduced the support vector machine (SVM), a powerful and versatile machine learning algorithm that has been applied to various classification problems, including iris flower prediction. SVM has been shown to achieve high accuracy in iris flower classification and has been widely used in the literature.

3

* + Breiman, L. (2001). "Random forests." - This seminal paper introduced the random forest algorithm, which is an ensemble method that combines multiple decision trees for improved classification performance. Random forests have been shown to be highly effective for iris flower prediction, as they can handle high-dimensional feature spaces and are resistant to overfitting.
  + Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." - Scikit-learn is a popular open-source machine learning library in Python that provides a wide range of tools for classification, including many algorithms for iris flower prediction. This paper describes the scikit-learn library and its implementation of various classification algorithms, including LDA, KNN, SVM, and random forests.
  + Palafox, L., et al. (2018). "A comparative study of machine learning algorithms for iris classification." - This study compared the performance of various machine learning algorithms, including LDA, KNN, SVM, and random forests, for iris flower prediction. It evaluated their accuracy, precision, recall, and F1 score, and provided insights into the strengths and weaknessesof different algorithms for this problem.
  + Ghasemi, E., & Najafi, A. (2019). "Comparative study of machine learning algorithms on iris dataset using R." - This study compared the performance of different machine learning algorithms, including decision tree, naive Bayes, SVM, and random forests, for iris flower prediction using the R programming language. It evaluated their accuracy, sensitivity
  1. **DATA DESCRIPTION**

## DATA PREPROCESSING

## 3.1 DATA DESCRIPTION

The Iris flower dataset is a popular multi-class classification problem that involves predicting the class of an iris plant based on the measurements of its petals and sepals. It is a small dataset that consists of 150 observations, where each observation corresponds to a different iris plant.

**The dataset has four features or attributes**: Sepal length in centimeters

Sepal width in centimeters Petal length in centimeters Petal width in centimeters

#### These features are used to predict the species of the iris plant, which are the target classes. There are three species in the dataset:

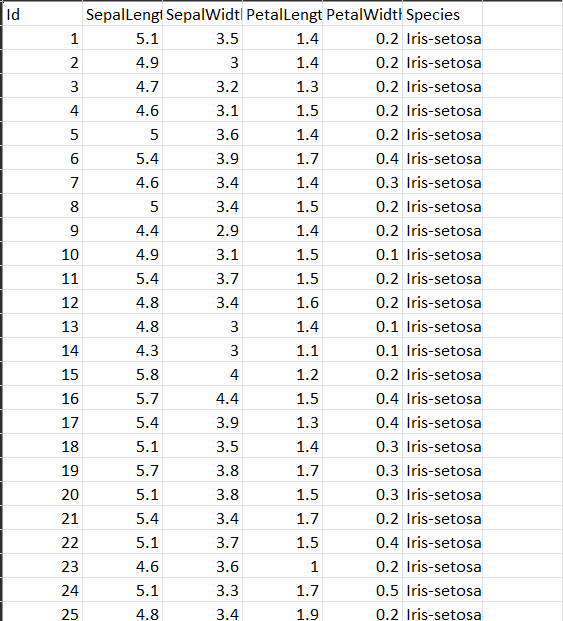
Iris Setosa

Iris Versicolor Iris Virginica

The dataset is evenly distributed, with 50 observations for each species. It is commonly used to test and evaluate classification algorithms, and it is a good dataset for beginners to practice machine learning techniques.

The iris flower dataset can be accessed through various libraries in Python, such as scikit-learn, seaborn, and pandas.

### DATA SET



### DATA AUGMENTATION

Data augmentation can be applied to the Iris dataset to increase the size of the training set and improve the performance of the model. Here are some possible data augmentation techniques that can be applied:

**Rotation**: Rotate the images of the flowers by a certain degree. This can help in making the model more robust to variations in the orientation of the flower images.

**Scaling**: Rescale the images of the flowers to different sizes. This can help in making the model more robust to variations in the size of the flower images.

**Flip**: Flip the images of the flowers horizontally or vertically. This can help in making the model more robust to variations in the orientation of the flower images.

**Translation**: Translate the images of the flowers by a certain number of pixels in different directions.

This can help in making the model more robust to variations in the position of the flower images.

**Gaussian noise**: Add some random Gaussian noise to the images of the flowers. This can help in making the model more robust to variations in the brightness of the flower images.

These data augmentation techniques can be applied to the Iris dataset to create new samples, which can be used to train a machine learning model for predicting the species of the flower. However, since the Iris dataset is relatively small, it is important to be careful not to overfit the model to the training data.

Therefore, it is recommended to use cross-validation and regularization techniques to ensure that the model generalizes well to unseen data.

### DATA VISUALIZATION

Data visualization is a powerful tool to gain insights from the Iris dataset and to understand the relationships between the features and the target variable. Here are some possible visualizations that can be used for Iris flower prediction:

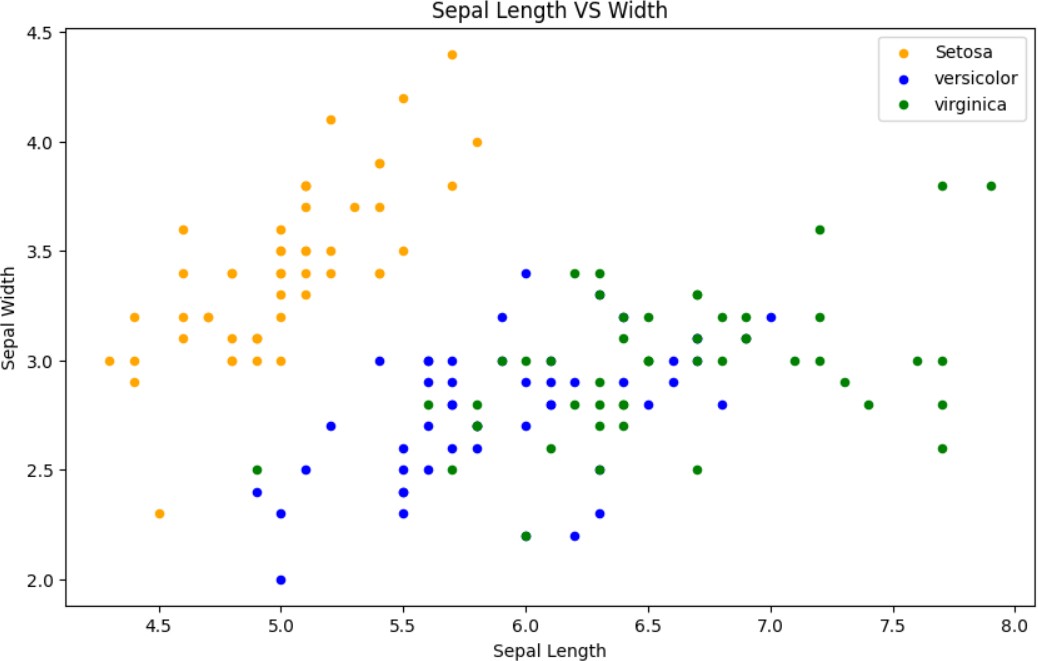
**Scatter plot**: A scatter plot can be used to visualize the relationship between two features and the target variable. For example, we can plot sepal length vs. petal length and color code the points based on the species of the flower.

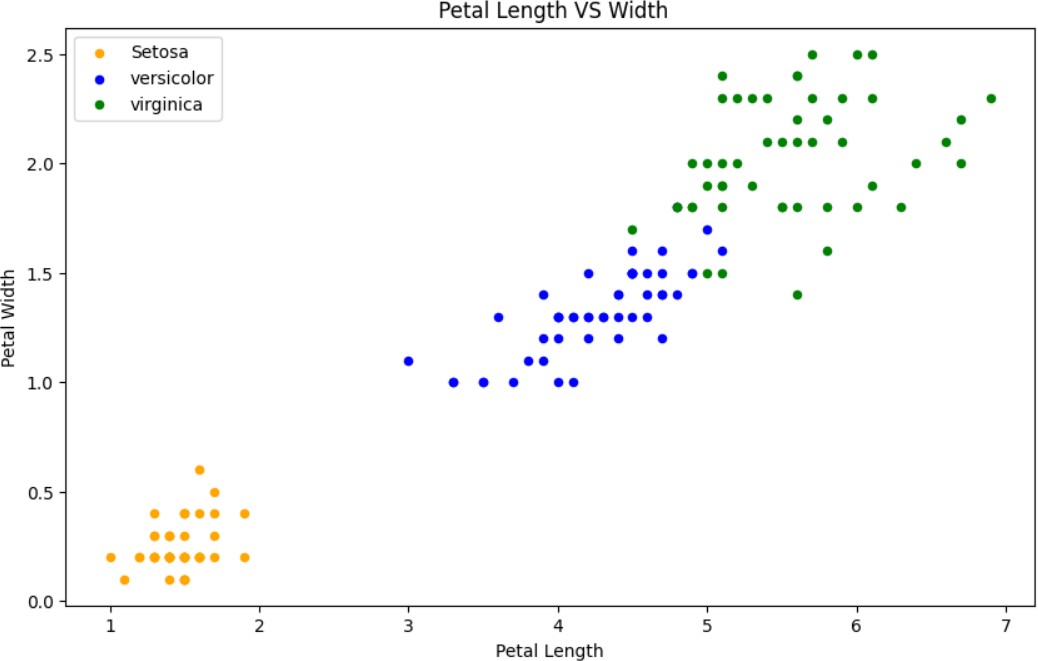
**Box plot**: A box plot can be used to visualize the distribution of a feature across different species of flowers. For example, we can plot the sepal length for each species using a box plot to see if there are any differences in the distribution.

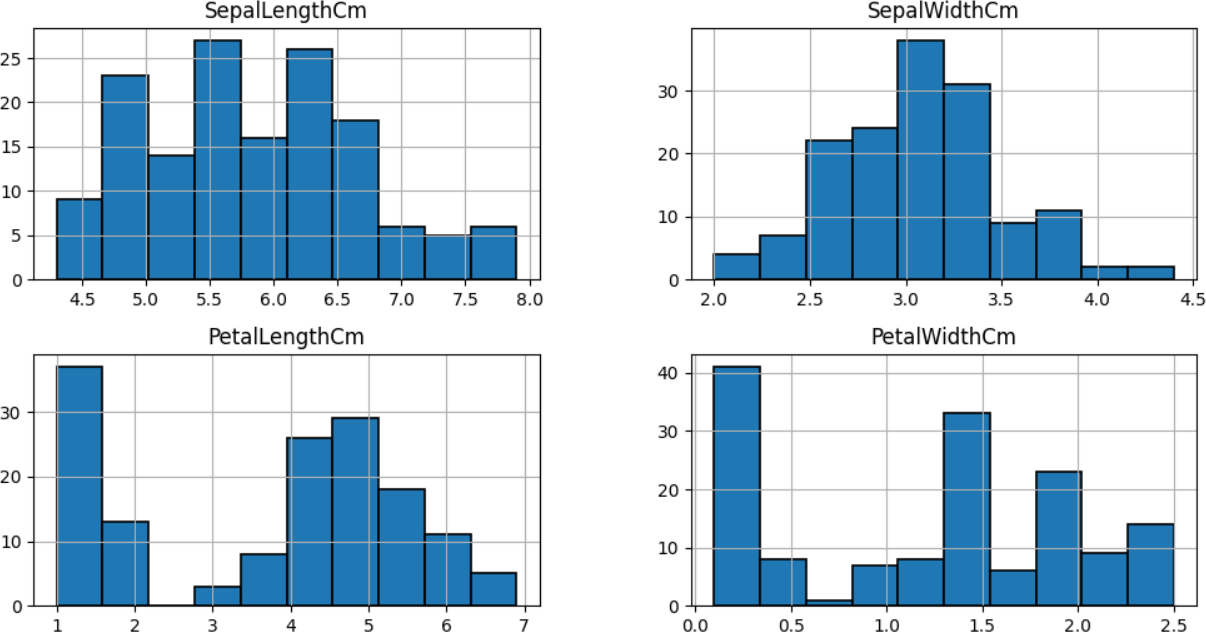
**Heatmap**: A heatmap can be used to visualize the correlation between different features. For example, we can plot a heatmap of the correlation matrix between the four features to see which features are highly correlated.

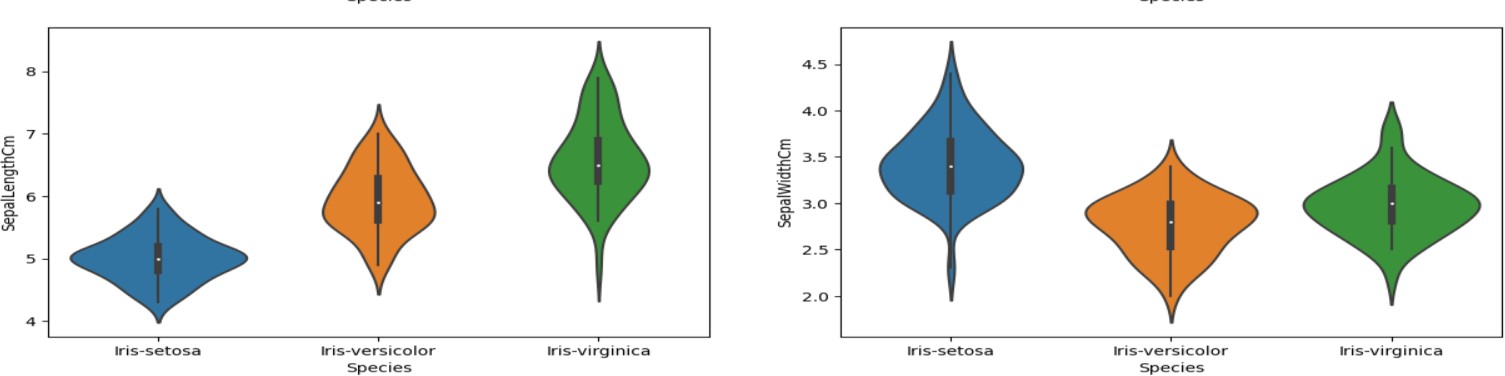
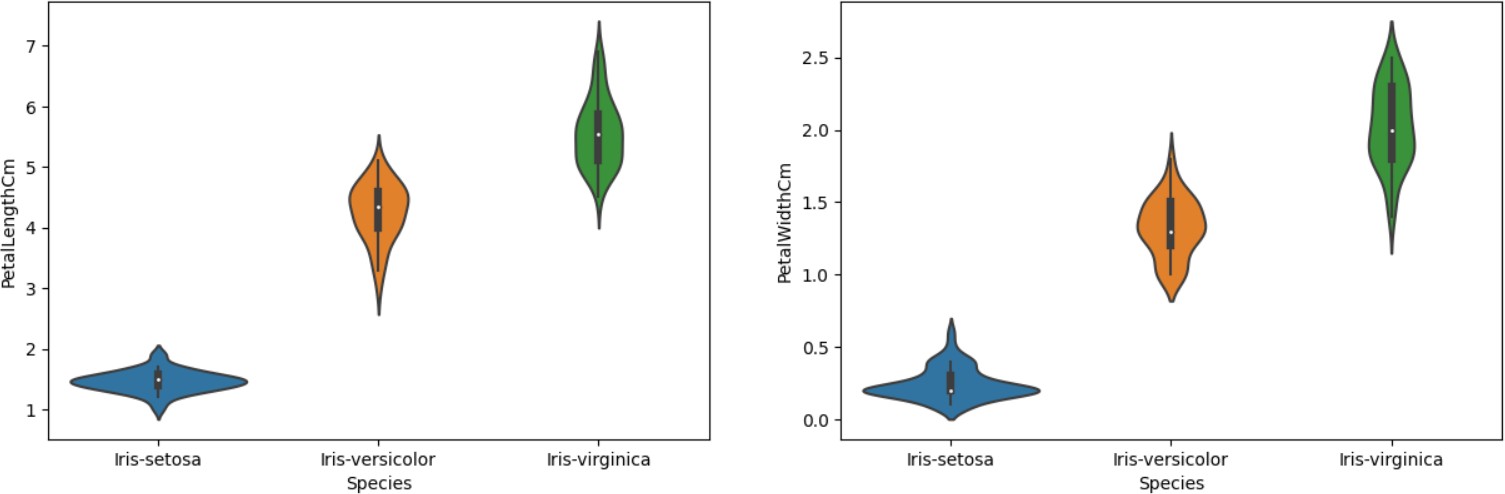
**Pair plot**: A pair plot can be used to visualize the pairwise relationships between all the features and the target variable. For example, we can plot a pair plot for all the features colored by the species to see how the features are related to each other and to the target variable.

By using these visualizations, we can gain insights into the Iris dataset and understand which features are important for predicting the species of the flower. We can also use these visualizations to identify any outliers or anomalies in the data and to decide which features to include in our machine learning model.









## METHODOLOGY

After Data pre-processing and data visualization the next step is to apply the models on the dataset. Our dataset comes under supervised learning as it contains the labeled data (target variables, feature variables). First the dataset is splitted into training set and testing set. Then the model is trained on training set and then tested on testing set.

#### Logistic regression algorithm:

Logistic regression is a machine learning algorithm which comes under supervised learning. It is a parametric method, where an equation is formed to solve. The equation returns continous values. These continues values should to converted to categorical values.so, we use a activation function called “sigmoid”.by using log error function we calculate the error.

from sklearn.linear\_model import LogisticRegression model = LogisticRegression() model.fit(train\_X,train\_y) prediction=model.predict(test\_X)

#### K-Nearest Neighbour algorithm:

K-Nearest Neighbour algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression. This algorithm is non parametric. This is also called as lazy learning algorithm. This algorithm works by first selecting the k value which is an integer value and less than the number of rows. When a new data point is given, KNN finds the nearest neighbors to that data point based on the distance using various methods like Euclidean distance or Manhattan distance. And assigns the data point to that class.

from sklearn.neighbors import KNeighborsClassifier model=KNeighborsClassifier(n\_neighbors=3) model.fit(train\_X,train\_y) prediction=model.predict(test\_X)

#### Decision Tree algorithm:

Decision tree algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. This algorithm is also known as ID3 algorithm. This algorithm is non parametric method. It forms a tree from the given dataset. It has two nodes decision nodes and leaf nodes. Decision nodes are used for taking decisions and leaf nodes are the output of that decisions. The attribute selection happens by entropy and information Gini.

from sklearn.tree import DecisionTreeClassifier model=DecisionTreeClassifier()

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

prediction=model.predict(test\_X)

#### Support Vector Machine Algorithm:

Support vector machine algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. SVM works by constructing a hyperplane or a line that separates the different classes of data points. SVM has support vectors. The distance between positive hyperplane and negative hyperplane is called margin.

from sklearn import svm model = svm.SVC() model.fit(train\_X,train\_y)

prediction=model.predict(test\_X)

## SOFTWARE DESCRIPTION

#### PYTHON

Python is an interpreted, high-level, general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and supports multiple programming paradigms, including procedural, object-oriented, and functional programming.

## GOOGLE COLAB

Google Colaboratory (also known as Google Colab or simply Colab) is a cloud-based service provided by Google that allows users to write, run, and share Python code in a web browser. It provides a Jupyter notebook environment with access to a free GPU, making it a popular tool for machine learning and data analysis tasks.

Users can create new notebooks or open existing ones from Google Drive, GitHub, or other cloud- based storage services. Colab also includes a code editor, a file browser, and a terminal interface for running shell commands.

One of the key features of Colab is its ability to run Python code on a remote server, which can be especially useful for computationally intensive tasks that require significant computing power or memory.

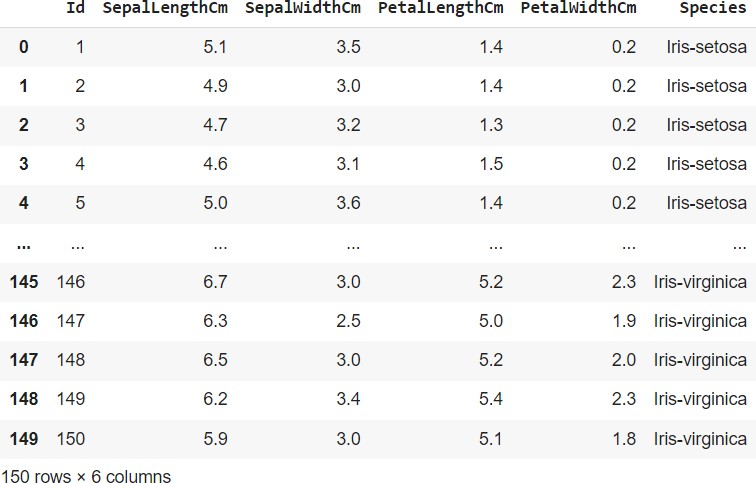
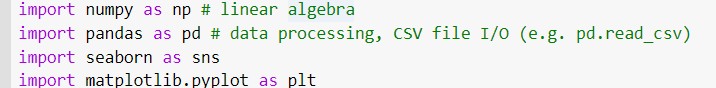
Users can install additional libraries using pip. In addition to Python code, Colab supports Markdown cells for creating formatted text, equations, and images, as well as interactive widgets for creating graphical user interfaces.

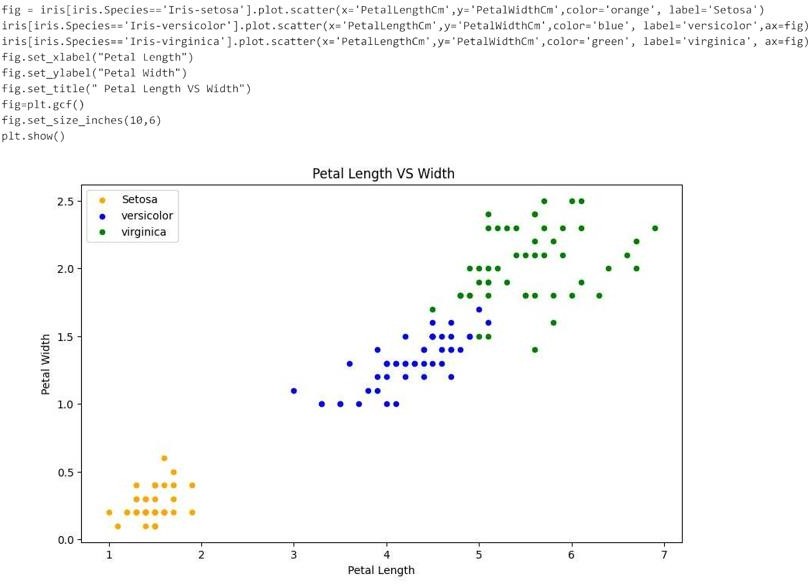
Overall, Google Colaboratory is a powerful and versatile tool for data scientists, machine learning engineers, and anyone else who needs to write and run Python code in the cloud.

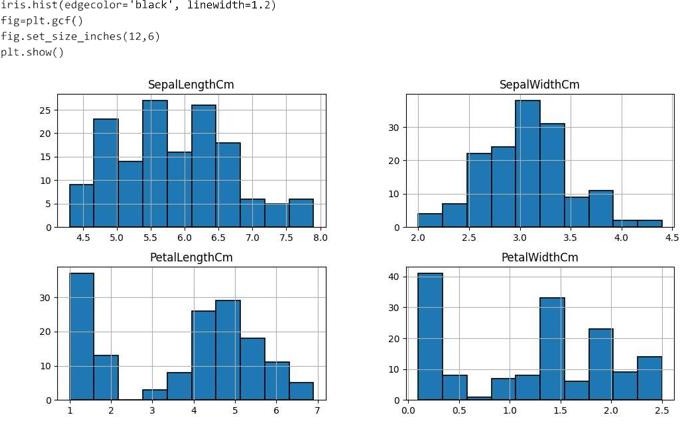
## RESULTS

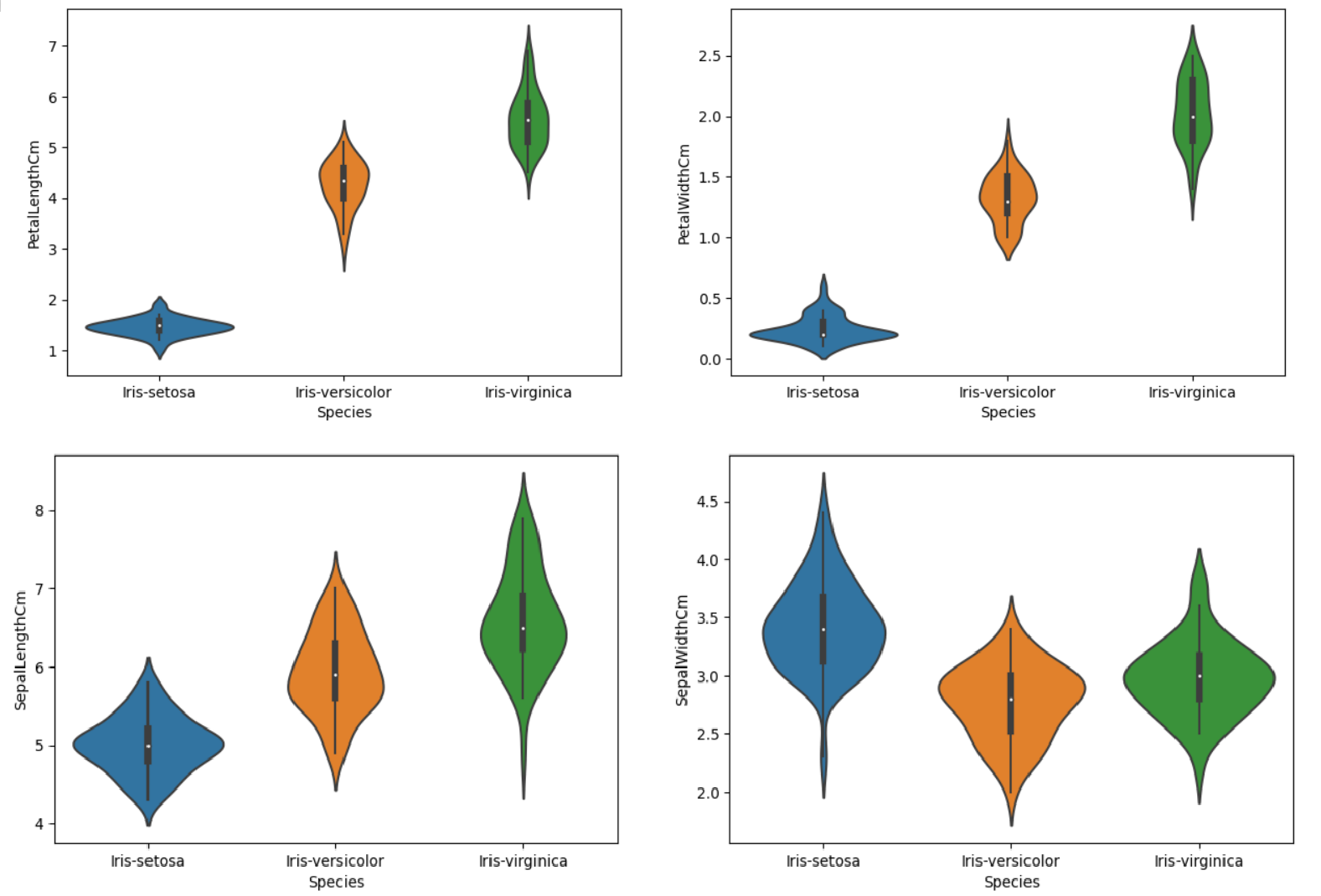
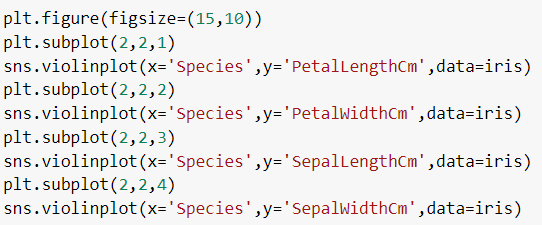
#### CODE

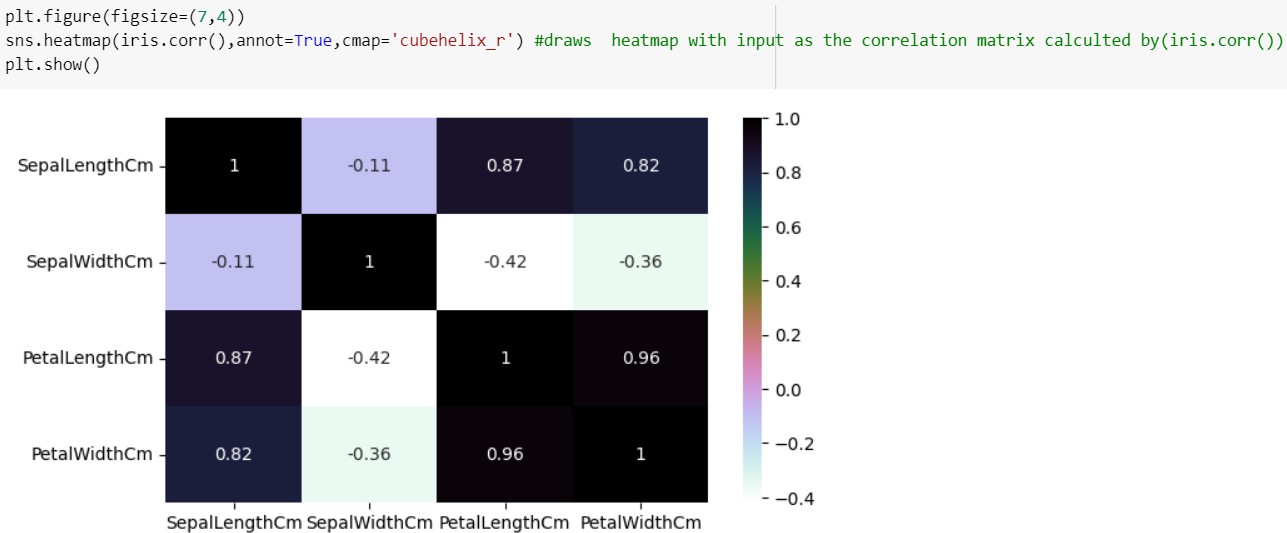


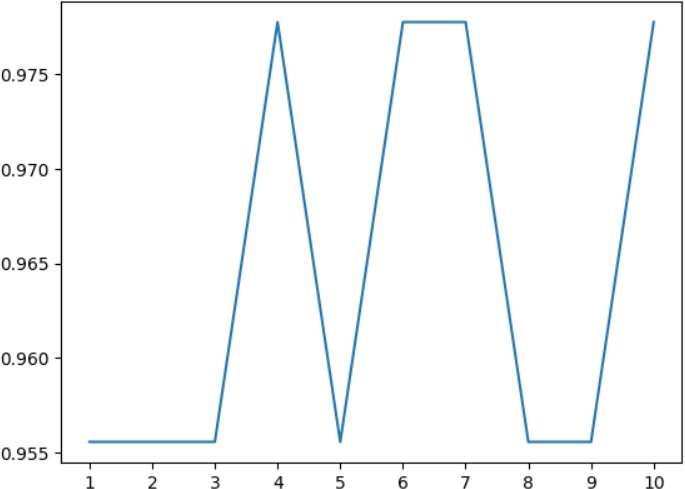
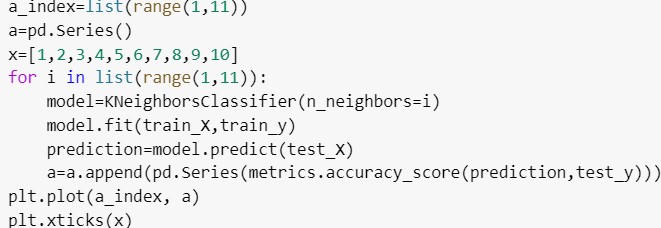
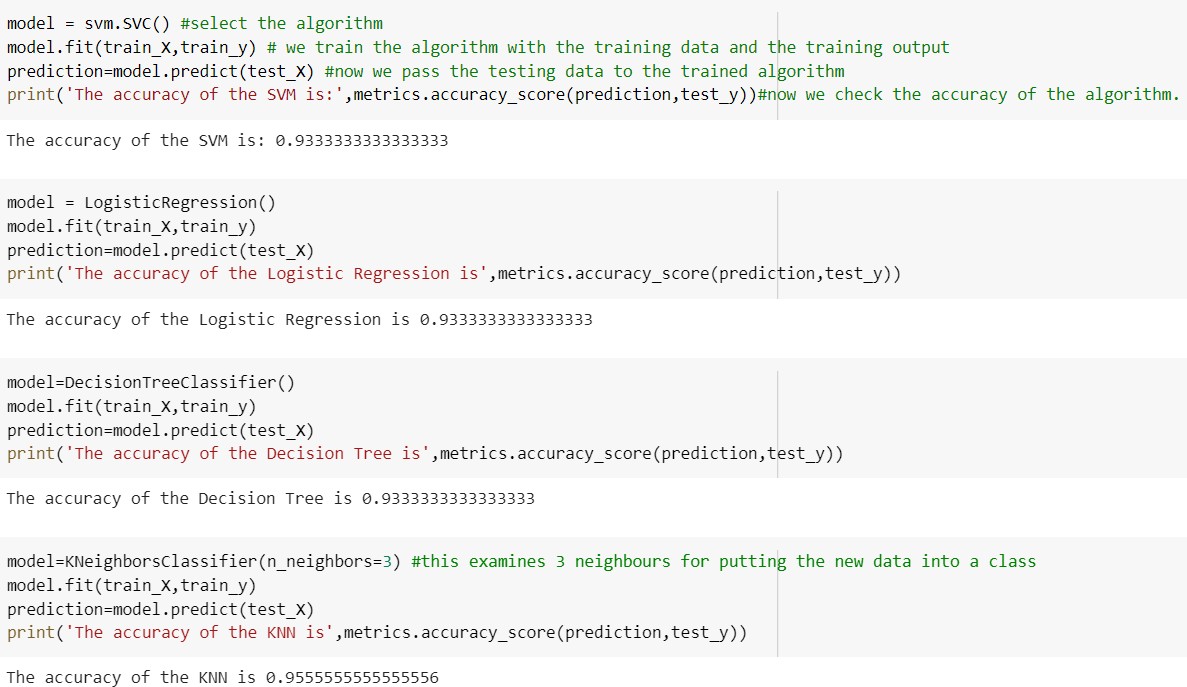












**Table 5.1 Table to represent Accuracy of all models**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **MACHINE LEARNING MODEL** | **ACCURACY** |
| 1 | Logistic regression | 0.9333333333333333 |
| **2** | K-Nearest Neighbour | 0.9555555555555555 |
| **3** | Support vector Machine | 0.9333333333333333 |
| **4** | Decision Tree | 0.9333333333333333 |

# 

# 6.CONCLUSION

In conclusion, iris flower prediction using machine learning techniques has proven to be an effective and accurate method for classifying iris flowers into different species based on their sepal length, sepal width, petal length, and petal width features. Various machine learning algorithms, such as decision trees, support vector machines, and deep learning models, have been used to achieve high prediction accuracy rates for iris flower classification tasks**.**

# 

# 7.FUTURE SCOPE

Iris flower prediction research has made significant progress and has great potential for future developments. Further research in feature engineering, model optimization, interpretability, real- world applications, and dataset diversity can contribute to the advancement of iris flower prediction techniques and their practical applications in various domains

## 

## 8.REFERENCES

https://www.kaggle.com/code/ash316/ml-from-scratch-with-iris

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>