Final Project Report

On

Iris Flower Classification

Submitted for the requirement of

Project course

BACHELOR OF ENGINEERING

COMPUTER SCIENCE & ENGINEERING

Submitted By:
Student Group
NAME
Aditya Sharma
Uday Garg
Mohd. Muaz
UID
20BCS9872
20BCS9864
20BCS9881

Submitted to:
Project Teacher (Supervisor)
(Ms. Ruby Garg &E-12271)

Radhika

Co-Supervisor Signature
(Ms. Radhika & E-12126)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
CHANDIGARH UNIVERSITY, GHARUAN
June 2022

ABSTRACT

Classification is a directed machine intelligence method that is used to conclude group participation for dossier instances. To create categorization easier, interconnected system everything has existed brought in. This Neural Network-located model focuses on Iris flower classification. The scikit gain form provisions will be used to streamline categorization. This project primarily focuses on taking advantage of scikit gain to categorize the iris dataset. The question is the determination of the Iris flower variety (Versicolor, Setosa, Virgin-ica) established the distance and breadth of the flower's whorl on flower and petal measures. We can form a categorization model by preparing the iris flower dataset accompanying various machine learning methods and therefore selecting the model accompanying the capital veracity to more precisely envision the iris flower class. The discovery of patterns from checking the whorl on flower and petal proportion of the Iris flower, in addition to by what method the forecast was made from analyzing the pattern to assemble the class of Iris flower, hopefully, used to categorize the Iris basic document file. An unseen dossier can be forecasted exactly from now on age by engaging this pattern and categorization. Pattern categorization, function approximations, addition, and select thought have all happened resolved using fake affecting animate nerve organs networks. The purpose search out pretends the odds of class enrolment established decorative traits. In this project, we'll use machine intelligence to train our model to predict the variety of iris flowers from the hidden dossier utilizing what it's well-informed from the trained dossier.

TABLE OF CONTENT

List of Figures

List of Tables

1. Introduction

1.1 Machine Learning

Machine learning (ML) is a sort of artificial intelligence (AI) that allows software applications to improve their prediction accuracy without being expressly designed to do so. Machine learning algorithms estimate new output values using historical data as input. The process of feeding a machine enough data to train and predict a likely outcome using algorithms is known as machine learning. The more processed or valuable data that is provided to the machine, the more efficient it becomes. It learns the data and constructs the prediction model when the data is difficult. It is said that the more data, the better the model, and the higher the accuracy. Machine learning can be done in a variety of ways, including supervised learning, unsupervised learning, and reinforcement learning.

1.2 Supervised Learning

Supervised learning is a method of developing artificial intelligence (AI) that involves training a computer algorithm on input data that has been labeled for a certain output. In supervised learning, the machine learning model learns from the object's features and labels. The machine knew the features of the object and the labels associated with those attributes, so supervised learning employs a set of data where the labels or desired outcomes are already known. It is possible to make predictions regarding unknown or future facts.

1.3 Classification

One of the most important data mining techniques is classification, which divides data into specified groups. It is classified as supervised learning since the classifications are defined before the data is examined. It is necessary to have some knowledge of the data in order to use all approaches to categorization. Knowing the data usually aids in the discovery of previously unknown patterns. The goal of pattern classification is to create a function that takes an input feature and outputs two or more classifications. The UCI Machine Learning Repository provided the dataset for this study. The Iris flower data set, commonly known as Fisher's Iris data set, was created by British statistician and biologist Ronald Fisher. It is a multivariate data set. An example of linear discriminant analysis is the use of various measurements in taxonomic problems. The expectation from mining the iris data set is to uncover patterns by studying the iris plant's sepal and petal size, as well as how the prediction was achieved by analyzing the pattern to predict the iris plant's class. In the coming years, different flowers will be able to be distinguished from one another utilizing classification and pattern recognition. A classification model is obviously expressed as the type of connection being mined using the iris dataset.

1.4 Software and library requirements

- python 3.7.2
- Google Colaboratory
- sklearn
- CSV
- numpy
- pandas
- matplotlib

1.5 Background

Since its invention, the computer has begun to influence our daily lives. It enhances the quality of our lives by making them more convenient and productive. Allowing a computer to think and learn like a person is an intriguing concept.
Machine learning is essentially allowing a computer to build learning skills on its own using pre-existing knowledge. Pattern recognition can be compared to a computer that can distinguish many types of items. As a result, pattern recognition and machine learning are closely linked.
The Iris flower is the subject of this project. Iris has three separate classifications in its data set: Setosa, Versicolor, and Virginica. These three separate kinds of Iris will be distinguished by the developed recognition mechanism.

1.6 Objective

After the project is completed, the computer should be able to combine three separate Iris flower classifications into three categories. The entire machine learning procedure should go seamlessly. Users do not need to inform the computer which class the Iris belongs to because the computer can recognize all of them on its own.

The ultimate goal of this effort is to provide a fundamental understanding of machine learning to anyone who reads this thesis. Even though they have never worked in this industry, they can see how the machine learning algorithm will grow in popularity and utility in the future. Furthermore, the Iris identification case study will demonstrate how to use Scikit-learn software to implement machine learning.

2. LITERATURE REVIEW

Joylin Priya Pinto uses KNN, SVM, and Logistic Regression algorithms in order to get good accuracy results. She has applied the technique of cross-validation in order to maximize the accuracy of her paper. She has focused on comparing the accuracy by including and excluding the technique of cross-validation. Three algorithms are used which are K-Nearest Neighbour, Logistic Regression, and SVM. Using these three algorithms we have found out that SVM is the most effective method among the others as it gives us the best accuracy.[1]

Patrick used the iris flower dataset to focus on the statistical analysis of it. In his paper, they are analyzing two different methods. The dataset is plotted in order to determine the various patterns in classification. Then they are able to extract statistical information by developing an application in java.[2]

Bin Shi used the classification of the iris flower problem as an example to show us that the photonic neural network concept permits us to get identical accuracy as compared to electronics. The final accuracy predicted gets reduced by 9.2%. A photonic DNN having three layers is used for image classification problems. Comprehensive error analysis suggests that a chip on optical neural networks is implemented to expect to improve photonic neural performance. An analysis is worked upon to get to know the on-chip-induced impairments.[3]

Deeptam Dutta in his work used a method to train artificial neural networks. In his paper, the data set is classified with the help of neural systems. The problem statement reviews the recognition of species based on the estimations of bloom quality assessments. The task is to search for designs by processing the sepal size and petal size of the iris. He used this example and this information can be used in the coming years. He processed those artificial neural systems can be effectively used for issues in work approximations, design arrangements, affiliated recollections, and advancement.[4]

Swain in his work using the iris data set presented the procedure of developing the artificial neural network based on a classifier that groups the iris data sets. The multilayer perceptron neural network is used by this classifier to solve the classification problem.[5]

Ettaouil used a process for the iris dataset. The proposed model is applied to the iris dataset. The iris plant is classified into three different species by using the pattern classification technique. The architecture used in neural networks by using the backpropagation algorithm is multilayer feedforward.[6]

Borovinskiy used three different neural systems procedures by connecting them. The base model and neural system is 98%. Further, the introduction of a coordinated grouping and characterization shows us that the iris dataset gives us 98.66 % precision.[7]

L.P. Gagnani and K.H. Wandra used the WEKA data mining instruments with various artificial intelligence calculations such as Naïve Bayes, multi-layer perceptron, and RBF on the iris dataset. Multilayer perceptron gives us better results of accuracy at 97.33 %.[8]

Chang in his work used bunching calculations for the order of iris flowers. Bunching is approached by the chart theory. The chart theory is utilized. He connected different sci-kit artificial intelligence apparatus with K-nearest neighbor. The calculated result on the iris data set was 96%.[9]

Kumar in his work suggested us the adjustment of system loads using Particle Swarm Optimisation. He proposed a device to push the performance of artificial neural networks in the implementation of the data sets that give us a 97.3 % accuracy.[10]

Rathee in his work showed us a feed-forward neural system model which is based on features connected to the iris data set which gave us 98.3 % accuracy. When multilayer perception is used on the iris data set it gives 98.82 % accuracy.[11]

Alejandro in his work used a hybrid system that is based on hypothetical assessments that have been already presented and assessed. A group of blocks such as uniocular weighted adder and uniocular subtractor is described to use the self-educating applications in equipment. The results indicate that self-learning and classification tasks can be performed by the proposed hybrid solution using simple digital blocks. The hypothetical block works properly if a high intensity of noise is provided at the inputs because of its hypothetical nature.[12]

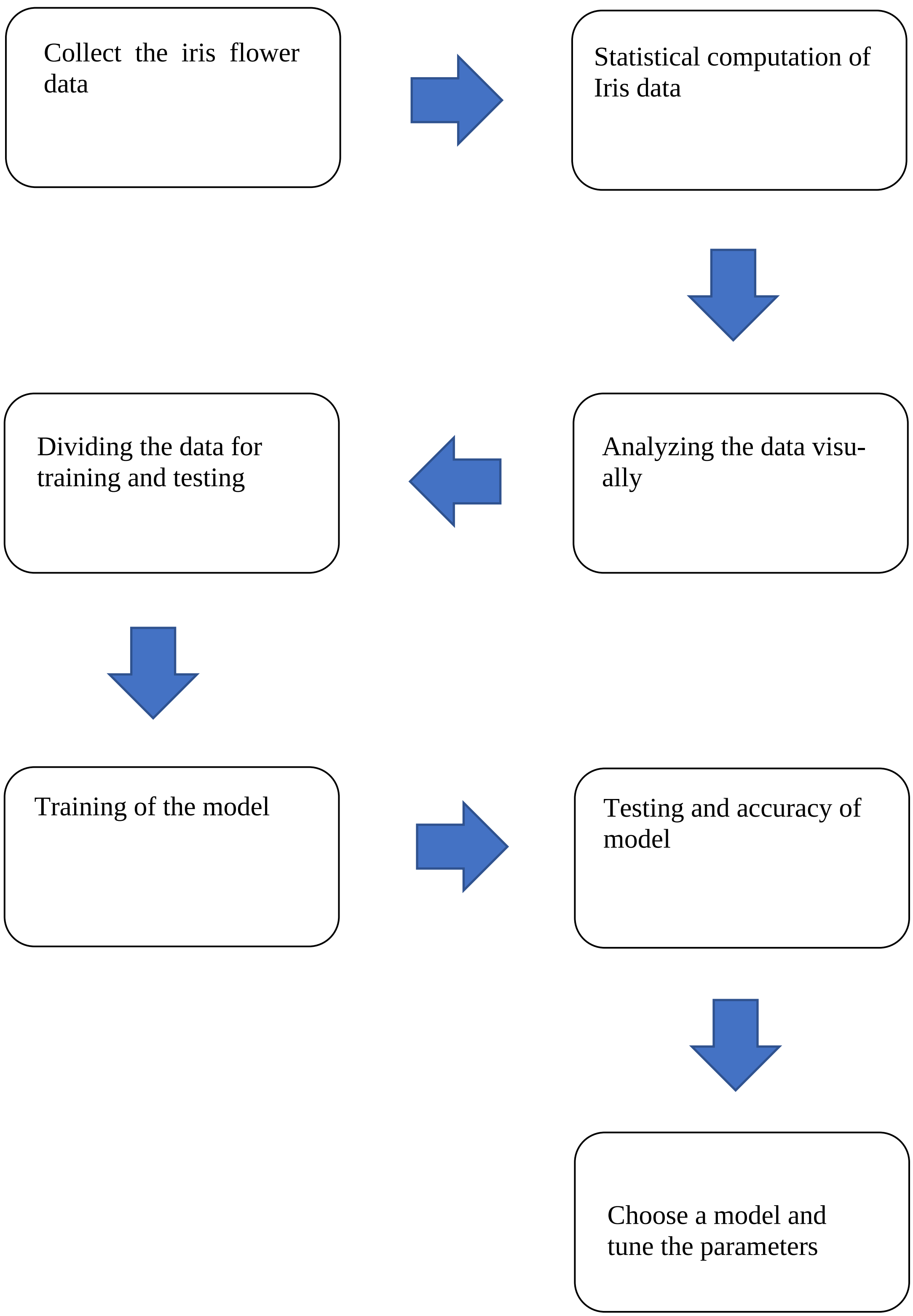
Many studies have been conducted using various strategies for the identification of the iris flower species. A different strategy is used in every study. The problem is the classification and recognition of iris flower species based on their features.

We classify the iris data set by determining patterns after inspecting the features of the iris flowers and then we predict the processing of the patterns to form the iris flower class.

After using this classification and pattern the unknown data can be predicted in the coming years more precisely. The machine learning prototype for the iris flower species technique is loaded with the dataset belonging to iris flowers.

3. Proposed Model

3.1 Block Diagram



Aditya Sharma
UID: 20BCS9872

Iris Data:

The Iris flower data set, often known as Fisher's Iris data set, is a multivariate data set provided as an example of linear discriminant analysis by British statistician and biologist Ronald Fisher in his 1936 paper The use of numerous measurements in taxonomic issues. Because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species, it is frequently referred to as Anderson's Iris data set.

This project's dataset comes from the UCI Machine Learning Repository. A multivariate data set is the Iris flower data set, often known as Fisher's Iris data set. The data set includes 50 samples from each of three Iris species (Iris virginica, Iris versicolor, and Iris setosa).

Four features were measured from each sample (in centimeters):

- Length of the petals
- Width of the petals
- Length of the sepals
- Width of the sepals

Understanding the data:

The observation data for the iris flower data set contains 150 samples. Since there are 150 samples in each of the three target classes, the data frame has four features (sepal width, sepal length, petal width, and petal length). In this stage, we'll look at the dataset's mathematics to determine the standard deviation, mean, minimum value, and four-quartile percentile. Since there are 150 samples in each of the three target classes, the data frame has four features (sepal width, sepal length, petal width, and petal length). In this stage, we'll look at the dataset's mathematics to determine the standard deviation, mean, minimum value, and four-quartile percentile.

Training and testing data:

The data we utilize will be divided into two categories: training data and test data. The training set comprises a known output, and the model learns from it in order to generalize to new data in the future. To test our model's prediction on this subset, we have the test dataset (or subset).

Training the model:

We will train our model with some of the most commonly used methods to see which approach is optimal for our model.

The algorithms that will be used are:

- K – Nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Random forest
- Logistic Regression

Aditya Sharma
UID: 20BCS9872

4. Implementation

4.1 Python

Guido van Rossum designed the Python programming language in 1989.
Python is a high-level programming language that is interpreted, object-oriented, and supports dynamic data.
The programming language is simple and straightforward, with a variety of strong classes.
Python can also easily incorporate other programming languages like C and C++.

Advantages:

- Simple & easy to learn
- Open Source
- Scalability

4.2 SciKit Learn

Scikit–learn is a Python machine learning library that is free and open source. It includes a number of classification, regression, and clustering algorithms and is designed to work with the NumPy and SciPy Python numerical libraries (Pedregosa et al. 2011). The Python-based Kmeans technique is included in SciKit-learn, which aids in figuring out how to use it in programming.

4.3Numpy, Scipy and Matplotlib

There is no such thing as an array in Python. Numpy and scipy are required libraries for analysing and calculating data in order to construct the array data type in Python.

They're all free and open source. Numpy is mostly used to calculate matrices. Scipy is a scientific research programming language built on top of numpy.

They may be utilized in Python programming with just two easy commands:
```
>>> import numpy
>>> import scipy
```
Then Python will call the methods from numpy and scipy.

Mathplotlib is a well-known Python plotting package. It comes with a set of APIs that may be used to create interactive maps. We'll apply it in this situation to find the greatest visual result.

4.4 Machine learning system design

In general, the principles of machine learning system design should follow two basic requirements:

- The model selection and creation
- The learning algorithm selection and design.

Furthermore, different models can have various learning mechanisms. On the other hand, different learning models have distinct objective functions. The machine can use the objective function to create a learning system.

Furthermore, the learning system's most crucial component would be the accuracy and complexity of distinct algorithms.

If the chosen algorithm is not particularly adaptable to the learning system, the learning system's efficiency and results will suffer.

Learning performance and feature selection can be influenced by the training data set chosen.

4.5 Setting up the environment

The virtual environment aids in the management of project dependencies. Its primary use is to provide a secure environment for Python projects.

To create a virtual environment, we must take the following steps:

1. Open the terminal
2. Setup the pip package manager
3. Create the virtual environment
4. Activate the virtual environment
5. After that install the appropriate libraries (numpy, pandas, etc.) using pip

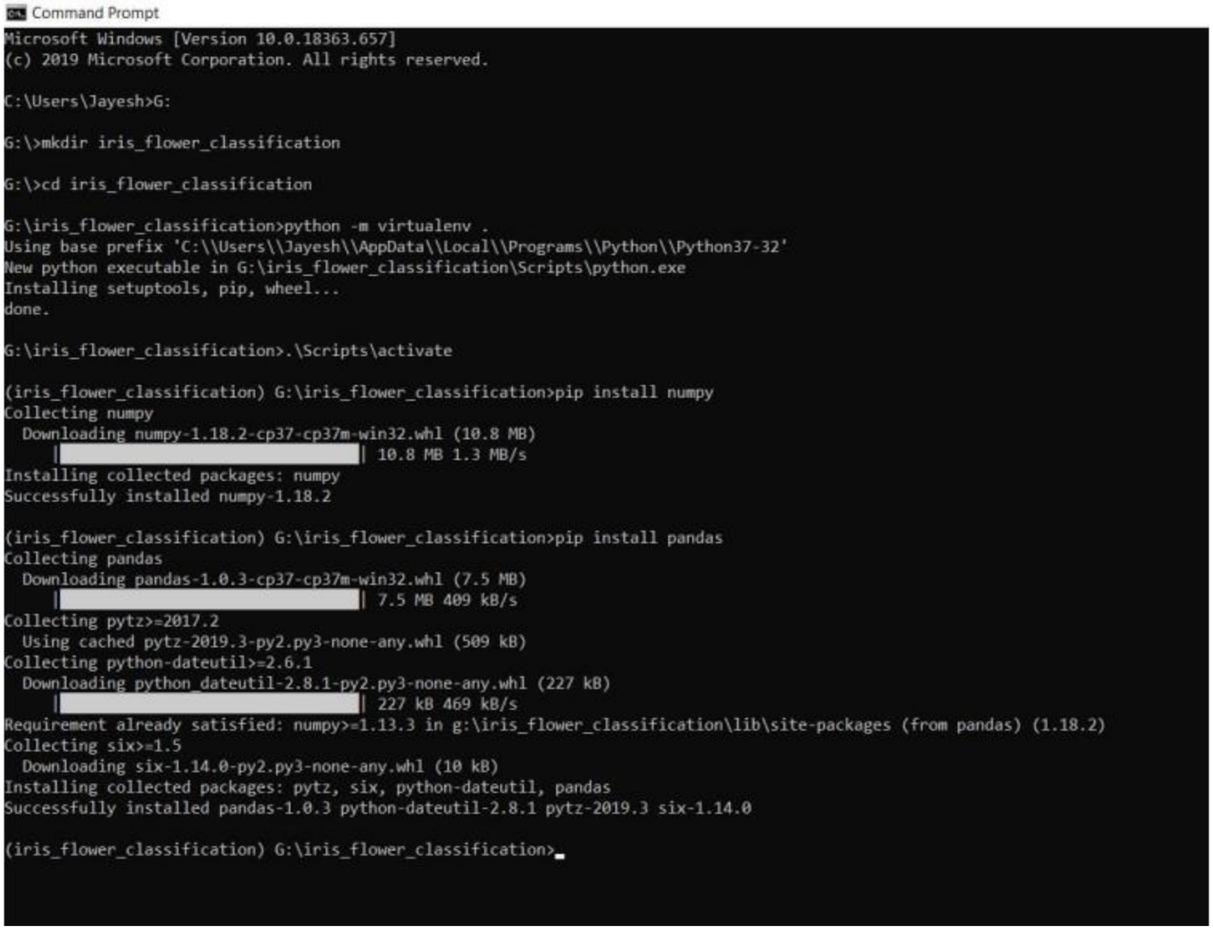


Fig.1 Process for setup virtual environment

4.6 Import the libraries

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
%matplotlib inline
```

Fig.2 Imported Libraries

The libraries NumPy, pandas, seaborn, matplotlib, and sklearn are imported here. NumPy is an array processing software that is commonly used in scientific computing. Pandas is based on the NumPy package, and the Data Frame is its primary data structure.

We may store and modify tabular data in rows of observations and columns of variables using Data Frames. Seaborn is a matplotlib-based library for statistical graphical representation and data display. Matplotlib is a visualization and charting package that may be used to create plots, histograms, bar charts, pie charts, and other graphs.

Scikit-learn provides a uniform Python interface for a variety of machine learning techniques, including both unsupervised and supervised learning algorithms.

The UCI Machine Learning Repository has the iris dataset. The data set has multiple characteristics.

This data collection includes four properties, including sepal length, sepal width, petal length, and petal width in centimeters, as well as three classifications, iris setosa, iris versicolor, and iris virginica.

The dataset downloaded from the UCI Machine Learning Repository is a CSV (Comma Separated Values) file named ‘iris.data’, which should be saved in the same directory as our project.

4.7 Data Exploration

We'll now go on to data exploration and analysis with the iris data. Let's import our data using the 'pandas' library, which will convert our data from CSV to tabular representation.

The convenience of using the pandas library is that we can read CSV files. We must add a column to the imported dataset that contains the attributes (sepal length, sepal width, petal length, petal width) in order to turn our data into a comprehensible format.

Fig 3. Types of Iris Flowers

This provides a heading for the imported data.

Connect G drive

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Load the data

```
[ ] columns = ['Sepal length', 'Sepal width', 'Petal length', 'Petal width', 'Class_labels']
    df = pd.read_csv('iris.data', names=columns)
    df.head()
```

	Sepal length	Sepal width	Petal length	Petal width	Class_labels
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Table 1: Showing Iris Dataset using pandas Library

Or we can use seaborn instead of pandas as:

```
iris=sns.load_dataset("iris")
print(iris.head())
```

```
   sepal_length  sepal_width  petal_length  petal_width 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
```

Table 2: Showing Iris Dataset using seaborn Library

4.8 Data Analysis

There are 150 samples in this collection. Because the dataframe comprises four features (petal length, petal width, sepal length, and sepal width) and 150 samples from each of the three target classes, each class has been equally distributed.

```
print(iris.groupby("species").size())
```

```
species
setosa        50
versicolor    50
virginica     50
dtype: int64
```

Table 3: Species in Iris Dataset

We can inspect the mathematics of the dataset using 'df.describe(),' which aids in determining the standard deviation, mean, lowest value, and four quartile percentile of the data.

Aditya Sharma
UID: 20BCS9872

```
df.describe()
```

	Sepal length	Sepal width	Petal length	Petal width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Table 4: Statistical description of iris dataset

4.9 Data Visualization

Visualization is an excellent approach to gain a better knowledge of your data and Python, and there are many excellent tools available for this purpose.

If you look closely at the image below, you'll notice that all of the traits are plotted against each other, and the three different colours represent the distribution of three different species
(setosa, versicolor and verginica).

It depicts the unique correlations that exist between the qualities.

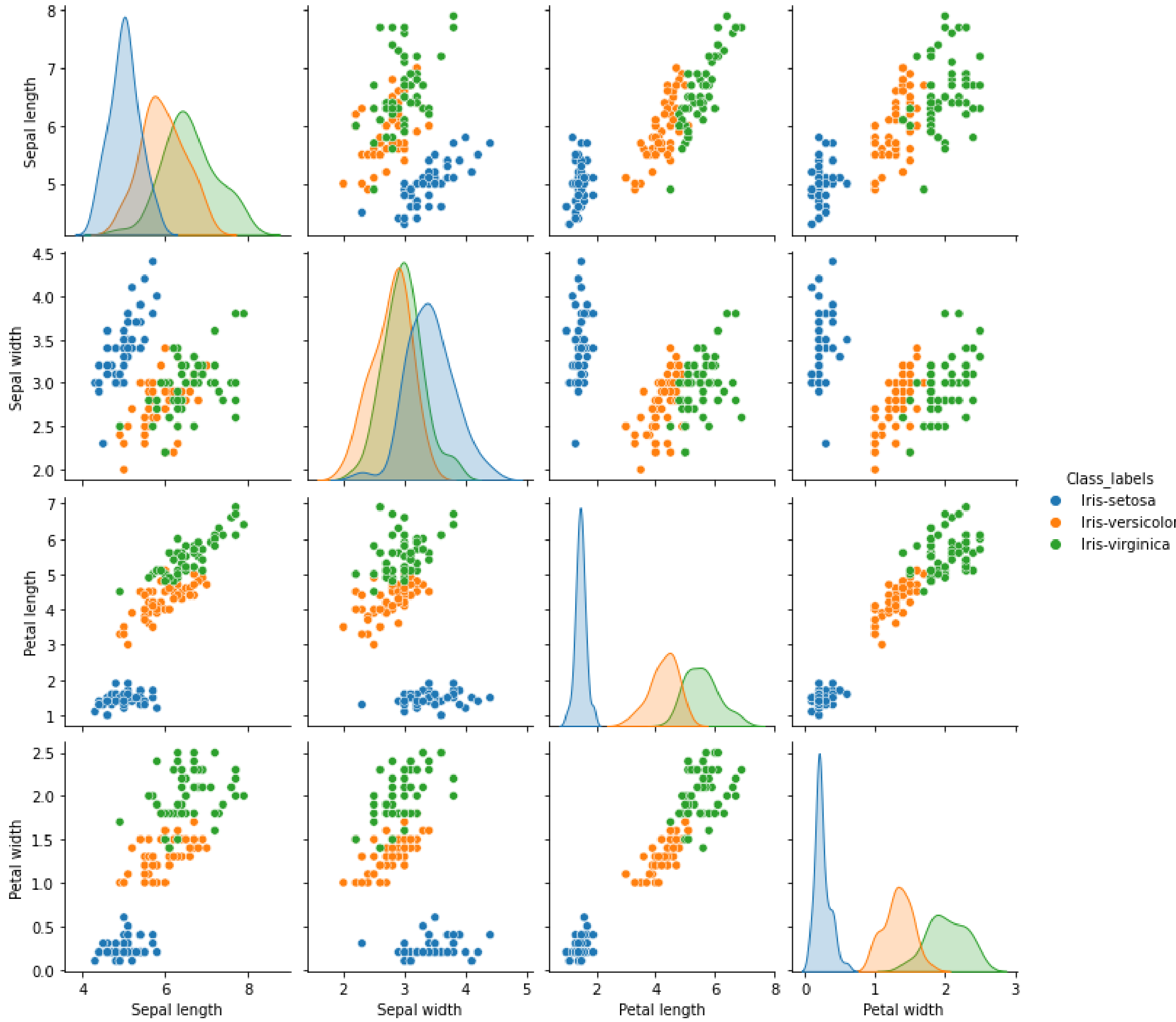


Fig.4 Pair-plot

Aditya Sharma
UID: 20BCS9872

5. Evaluating Results

The result is shown in four images for the clustering results.

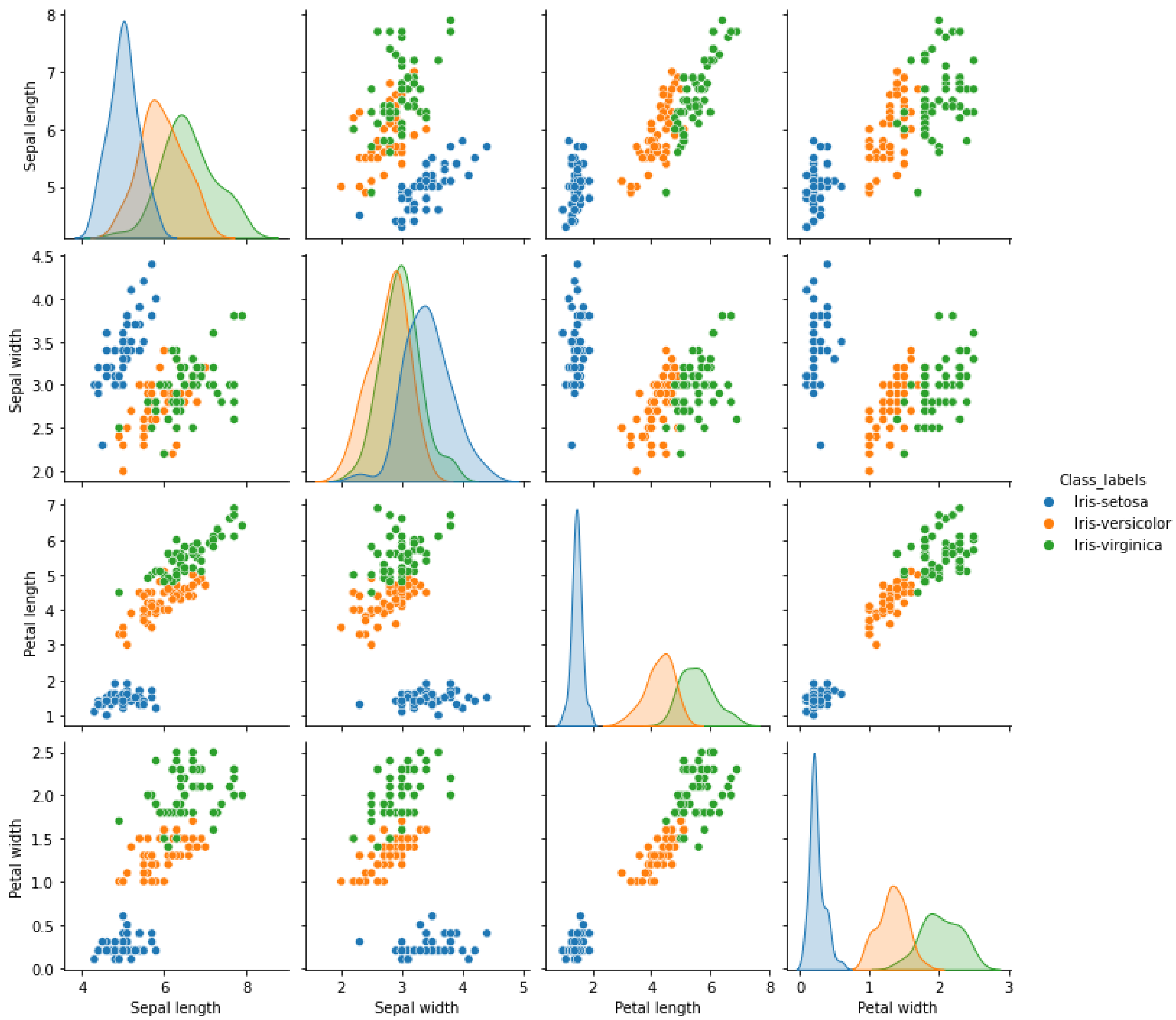


Fig.5 Result

The cluster result in Figure displays three clusters with poor initialization. In comparison to the Figure, we can observe that some of the samples alter their class.

The system will produce varied cluster outcomes with a random initialization number. As a result, for a successful cluster result, a random initialization number is critical.

Aditya Sharma
UID: 20BCS9872

However, we have no idea what a decent initialization number might be. In this instance, scientists will use GA (Genetic Algorithm) as the initiation point in various machine learning systems.

A standard result of K-means clustering of Iris recognition is shown in the figure below. In supervised learning, the term "ground truth" refers to the classification of training datasets. There are three clusters, each having a decent initiation point.

This is the most accurate classification of all the options presented. The entire dataset has been carefully divided, and each dataset contains significant differences. The standard outcome of classification in unsupervised learning is shown in Figure.

The figure has several minor modifications from this figure, but it still functions effectively. Almost every piece of information has a proper home.

These findings demonstrate the impact of the clustering outcome on the number of k and the random initialization number. The advantages and disadvantages of the K-means clustering algorithm can also be shown.

6. The Future Prospects

The above Iris identification case study demonstrates that the Machine Learning method is effective at pattern recognition. Computing is quick, and the outcome is satisfactory. However, in unsupervised learning, the K-means clustering algorithm is merely one of many clustering algorithms. In several scientific domains, there are additional algorithms for various job objectives.

Machine learning can be divided into supervised and unsupervised learning, as previously stated. However, a dataset can contain both labeled and unlabeled data at times.

A new learning method is known as Semi-supervised (SSL) Learning has become a research hotspot in order to process this type of dataset. Both machine learning and pattern recognition have a new study direction as a result of this learning method.

Labeling those vast amounts of unlabeled data saves a lot of time and human resources. The SSL has a substantial impact on computer learning performance.

Furthermore, every learning system has two components: learning and environment. The computer receives knowledge from the environment, which it then transfers, stores, and selects important information to implement various learning objectives.

As a result, rote learning, learning from instruction, learning by deduction, learning by analog, explanation-based learning, and learning by induction are all distinct learning processes. They all use different algorithms to accomplish different tasks.

The implemented scenario in this thesis is just a simple example of pattern recognition and machine learning. The K-means method utilized in this thesis is also a fundamental algorithm. The K-means technique, on the other hand, cannot be employed if the data set contains many feature dimensions and is difficult and if the learning objective is not straightforward.

GA (Genetic Algorithm), ANN (Artificial Neural Network), and other machine learning methods are becoming increasingly robust and useful in recent years. The K-means method cannot be employed in many cases because they are too simple.

Many scientists are attempting to improve machine learning algorithms' performance. The K-means has its own enhanced components. Other algorithms, such as ISODATA, EM, and Kmeans++, can be combined with K-means.

Pattern recognition can be improved with a better machine learning method. As pattern recognition technology advances, more professional and flawless machine learning algorithms are required. Machine learning has a lot of room to expand in this scenario.

Machine learning is widely employed in many domains of computer science and artificial intelligence, in addition to pattern recognition. Artificial Intelligence goods are becoming increasingly popular. People can now use Artificial Intelligence items on a daily basis.

People, for example, utilize Google search to find information, which is also based on a machine learning clustering method. Overall, machine learning appears to have a promising future.

Aditya Sharma
UID: 20BCS9872

7. Objective

This project is basically to classify iris flowers in their different types using machine learning which feeds training data in its machine learning model and then the machine learning model predicts the value that is demanded from it.

Machine learning's main goal is to find patterns in user data and then make predictions based on these and other complex patterns in order to answer business queries and solve problems. Machine learning aids in the analysis of data and the detection of trends.

In business and other sectors, machine learning is a method of data analysis that automates the process of developing data models.

Machine learning's purpose is to train a model on historical, labeled data (data with known outcomes) in order to estimate the value of a quantity based on a new data item with an unknown target value or categorization.

For example, we might wish to estimate client XYZ's lifetime worth or determine whether a transaction is fraudulent.

An analytical model must first pass two conditions before it can be used to modify how we do business.

It must, first and foremost, be adequately accurate. Second, we must be able to install it such that it can make recommendations and predictions based on data that we have access to — and do so rapidly enough that we can act on them.

To support near-real-time data collecting and scoring, we need to build a very robust data pipeline, which is why effective data engineering is such a vital complement to good data science in bringing analytics out of the lab and into the frontlines of your business.

After the project is completed, the computer should be able to combine three separate Iris flower classifications into three categories. The entire machine learning operation should go seamlessly. Users do not need to inform the computer which class the Iris belongs to; the computer is capable of recognizing all of them on its own.

The ultimate goal of this effort is to provide a fundamental understanding of machine learning to anyone who reads this thesis. Even though they have never worked in this industry, they can see how the machine learning algorithm will grow in popularity and utility in the future.

Furthermore, the Iris identification case study will demonstrate how to use Scikit-learn software to implement machine learning.

Aditya Sharma
UID: 20BCS9872

8. Conclusion

AI has been used in a variety of industries as technology has advanced. Machine learning is the most basic method for achieving AI. This thesis discusses the machine learning work concept, two different machine learning forms, and a machine learning application. A case study of Iris flower recognition is also provided to demonstrate the workflow of machine learning in pattern recognition. The definition of pattern recognition and how machine learning works in pattern recognition have been presented in this case. The unsupervised learning method employs the K-means algorithm, which is a fairly simple machine learning algorithm. The paper also demonstrates how to learn machine learning using the SciKit-learn package.

Iris Flower Dataset

	A	B	C	D	E	F
1	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
2	1	5.1	3.5	1.4	0.2	Iris-setosa
3	2	4.9	3	1.4	0.2	Iris-setosa
4	3	4.7	3.2	1.3	0.2	Iris-setosa
5	4	4.6	3.1	1.5	0.2	Iris-setosa
6	5	5	3.6	1.4	0.2	Iris-setosa
7	6	5.4	3.9	1.7	0.4	Iris-setosa
8	7	4.6	3.4	1.4	0.3	Iris-setosa
9	8	5	3.4	1.5	0.2	Iris-setosa
10	9	4.4	2.9	1.4	0.2	Iris-setosa
11	10	4.9	3.1	1.5	0.1	Iris-setosa
12	11	5.4	3.7	1.5	0.2	Iris-setosa
13	12	4.8	3.4	1.6	0.2	Iris-setosa
14	13	4.8	3	1.4	0.1	Iris-setosa
15	14	4.3	3	1.1	0.1	Iris-setosa
16	15	5.8	4	1.2	0.2	Iris-setosa
17	16	5.7	4.4	1.5	0.4	Iris-setosa
18	17	5.4	3.9	1.3	0.4	Iris-setosa
19	18	5.1	3.5	1.4	0.3	Iris-setosa
20	19	5.7	3.8	1.7	0.3	Iris-setosa
21	20	5.1	3.8	1.5	0.3	Iris-setosa
22	21	5.4	3.4	1.7	0.2	Iris-setosa
23	22	5.1	3.7	1.5	0.4	Iris-setosa
24	23	4.6	3.6	1	0.2	Iris-setosa
25	24	5.1	3.3	1.7	0.5	Iris-setosa
26	25	4.8	3.4	1.9	0.2	Iris-setosa
27	26	5	3	1.6	0.2	Iris-setosa
28	27	5	3.4	1.6	0.4	Iris-setosa
29	28	5.2	3.5	1.5	0.2	Iris-setosa
30	29	5.2	3.4	1.4	0.2	Iris-setosa

31	30	4.7	3.2	1.6	0.2	Iris-setosa
32	31	4.8	3.1	1.6	0.2	Iris-setosa
33	32	5.4	3.4	1.5	0.4	Iris-setosa
34	33	5.2	4.1	1.5	0.1	Iris-setosa
35	34	5.5	4.2	1.4	0.2	Iris-setosa
36	35	4.9	3.1	1.5	0.1	Iris-setosa
37	36	5	3.2	1.2	0.2	Iris-setosa
38	37	5.5	3.5	1.3	0.2	Iris-setosa
39	38	4.9	3.1	1.5	0.1	Iris-setosa
40	39	4.4	3	1.3	0.2	Iris-setosa
41	40	5.1	3.4	1.5	0.2	Iris-setosa
42	41	5	3.5	1.3	0.3	Iris-setosa
43	42	4.5	2.3	1.3	0.3	Iris-setosa
44	43	4.4	3.2	1.3	0.2	Iris-setosa
45	44	5	3.5	1.6	0.6	Iris-setosa
46	45	5.1	3.8	1.9	0.4	Iris-setosa
47	46	4.8	3	1.4	0.3	Iris-setosa
48	47	5.1	3.8	1.6	0.2	Iris-setosa
49	48	4.6	3.2	1.4	0.2	Iris-setosa
50	49	5.3	3.7	1.5	0.2	Iris-setosa
51	50	5	3.3	1.4	0.2	Iris-setosa
52	51	7	3.2	4.7	1.4	Iris-versicolor
53	52	6.4	3.2	4.5	1.5	Iris-versicolor
54	53	6.9	3.1	4.9	1.5	Iris-versicolor
55	54	5.5	2.3	4	1.3	Iris-versicolor
56	55	6.5	2.8	4.6	1.5	Iris-versicolor
57	56	5.7	2.8	4.5	1.3	Iris-versicolor
58	57	6.3	3.3	4.7	1.6	Iris-versicolor
59	58	4.9	2.4	3.3	1	Iris-versicolor
60	59	6.6	2.9	4.6	1.3	Iris-versicolor

61	60	5.2	2.7	3.9	1.4	Iris-versicolor
62	61	5	2	3.5	1	Iris-versicolor
63	62	5.9	3	4.2	1.5	Iris-versicolor
64	63	6	2.2	4	1	Iris-versicolor
65	64	6.1	2.9	4.7	1.4	Iris-versicolor
66	65	5.6	2.9	3.6	1.3	Iris-versicolor
67	66	6.7	3.1	4.4	1.4	Iris-versicolor
68	67	5.6	3	4.5	1.5	Iris-versicolor
69	68	5.8	2.7	4.1	1	Iris-versicolor
70	69	6.2	2.2	4.5	1.5	Iris-versicolor
71	70	5.6	2.5	3.9	1.1	Iris-versicolor
72	71	5.9	3.2	4.8	1.8	Iris-versicolor
73	72	6.1	2.8	4	1.3	Iris-versicolor
74	73	6.3	2.5	4.9	1.5	Iris-versicolor
75	74	6.1	2.8	4.7	1.2	Iris-versicolor
76	75	6.4	2.9	4.3	1.3	Iris-versicolor
77	76	6.6	3	4.4	1.4	Iris-versicolor
78	77	6.8	2.8	4.8	1.4	Iris-versicolor
79	78	6.7	3	5	1.7	Iris-versicolor
80	79	6	2.9	4.5	1.5	Iris-versicolor
81	80	5.7	2.6	3.5	1	Iris-versicolor
82	81	5.5	2.4	3.8	1.1	Iris-versicolor
83	82	5.5	2.4	3.7	1	Iris-versicolor
84	83	5.8	2.7	3.9	1.2	Iris-versicolor
85	84	6	2.7	5.1	1.6	Iris-versicolor
86	85	5.4	3	4.5	1.5	Iris-versicolor
87	86	6	3.4	4.5	1.6	Iris-versicolor
88	87	6.7	3.1	4.7	1.5	Iris-versicolor
89	88	6.3	2.3	4.4	1.3	Iris-versicolor
90	89	5.6	3	4.1	1.3	Iris-versicolor

91	90	5.5	2.5	4	1.3	Iris-versicolor
92	91	5.5	2.6	4.4	1.2	Iris-versicolor
93	92	6.1	3	4.6	1.4	Iris-versicolor
94	93	5.8	2.6	4	1.2	Iris-versicolor
95	94	5	2.3	3.3	1	Iris-versicolor
96	95	5.6	2.7	4.2	1.3	Iris-versicolor
97	96	5.7	3	4.2	1.2	Iris-versicolor
98	97	5.7	2.9	4.2	1.3	Iris-versicolor
99	98	6.2	2.9	4.3	1.3	Iris-versicolor
100	99	5.1	2.5	3	1.1	Iris-versicolor
101	100	5.7	2.8	4.1	1.3	Iris-versicolor
102	101	6.3	3.3	6	2.5	Iris-virginica
103	102	5.8	2.7	5.1	1.9	Iris-virginica
104	103	7.1	3	5.9	2.1	Iris-virginica
105	104	6.3	2.9	5.6	1.8	Iris-virginica
106	105	6.5	3	5.8	2.2	Iris-virginica
107	106	7.6	3	6.6	2.1	Iris-virginica
108	107	4.9	2.5	4.5	1.7	Iris-virginica
109	108	7.3	2.9	6.3	1.8	Iris-virginica
110	109	6.7	2.5	5.8	1.8	Iris-virginica
111	110	7.2	3.6	6.1	2.5	Iris-virginica
112	111	6.5	3.2	5.1	2	Iris-virginica
113	112	6.4	2.7	5.3	1.9	Iris-virginica
114	113	6.8	3	5.5	2.1	Iris-virginica
115	114	5.7	2.5	5	2	Iris-virginica
116	115	5.8	2.8	5.1	2.4	Iris-virginica
117	116	6.4	3.2	5.3	2.3	Iris-virginica
118	117	6.5	3	5.5	1.8	Iris-virginica
119	118	7.7	3.8	6.7	2.2	Iris-virginica
120	119	7.7	2.6	6.9	2.3	Iris-virginica

121	120	6	2.2	5	1.5	Iris-virginica
122	121	6.9	3.2	5.7	2.3	Iris-virginica
123	122	5.6	2.8	4.9	2	Iris-virginica
124	123	7.7	2.8	6.7	2	Iris-virginica
125	124	6.3	2.7	4.9	1.8	Iris-virginica
126	125	6.7	3.3	5.7	2.1	Iris-virginica
127	126	7.2	3.2	6	1.8	Iris-virginica
128	127	6.2	2.8	4.8	1.8	Iris-virginica
129	128	6.1	3	4.9	1.8	Iris-virginica
130	129	6.4	2.8	5.6	2.1	Iris-virginica
131	130	7.2	3	5.8	1.6	Iris-virginica
132	131	7.4	2.8	6.1	1.9	Iris-virginica
133	132	7.9	3.8	6.4	2	Iris-virginica
134	133	6.4	2.8	5.6	2.2	Iris-virginica
135	134	6.3	2.8	5.1	1.5	Iris-virginica
136	135	6.1	2.6	5.6	1.4	Iris-virginica
137	136	7.7	3	6.1	2.3	Iris-virginica
138	137	6.3	3.4	5.6	2.4	Iris-virginica
139	138	6.4	3.1	5.5	1.8	Iris-virginica
140	139	6	3	4.8	1.8	Iris-virginica
141	140	6.9	3.1	5.4	2.1	Iris-virginica
142	141	6.7	3.1	5.6	2.4	Iris-virginica
143	142	6.9	3.1	5.1	2.3	Iris-virginica
144	143	5.8	2.7	5.1	1.9	Iris-virginica
145	144	6.8	3.2	5.9	2.3	Iris-virginica
146	145	6.7	3.3	5.7	2.5	Iris-virginica
147	146	6.7	3	5.2	2.3	Iris-virginica
148	147	6.3	2.5	5	1.9	Iris-virginica
149	148	6.5	3	5.2	2	Iris-virginica
150	149	6.2	3.4	5.4	2.3	Iris-virginica
151	150	5.9	3	5.1	1.8	Iris-virginica

Code

DataFlair Iris Flower Classification

```
[ ] import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    import pandas as pd
    %matplotlib inline
```

Connect G drive

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Load the data

```
[ ] columns = ['Sepal length', 'Sepal width', 'Petal length', 'Petal width', 'Class_labels']
    df = pd.read_csv('iris.data', names=columns)
    df.head()
```

	Sepal length	Sepal width	Petal length	Petal width	Class_labels
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Statistical analysis about the data

```
[ ] df.describe()
```

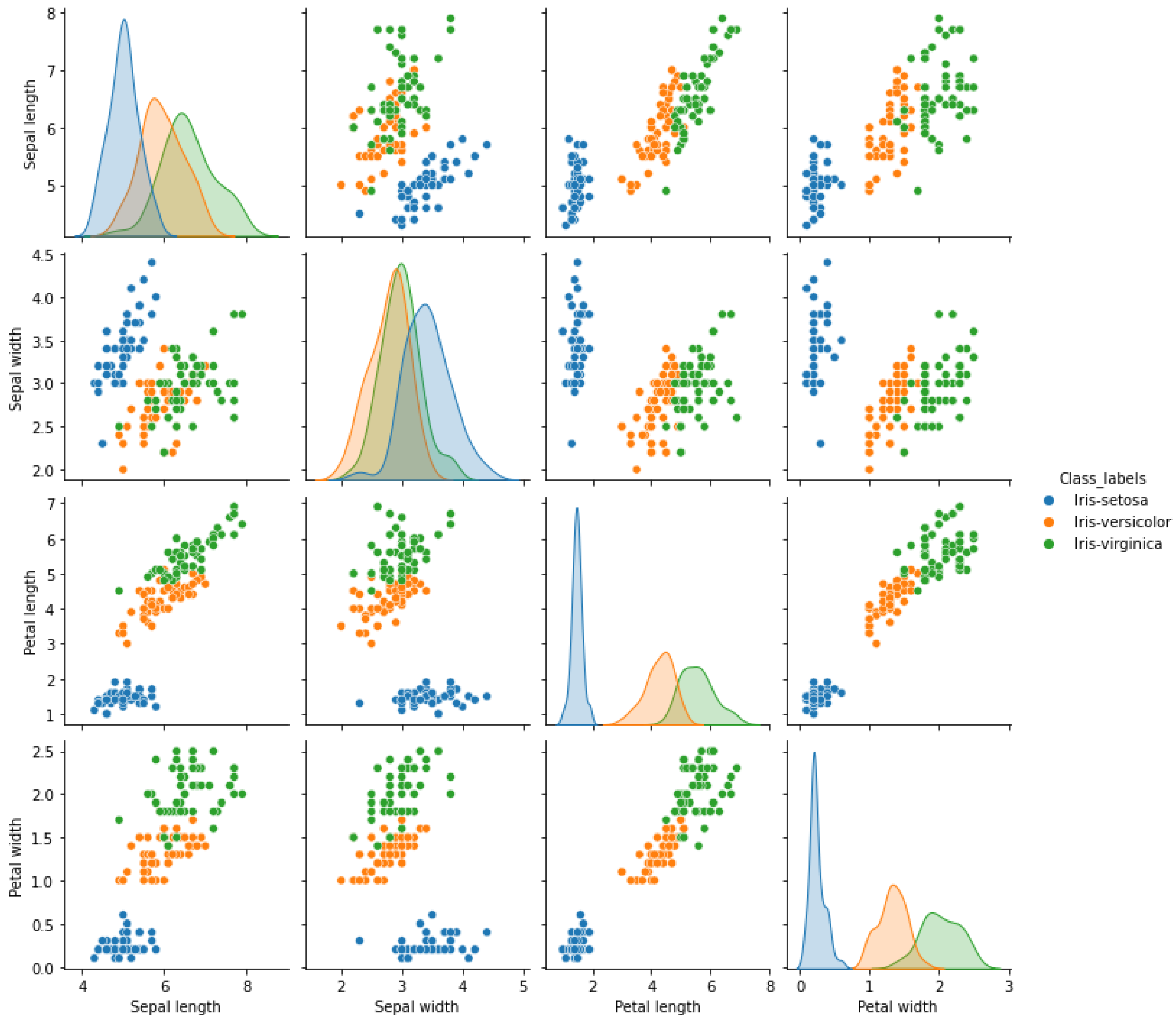
	Sepal length	Sepal width	Petal length	Petal width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Aditya Sharma
UID: 20BCS9872

Visualize the dataset

```
[ ]  sns.pairplot(df, hue='Class_labels')

     <seaborn.axisgrid.PairGrid at 0x7f9732ba1210>
```



Aditya Sharma
UID: 20BCS9872

Separating features and target

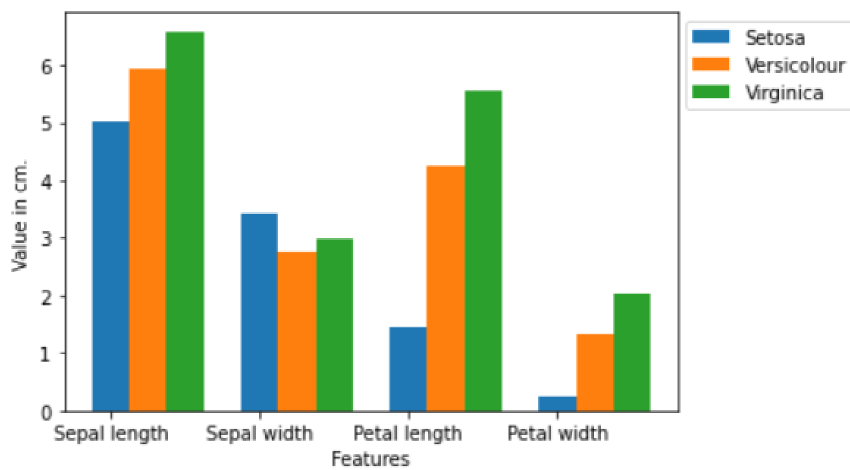
```
[ ] data = df.values
    X = data[:,0:4]
    Y = data[:,4]
```

Calculating average of each feature (for all classes)

```
[ ] Y_Data = np.array([np.average(X[:, i][Y==j].astype('float32')) for i in range (X.shape[1])
     for j in (np.unique(Y))])
    Y_Data_reshaped = Y_Data.reshape(4, 3)
    Y_Data_reshaped = np.swapaxes(Y_Data_reshaped, 0, 1)
    X_axis = np.arange(len(columns)-1)
    width = 0.25
```

Plotting the average

```
[ ] plt.bar(X_axis, Y_Data_reshaped[0], width, label = 'Setosa')
    plt.bar(X_axis+width, Y_Data_reshaped[1], width, label = 'Versicolour')
    plt.bar(X_axis+width*2, Y_Data_reshaped[2], width, label = 'Virginica')
    plt.xticks(X_axis, columns[:4])
    plt.xlabel("Features")
    plt.ylabel("Value in cm.")
    plt.legend(bbox_to_anchor=(1.3,1))
    plt.show()
```



Aditya Sharma
UID: 20BCS9872

Split the data to train and test dataset.

```
[ ] from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2)
```

Support vector machine algorithm

```
[ ] from sklearn.svm import SVC
    svn = SVC()
    svn.fit(X_train, y_train)
```

```
SVC()
```

This is formatted as code and then accuracy is calculated

```
[ ] predictions = svn.predict(X_test)

    from sklearn.metrics import accuracy_score
    accuracy_score(y_test, predictions)
```

```
0.9666666666666667
```

Detailed classification report

```
[ ] from sklearn.metrics import classification_report
    print(classification_report(y_test, predictions))
```

```
                 precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00         8
Iris-versicolor       0.92      1.00      0.96        11
 Iris-virginica       1.00      0.91      0.95        11

       accuracy                           0.97        30
      macro avg       0.97      0.97      0.97        30
   weighted avg       0.97      0.97      0.97        30
```

Prediction of the species from the input vector

```
[ ] X_new = np.array([[3, 2, 1, 0.2], [  4.9, 2.2, 3.8, 1.1 ], [  5.3, 2.5, 4.6, 1.9 ]])
    prediction = svn.predict(X_new)
    print("Prediction of Species: {}".format(prediction))
```

```
Prediction of Species: ['Iris-setosa' 'Iris-versicolor' 'Iris-versicolor']
```

Saving and loading the model

```
[ ] import pickle
    with open('SVM.pickle', 'wb') as f:
        pickle.dump(svn, f)

    with open('SVM.pickle', 'rb') as f:
        model = pickle.load(f)
    model.predict(X_new)
```

```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-versicolor'], dtype=object)
```


Aditya Sharma
UID: 20BCS9872

REFERENCES

[1] Lichman and K, Bache M. 2013. Machine Learning Repository. Irvine, CA: School of information, Computer Science and the University of California, vol. 516, 2015.

[2] K R Rathy, Arya Vaishali, "Classification of Dataset using Efficient Neural Fuzzy Approach", vol. 099, August 2019.

[3] Lynch C, "How do your data grow? By Big Data" Nature, vol. 455, no. 7209, pp. 28–29, 2008.

[4] Kaustav Choudary, Arga Roay, Dipptam Duta, "By Using Particle Swarm Optimization Algorithm Training the Artificial Neural Network", Proc. IEEE, vol. 110, July 2020.

[5] Shen Y. et al., "extreme-scale systems are measured by Silicon Phonetics," Lightw. J Technol., vol. 37, no. 2, pp. 245–259, Jan. 2019.

[6] Martin C. K. A, Douglas J. R, "Recurrent neuronal circuits in the neocortex," Current Biol., vol. 17, no. 13, pp. 496–500, 2004.

[7] Shi P, He. X, An S. "Fake iris detection using support vector machines advances in biometrics by using Statistical texture analysis", vol. 501, 2009.

[8] R.A, Fisher. 1936. "Iris Data Set by using Machine Learning Repository", vol. 116, May 2020.

[9] Wong P. H. S, Theis N. T, "A new beginning for information technology by Moore's law," Comput. Sci. Eng., vol. 19, no. 2, pp. 41–50, 2016.

[10] Hines J, "Stepping up to the summit," Comput. Sci. Eng., vol. 20, no. 2, pp. 78–82, 2018.

[11] Chhugani J et al., "CPU myth vs Debunking the 100X GPU," ACM SIGARCH Comput. Archit. News, vol. 38, no. 3, pp. 451–460, 2012.

[12] Gray J, Szalay A. "Science in an exponential world," Nature, vol. 440, no. 7083, pp. 413–414, Mar. 2006.