

# NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA



## ML LA REPORT

On

**Analyzing Decision Tree and K-means Clustering using Iris dataset**

*Submitted in partial fulfilment of the requirement for the award of Degree of*

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

*Submitted by:*

Guruprasad K

1NT19CS407

Agil Srinivasan

1NT19CS400

Under the Guidance of  
Dr. Vani V  
Professor, Dept. of CS&E, NMIT



Department of Computer Science and Engineering  
(Accredited by NBA Tier-1)

2021-22

# NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM  
, APPROVED BY AICTE & GOVT.OF KARNATAKA)

## Department of Computer Science and Engineering (Accredited by NBA Tier-1)



### CERTIFICATE

This is to certify that **Analyzing Decision Tree and K-means Clustering using Iris** is an authentic work carried out by **Guruprasad k (1NT19CS407) Agil Srinivasan (1NT19CS400)** bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of **Bachelor of Engineering** in **COMPUTER SCIENCE AND ENGINEERING** of Visvesvaraya Technological University, Belagavi during the academic year **2021-2022**. It is certified that all corrections and suggestions indicated during the internal assessment has been incorporated in the report.

Internal Guide

Signature of HOD

---

Dr. Vani V  
Professor, Dept. CSE,  
NMIT Bangalore

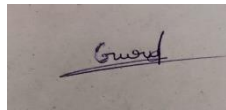
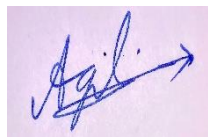
---

Dr. Sarojadevi H  
Professor, Head, Dept. CSE,  
NMIT Bangalore

## DECLARATION

We are hereby declaring  
that

- (i) The project work is our original work
- (ii) This Project work has not been submitted for the award of any degree or examination at any other university/College/Institute.
- (iii) This Project Work does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
- (iv) This Project Work does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
  - a) their words have been re-written, but the general information attributed to them has been referenced.
  - b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.
- (v) This Project Work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

NAME	USN	SIGNATURE
Guruprasad K	1NT19CS407	
Agil Srinivasan	1NT19CS400	

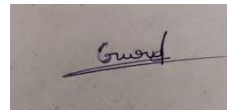
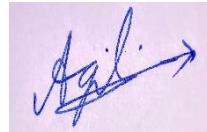
Date: 20-01-2022

## ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal Dr. H. C. Nagaraj, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD, Dr. Sarojadevi H. for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

NAME	USN	SIGNATURE
Guruprasad K	1NT19CS407	
Agil Srinivasan	1NT19CS400	

Date: 20-01-2022

## TABLE OF CONTENT

SERIAL NO.	CONTENT	PAGE NO.
	Certificate	
1	Declaration	3
2	Acknowledgement	4
3.	Chapter 1 Abstract	6
4	Chapter 2 Introduction 2.1. Motivation 2.2. Problem Domain 2.3. Aims &Objectives	7
5	Chapter 3 Data source & Data quality	9
6	Chapter 4 Data Pre-processing	10
7	Chapter 5 Machine Learning Methods	11
8	Chapter 6 Results	13
9	Chapter 7 Conclusion and Future Prospects	16
10	Chapter 8 References	17

## **CHAPTER 1: ABSTRACT**

As we all know from the nature, most of creatures have the ability to recognize the objects in order to identify food or danger. Human beings can also recognize the types and application of objects. An interesting phenomenon could be that machines could recognize objects just like us someday in the future.

This thesis mainly focuses on machine learning in pattern recognition applications. Machine learning is the core of Artificial Intelligence and pattern recognition is also an important branch of AI. In this thesis, the conception of machine learning and machine learning algorithms are introduced. Moreover, a typical and simple machine learning algorithm called K-means is introduced. A case study about Iris classification is introduced to show how the K-means works in pattern recognition.

## **CHAPTER 2: Introduction**

Machine learning, as a powerful approach to achieve Artificial Intelligence, has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

### **2.1. MOTIVATION**

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives, it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition.

In this project, the object is the Iris flower. The data set of Iris contains three different classes: Setosa, Versicolour, and Virginica.

The designed recognition system will distinguish these three different classes of Iris.

### **2.2.Problem Domain**

How many of we have actually seen Iris flowers, particularly those three species included in the dataset, in gardens or pictures? I doubt the number can be high because based on the responses from my surrounding friends and colleagues, only a small portion of them told me that they had seen Iris followers before. Certainly, I had to show them some pictures of Iris, otherwise many of them probably didn't even know what Iris flowers were.

All this aside, we know that the three Iris species in the dataset: Iris setosa, Iris virginica, and Iris versicolor, and we also know that the dataset records the lengths and widths of sepals and petals for these flowers

## 2.3. Aims and Objectives

### Aims:

- The aim of the case study is to design and implement a system of pattern recognition for the Iris flower based on Machine Learning.
- This project shows the workflow of pattern recognition and how to use machine learning approach to achieve this goal.
- The data set was collected from an open source website of machine learning.
- The programming language used in this project was Python.

### Objectives:

- After the project has been settled, the computer should have the ability to aggregate three different classifications of Iris flower to three categories.
- The whole workflow of machine learning should work smoothly.
- The users do not need to tell the computer which class the Iris belongs to, the computer can recognize them all by itself.



## CHAPTER 3: Data source and Data quality

### Data Source:

We will use the dataset from the Kaggle. The dataset is obtained from the following link –

<https://www.kaggle.com/uciml/iris>

### Data Quality :

The data set contains 3 classes with 50 instances each, and 150 instances in total, where each class refers to a type of iris plant.

Class : Iris Setosa, Iris Versicolour, Iris Virginica

The format for the data: (sepal length, sepal width, petal length, petal width)



We will be training our models based on these parameters and further use them to predict the flower classes.

## **CHAPTER 4: Data Pre-processing**

Using an inbuilt library called ‘train\_test\_split’, which divides our data set into a ratio of 80:20. 80% will be used for training, evaluating, and selection among our models and 20% will be held back as a validation dataset.

The data set of Iris flower can be also found in the Scikit-learn library. In site packages, there is a folder named sklearn. In this folder, there is a datasets subfolder to contain many kinds of data sets for machine learning study. The data set can be found in Appendix 1. In the species of this table, 0 represents setosa, 1 represents versicolor, 2 represents virginica.

## Chapter 6: Machine Learning Methods

### Supervised Learning:

Supervised machine learning algorithms are trained to find patterns using a dataset. The process is simple, It takes what has been learned in the past and then applies that to the new data. Supervised learning uses labelled examples to predict future patterns and events.

For example – when we teach a child that  $2+2=4$  or point them to the image of any animal to let them know what it is called.

### Supervised learning is further divided into:

- **Classification:** Classification predicts the categorical class labels, which are discrete and unordered. It is a two-step process, consisting of a learning step and a classification step. There are various classification algorithms like – “Decision Tree Classifier”, “Random Forest”, “Naive Bayes classifier” etc.
- **Regression:** Regression is usually described as determining a relationship between two or more variables, like predicting the job of a person based on input data X. Some of the regression algorithms are: “Logistic Regression”, “Lasso Regression”, “Ridge Regression” etc.

### Decision Tree Classifier:

The general motive of using a Decision Tree is to create a training model which can be used to predict the class or value of target variables by learning decision rules inferred from prior data(training data).

It tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

### Unsupervised Learning:

Unsupervised learning is used against data without any historical labels. The system is not subjected to a pre-determined set of outputs, correlations between inputs and outputs or a “correct answer.” The algorithm must figure out what it is viewing by itself, as it does not have any storage of reference points. The goal is to explore the data and find some sort of patterns or structures.

Unsupervised learning can be classified into:

- **Clustering:** Clustering is the task of dividing the population or data points into several groups, such that data points in a group are homogenous to each other than those in different groups. There are numerous clustering algorithms, some of them are – “K-means clustering algorithms”, “mean shift”, “hierarchal clustering”, etc.
- **Association:** An association rule is an unsupervised learning method that is used for finding the relationships between variables in a large database. It determines the set of items that occurs together in the dataset

### **K-means Clustering:**

The goal of the K-means clustering algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of the K groups based on the features that are provided.

The outputs of executing a K-means on a dataset are:

- K centroids: Centroids for each of the K clusters identified from the dataset.
- Labels for the training data: Complete dataset labelled to ensure each data point is assigned to one of the clusters.

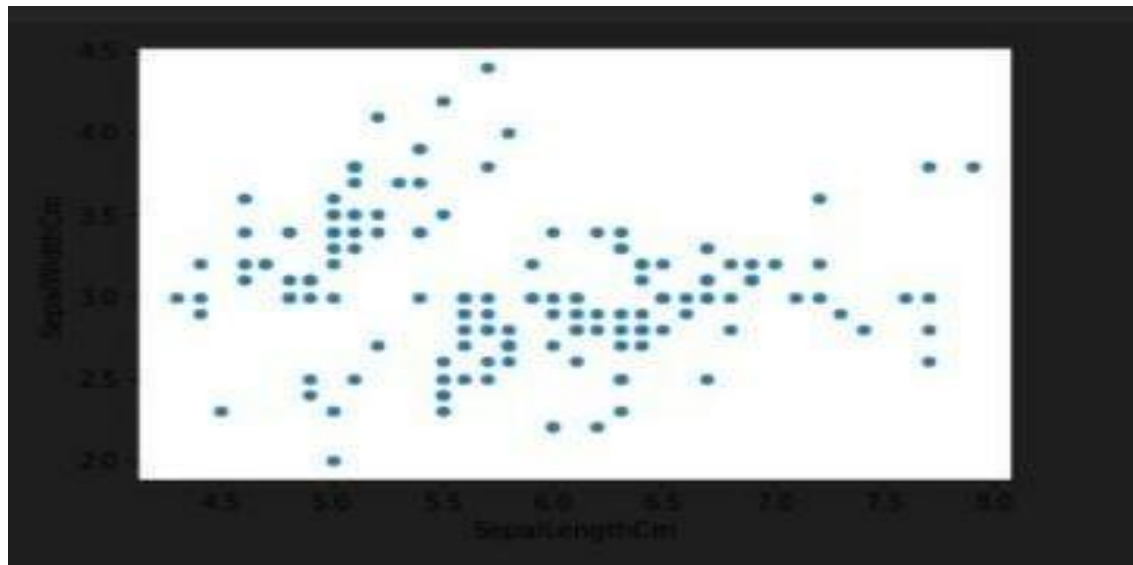
## CHAPTER 6: Results and Discussions

...	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

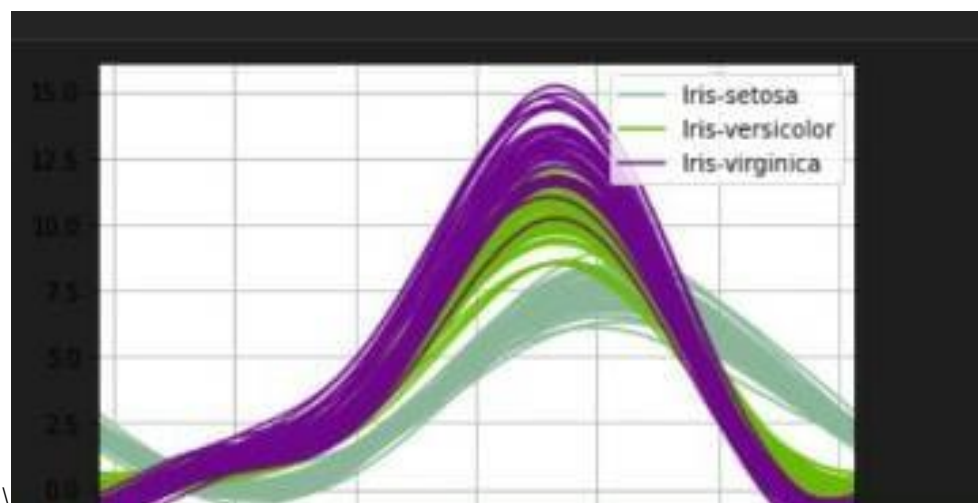
Sample data from the iris dataset, it contains the features such as Sepal length, Sepal width, Petal length, Petal width and we have taken 5 sample rows.

...	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

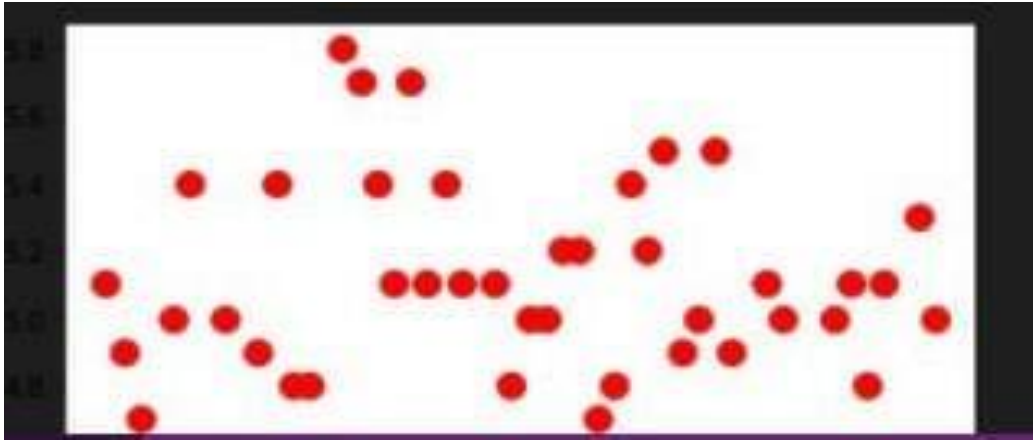
Iris.describe(), this method describes the columns and row values of iris dataset.



Data visualization is the presentation of data in an accessible manner through visual tools like graphs or charts. These visualizations aid the process of communicating insights and relationships within the data, and are an essential part of data analysis



Data visualization is done using Andrew curves, where it represents the curve status of iris. Purple curve represents Virginica, Green represents Versicolor, Cyan represents Setosa.



Clustering of data



Representation of dataset using Confusion matrix, an insight we can get from the matrix is that the model was very accurate at classifying Setosa and Virginica (True Positive/All = 1.0). However, accuracy for Versicolor was lower ( $13/14=0.928$ ).

## **CHAPTER 7: Conclusion and Future Prospects**

### **Conclusion**

With the rapid development of technology, AI has been applied in many fields. Machine learning is the most fundamental approach to achieve AI. This thesis describes the work principle of machine learning, two different learning forms of machine learning and an application of machine learning. In addition, a case study of Iris flower recognition to introduce the workflow of machine learning in pattern recognition is shown. In this case, the meaning of pattern recognition and how the machine learning works in pattern recognition has been described. The K-means algorithm, which is a very simple machine learning algorithm from the unsupervised learning method is used.

### **Future Prospects**

The Iris recognition case study above shows that the Machine Learning algorithm works well in this pattern recognition. The speed of computing is fast and the result is acceptable. However, the K-means clustering algorithm is just one of the clustering algorithms in unsupervised learning. There are more algorithms for different work objectives in different scientific fields.



## CHAPTER 8: References

### References:

- <https://www.kaggle.com/sixteenpython/machine-learning-with-iris-dataset>
- <https://towardsdatascience.com/exploring-classifiers-with-python-scikit-learn-iris-dataset-2bcb490d2e1b>
- <https://scikit-learn.org/stable/>
- <https://certes.co.uk/types-of-artificial-intelligence-a-detailed-guide/>
- <https://www.kaggle.com/uciml/iris>