IRIS FLOWER CLASSIFICATION
BY
USING MACHINE LEARNING TECHNIQUES

ARDENT
COMPUTECH PVT. LTD.
High-End Technology Training and Project

SUBMITTED BY:

ABHIJIT SAU
SUFAL DUTTA
KAUSTAV BHADURI
SOHAM SAHA
KRISHNENDU CHOWDHURY

PROJECT GUIDE:

SOFIKUL MULLICK

Submission Date: 2 JULY ,2019

CERTIFICATE FROM SUPERVISOR

This is to certify that Abhijit Sau,Sufal Dutta,Kaustav Bhaduri,Soham Saha,Krishnendu Chowdhury successfully completed the project titled **“Iris Flower Classification using Machine Learning”** under my supervision during the period from June to July which is in partial fulfilment of the requirements for the award of B.Tech and submitted to CSE department of, Netaji Subhash Engineering College,Garia,Kolkata.

Date: ----------------------------

Signature of Supervisor

Sofuikul Mullick

(Project Engineer,

Ardent Collaboratioins Pvt.Ltd.)

CONTENTS

Acknowledgement

The achievement that is associated with the successful completion of any task would be incomplete without mentioning the names of those people whose endless cooperation made it possible.

We take this opportunity to express our deep gratitude towards our project mentor, ***Mr. Sofikul Mullick*** for giving such valuable suggestions, guidance and encouragement during the development of this project work.

Last but not the least we are grateful to all the faculty members of Ardent Computech Pvt. Ltd. for their support.

ABSTRACT

The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. With the emerging increase of diabetes, that recently affects around 346 million people, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision making process is generated.

Iris is a genus of 260–300 species of flowering plants with showy flowers. It takes its name from the Greek word for rainbow, which is also the name for the Greek goddess of the rainbow, Iris. Three Iris varieties are used in the Iris flower data set outlined by Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. The data set consists of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

In this project, data mining methods, and decision tree classifier is used to predict the species of a new sample of Iris flower. Decision Tree Classifier are considered as helpful methods for the iris flower classification project. It is, in fact, probable model which have been proved useful in displaying complex systems and showing the relationships between different entities.

INTRODUCTION

The dataset for this project originates from the UCI Machine Learning Repository. The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.

. This program applies basic machine learning (classification) concepts on *Fisher's Iris Data* to predict the species of a new sample of Iris flower.

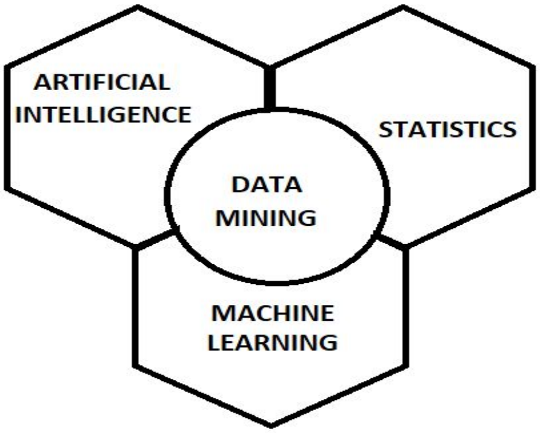
- The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).
- Four features were measured from each sample (in centimetres):
  - Length of the sepals
  - Width of the sepals
  - Length of the petals
  - Width of the petals

DATA MINING

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.

Data mining is also known as data discovery and knowledge discovery.

Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of Machine Learning, statistics, and database systems.



RELATIONSHIP DIAGRAM OF DATA MINING AND MACHINE LEARNING

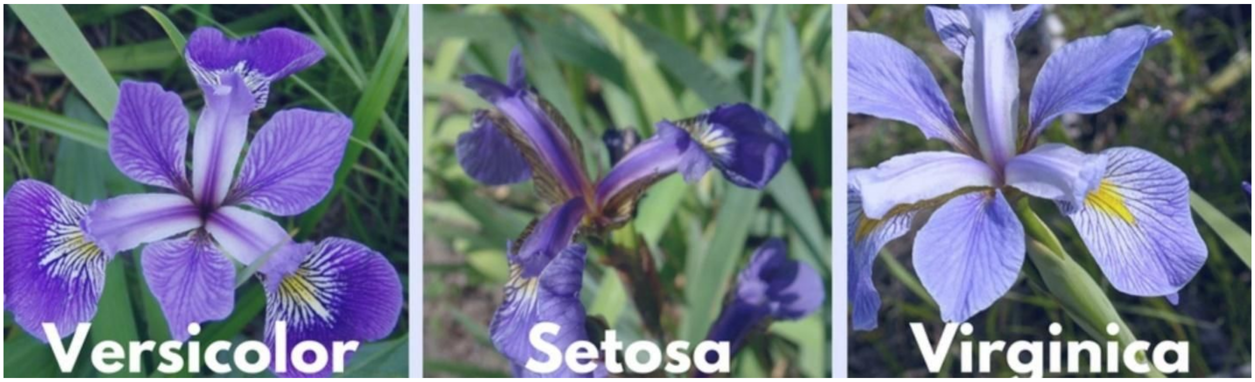
IRIS FLOWER

Iris is a genus of 260–300 species of flowering plants with showy flowers. It takes its name from the Greek word for a rainbow, which is also the name for the Greek goddess of the rainbow, Iris. Some authors state that the name refers to the wide variety of flower colors found among the many species.

As well as being the scientific name, iris is also widely used as a common name for all Iris species, as well as some belonging to other closely related genera. A common name for some species is 'flags', while the plants of the subgenus Scorpiris are widely known as 'junos', particularly in horticulture. It is a popular garden flower.

CLASSIFICATION OF IRIS FLOWER:

- ⇛ VERSICOLOR
- ⇛ SETOSA
- ⇛ VERGINICA



MACHINE LEARNING

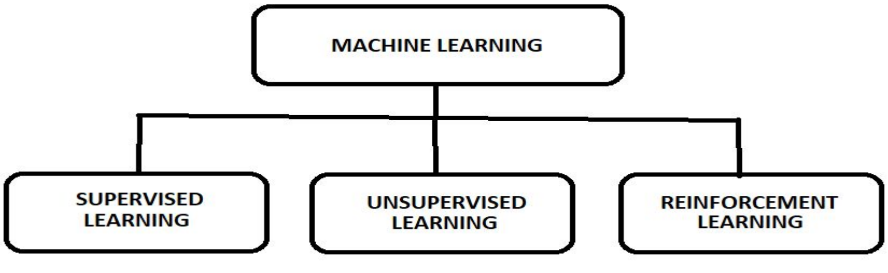
Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Two definition of Machine Learning is offered.Arthur Samuel described it as:"the field of study gives computers the ability to learn without being explicitly programmed."

Tom Mitchell provides a more modern definition:"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P,if its performance at tasks in T,as measured by P,improves with experience E."

CLASSIFICATION OF MACHINE LEARNING:



Supervised Learning :

In Supervised Learning, we are given a data set and already know what our correct output should look like,having the idea that there is a relationship between the input and the output.

Supervised Learning problems are categorised into "Regression" and "Classification". In a Regression Problem,we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a Classification problem,we are instead trying to predict results in a discrete output. In other words,we are trying to map input variables into discrete categories.

Unsupervised Learning:

Unsupervised Learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.

We can derive this structure by clustering the data based on relationships among the variables in the data. With unsupervised learning there is no feedback based on the prediction results.

Reinforcement Learning:

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Reinforcement learning is considered as one of three machine learning paradigms, alongside supervised learning and unsupervised learning.

PROJECT OBJECTIVE

- The main objective of this project is to predict the species of a new sample of Iris flower.
- The proposed work focuses on to predict species of a new sample of Iris flower using Decision tree classifier.
- This integrated technique of classification gives a promising classification results with utmost accuracy rate.
- For detecting a sample of Iris flower a dataset is required.
- But using data mining techniques the number of test should be reduced.This reduced test plays an important role in time and performance.
- The program divides the dataset into training and testing samples in 80:20 ratio randomly using train_test_learn() function available in sklearn module.
- The training sample space is used to train the program and predictions are made on the testing sample space.
- Accuracy score is then calculated by comparing with the correct results of the training dataset.
- The program creates a decision tree based on the dataset for classification
- The central goal here is to design a model which makes good classifications for new flowers or, in other words, one which exhibits good generalization.

PROBLEM STATEMENT & DESCRIPTION

PROBLEM STATEMENT:

Create a model that can classify the different species of the Iris flower and check the accuracy of the algorithm.

DATASET DESCRIPTION:

Dataset contains three classes (Iris-sentosa, Iris-versicolor and Iris-virginica) of iris flower You can clearly see that the sizes of sepals and petals of flowers varies with the class. Hence based on these features we are going to classify the flowers. You can see that there are five columns in the dataset. Each representing following attribute. Column Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

HARDWARE & SOFTWARE REQUIREMENTS

Hardware Requirements:

1.Intel Core i5 (7th generation)
2. 4gb DDR3 Ram
3. Hard Disk
4.Intel HD Graphics

Software used:

1.Anaconda Navigator(Jupyter Notebook)

PROJECT

```
#Program takes data from the load_iris() function available in sklearn module
    from sklearn.datasets import load_iris
    import numpy as np
    from sklearn import tree
# Prints the name of iris species from the predicted number
    def decode(num):
        for i in num:
            if i==0:
                print("setosa")
            elif i==1:
                print("versicolor")
            else:
                print("virginica")
#Load the data
    iris = load_iris()
    test_ids = []

    for i in range (0, 20):
        test_ids.append(i)
    for i in range (50, 70):
        test_ids.append(i)
    for i in range (100, 120):
        test_ids.append(i)

# Training data
    train_data = np.delete(iris.data, test_ids, axis=0)
    train_target = np.delete(iris.target, test_ids)

#User is then asked to enter the four parameters of his sample
#Prediction about the species of the flower is printed to the user
```

```
d1 = float(input("Enter sepal length : "))
d2 = float(input("Enter sepal width : "))
d3 = float(input("Enter petal length : "))
d4 = float(input("Enter petal width : "))

data = [[d1, d2, d3, d4]]
decode(clf.predict(data))

clf = tree.DecisionTreeClassifier()
clf.fit(train_data, train_target)
prediction=clf.predict(train_data)

from sklearn.metrics import accuracy_score as acs
print(acs(prediction, train_target))
```

EXPERIMENT RESULT

```
.9666747
Enter sepal length : 7.5
Enter sepal width : 5.5
Enter petal length : 2.5
Enter petal width : .8
versicolor
```

Summary

⇒The program creates a decision tree based on the dataset for classification.

⇒The user is then asked to enter the four parameters of his sample and prediction about the species of the flower is printed to the user.

⇒The variables are:

- **sepal_length**: Sepal length, in centimeters, used as input.
- **sepal_width**: Sepal width, in centimeters, used as input.
- **petal_length**: Petal length, in centimeters, used as input.
- **petal_width**: Petal width, in centimeters, used as input.
- **class**: Iris Setosa, Versicolor or Virginica, used as target.

⨯ The program then divides the dataset into training and testing samples in 80:20 ratio randomly using train_test_learn() function available in sklearn module.

⨯ The training sample space is used to train the program and predictions are made on the testing sample space.

Accuracy score is then calculated by comparing with the correct results of the training dataset.

CONCLUSION

Lately, machine learning has gained in interest by the scientific and research communities. Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper.The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. The data set consists of 50 samples from each of three species of Iris: Iris setosa, Iris virginica, Iris versicolor.Using decision tree classification approach, a decision tree is created based on the dataset given and the four parameters, i.e., petal length, petal width, sepal length, sepal width are entered by the user which are eventually used by the decision tree approach to predict the species of the new sample of iris flower.The project is made more efficient by introducing accuracy check by KNN approach.

Results have been obtained. Moreover, we recommend the proposed models to be tested on a larger dataset.

BIBLIOGRAPHY

- UCI Machine Learning Repository. Iris Data Set
- Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- https://github.com/Pranav-Rastogi/Iris-flower-classification