GIFT UNIVERSITY
Converting Knowledge into Practical Experience

DS-312 Applications of Data Science

Fall 2022

Assignment 2

Assignment Report for:
Machine Learning on Iris Dataset

NAME: MUHAMMAD ABDULLAH
ROLL NO: 201980007

Group Members:

Name	Roll no.
Sufyan Ahmad	201980013
Abdul basit	201980014
Zain Ramzan	201980028

PROBLEM STATEMENT

The iris data set consists of the physical parameters of three species of flower — **Versicolor**, **Setosa** and **Virginica**. The numeric parameters which the dataset contains are Sepal width, Sepal length, Petal width and Petal length. In this data we will be predicting the classes of the flowers based on these parameters. The data consists of continuous numeric values which describe the dimensions of the respective features. you will be training the ML models based on these features.

Abstract:

Isir Flovers Species:

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. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris Flower of three related species. Two of the three species were collected in Gaspe Peninsula all from the same pasture, and picked on the same day and measured at the same time by the same person with same apparatus.

The data set consists of 50 samples from each of three species of Iris that is 1) Iris Setosa 2) Iris Virginica 3) Iris Versicolor. Four features were measured from each sample. They are 1) Sepal Length 2) Sepal Width 3) Petal Length4) Petal Width. All these four parameters are measured in Centimeters. Based on the combination of these four features, the species among three can be predicted.

IMPLEMENTATION OF ALGORITHMS

1. K-Nearest Neighbors Algorithm

The k-Nearest Neighbors algorithm (or KNN for short) is an easy algorithm to understand and to implement, and a powerful tool to have at your disposal. The implementation will be specific for classification problems and will be demonstrated using the Iris flowers classification problem.

2. Decision Tree Algorithm

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes

Datasets Uses in this Assignment:

- # Iris Dataset

Information of data

- **Number of Attributes:** 4 numeric, predictive attributes and the class
- **Number of Instances:** 150 (50 in each of three classes)
- **Missing Values?** NO
- **Associated Tasks:** Classification

- **Attribute Information**
    1. **sepal length in cm**
    2. **sepal width in cm**
    3. **petal length in cm**
    4. **petal width in cm**
    5. **class:**
        - **Iris Setosa**
        - **Iris Versicolour**
        - **Iris Virginica**

Summary Statistics:

	Min	Max	Mean	SD	ClassCorrelation	
sepal length:	4.3	7.9	5.84	0.83	0.7826	
sepal width:	2.0	4.4	3.05	0.43	-0.4194	
petal length:	1.0	6.9	3.76	1.76	0.9490	**(high!)**
petal width:	0.1	2.5	1.20	0.76	0.9565	**(high!)**

Libraries used in this Assignment:

Pandas:

Pandas is a software library written for the Python programming language for data manipulation and analysis.

Scikit-learn:

Scikit-learn is an open source machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms and is designed to interoperate with the Python numerical libraries NumPy and SciPy (Pedregosa et al. 2011). Scikit-learn contains the K-means algorithm based on Python and it helps to figure out how to implement this algorithm in programming.

Numpy, Scipy and Matplotlib

In Python, there is no data type called array. In order to implement the data type of array with python, numpy and scipy are the essential libraries for analyzing and calculating data. They are all open source libraries. Numpy is mainly used for the matrix calculation. scipy is developed based on numpy and it is mainly used for scientific research.

By using them in Python programming, they can be used with two simple commands:

```
import numpy
import scipy
```

Then Python will call the methods from numpy and scipy.

Mathplotlib is a famous library for plotting in Python. It provides a series of API and it is suitable for making interactive mapping. In this case, we need to use it to find the best result visulisly.

The End