**MACHINE LEARNING MINI PROJECT**

Every machine learning project begins by understanding what the data and drawing the objectives. While applying machine learning algorithms to your data set, you are understanding, building and analyzing the data as to get the end result.

Following are the steps involved in creating a well-defined ML project:

1] Understand and define the problem

2] Prepare the data

3] Explore and Analyse the data

4] Apply the algorithms

5] Reduce the errors

6] Predict the result

**PROBLEM STATEMENT :-**

This 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. We will be training the model based on these features.

**TOOLS AND LIBRARIES :-**

1. Python version 3.0 and above.
2. Jupyter or any other IDES.
3. Pandas library importing.
4. Matplotlib library importing.
5. Sklearn library for model selection.

**IRIS DATASET DESCRIPTION :-**

**Features**:

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

**Class:**

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

**DATA ANALYSIS :-**

Data analysis is a big task in machine learning most of the time in building a model is taken by this process. Because, for machine learning models training and testing of data used should be clean and filtered according to model criteria.

Generally, cleaning and filtering of data is done by using Numpy , Pandas and Matplotlib libraries.

**MODEL SELECTION :-**

For given iris dataset the class/label values are discrete and characterised so its better to use classification models to achieve better performance of a model. I have selected a KNN model and Decision Tree model.

1. **KNN Model :-**

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
* It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
* **How it works:-**

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbours

Step-2: Calculate the Euclidean distance of K number of neighbours

Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.

Step-4: Among these k neighbours, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.

Step-6: Our model is ready.

1. **Decision Tree Classification Model :-**

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
* **How it works :-**

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**PERFORMANCE COMPARISION BETWEEN MODELS :-**

In model selection I have taken 70 percent of data as training set and remaining 30 percent as test set.

So, the accuracy of KNN algorithm is 100 percent and the accuracy of Decision Tree algorithm is 96.666 percent.

Hence we can conclude that for given iris dataset KNN model perform better than the Decision Tree model for 70 and 30 percent ratio of training and test data.