Assignment & Projects

Instructions

- Please submit the assignment & projects before 14th October, 2020, 08:00 PM. It should be submitted to LMS only, submission through other means will not be accepted.
- Any submission after that will not be accepted.
- All assignment & projects should be submitted in a single file (No separate file for each assignment & projects)
- File should be in .ipynb format only. Any other format will not be accepted.
- Copy and Paste in any form is not recommended and will have negative impact on the evaluation of project.
- New idea and concept will be highly appreciated.
- Practice: Exercise 1 is for practice and is not Considered for Evaluation.

Assignment 1:

- Consider a case of 50 student, out of which 35Passed (P) and 15 Failed (F) in the Certification Exam. Write a program to compute Entropy for the given information.

- If we have the information that some of the candidates have undergone training and some didn't. Based on it split the dataset, let say 20 people undergone the training and 30 didn't. Out of this 20 trained candidate 18P and 2F, while in case of 30 untrained 17P and 13F. Write a program to calculate the information gain after split of data.

Project 1:

From Data given it is required to develop the model to predict the profit of the company. Dataset file name: Companies-profit.csv

Task:

1) Load the dataset
2) Analyse the data visually and give justification for selecting particular algorithm for developing the model.
3) Split it into test and train.
4) Develop a model for predicting the profit of the company
5) Measure all the parameters of the model
6) Measure the performance of the model

Project 2:

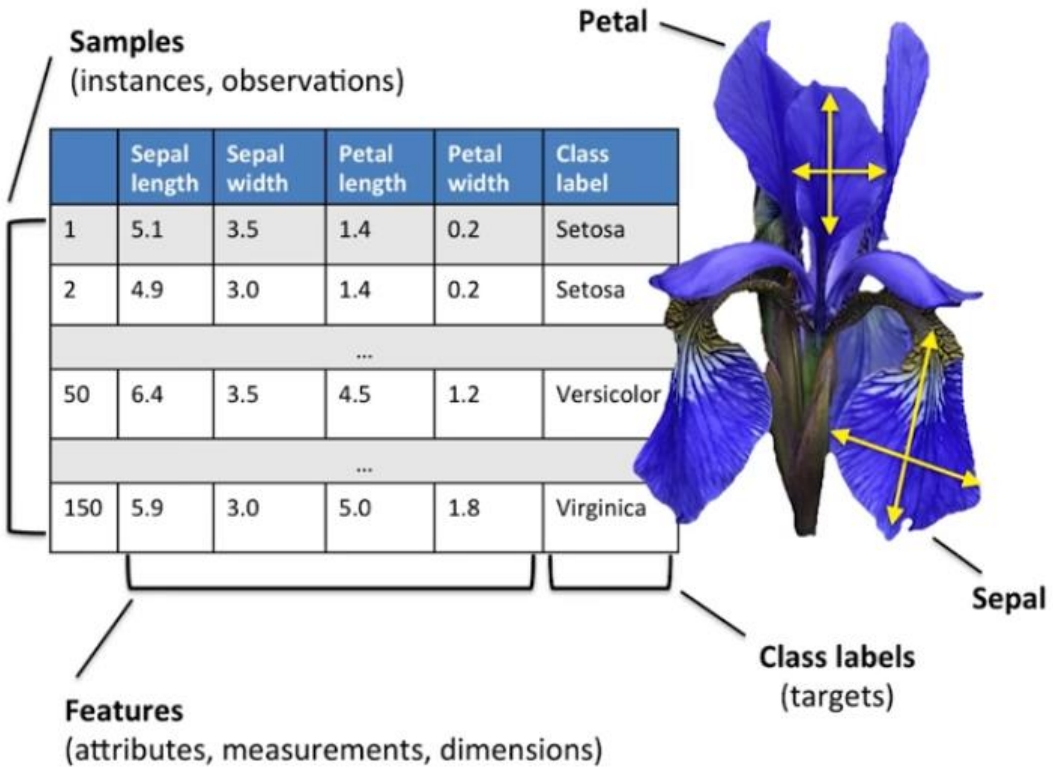
- Use wine dataset from sklearn
- Load the dataset and split it into test and train.
- Develop the model to Classify wines into different categories.
- Train the model using Gaussian and Multinominal classifier, please mention which model is better and why.
- Use the trained model to perform some predictions on test data.

Project 3:

- Datasets for Disease operation is provided. Dataset file name: Disease.csv
  1. Load the dataset and split it into test and train.
  2. Analyse the Data Visually.
  3. Find limitation with data if any in developing the model
  4. Develop the model to identify whether operation successful or not using Decision Tree algorithm
  5. Measure the performance of the model
  6. Try to improvise the model by suitable technique.
  7. Measure the performance of the model

Project: 4

- Use iris flower dataset from sklearn library and try to form clusters of flowers using petal width and length features.
- Analyse the Data Visually before training
- Drop other two features for simplicity.
- Try to Use K Mean Clustering
- Analyse the Data Visually by comparing before and after the training.



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Practice: Exercise 1

1. What is K-Nearest Neighbour(KNN) Algorithm for Machine Learning
2. How does K-NN work?
3. Implementation of the KNN algorithm- Take dataset: diabetes.csv
   - Data Pre-processing step-if any
   - Fitting the K-NN algorithm to the Training set
   - Predicting the test result
   - Test accuracy of the result