Requirements

- Machine Learning Understanding
- Python 3.5+
- Python libraries : Scikit Learn, Pandas, Numpy

1 Introduction

This is the beginner project 2 for the ML Course. The aim of the project is to predict predict the price of diamonds given the features of it. Essentially this is a **Regression** problem.

2 Description

"Diamonds are forever". Diamonds are one of the most expensive jewellery items. Formed by only one element, they are found in every color of the rainbow. The Diamonds dataset is a classic and a great dataset for beginners for understanding regression.

The dataset can be downloaded from https://www.kaggle.com/shivam2503/diamonds.

3 Data set Description

The attributes of the dataset are:

- price price in US dollars (\$326–\$18,823)
- carat weight of the diamond (0.2–5.01)
- cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color diamond colour, from J (worst) to D (best)
- clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- x length in mm (0–10.74)

- y width in mm (0–58.9)
- z depth in mm (0–31.8)
- depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43–79)
- table width of top of diamond relative to widest point (43–95)

Thus we can see that the **Price** is the target to be predicted. All the rest are features.It must be noted Index Counter in the CSV is just the index.

4 Problem Statement

Given the features and the target, build a **Machine Learning Regression model** that can **predict** the price of the diamond from a given set of features.

Note 1: Please divide the dataset into 70% for training, 20% for validation and keep the rest 10% for testing.

Note 2: Upon observation some of the data is of string type(Eg: Color). Since python or scikit learn works with numbers, it will be required to convert the string type of data into numbers. ***Clue: Check out Label Encoding using Scikit Learn***

You can use any regression algorithm. Please make sure you answer the following are in your report with the code:

- Briefly explain the algorithm.
- What is the Mean Sqaured Error and Mean Absolute Error obtained?
- Observe and note changes in accuracy as you **vary parameters**. Ex. Number of Nodes in Random Forest regressor.

Optional: An interesting extra task that can be done after you are able to predict is **compare the performance** on different algorithms. Also looking into **Exploratory data Analysis** will be useful.

5 Report

A simple report containing the following can be submitted:

- Name, SRN, Section
- Answers to the questions above
- Code : Direct python or jupyter notebook screenshots
- Results (Can be tables or Graphs or simple sentences)

6 Contact for Doubts:

- Dr. B N Krupa : bnkrupa@pes.edu
- K Venkat Ramnan (TA) : venkatramnank@pesu.pes.edu