THE UNIVERSITY OF HONG KONG

MSc in E-Commerce and Internet Computing

ECOM7126: Machine Learning for Business and E-Commerce
(2022-2023)

Assignment 2 – Diamond Classification

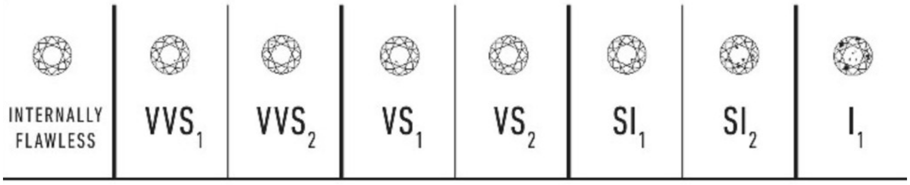
Diamond is one of the most valued gem stones in the world. The quality of a diamond is often graded according to what is well-known as the 4Cs: Carat, Cut, Colour and Clarity. There are also other objective and physical attributes, especially those that relates to the dimensions of the diamond. You are asked by your company to build a classification engine to classify any diamond into three classes: Low, Mid and High, based on data accumulated of grading by experts over the years. The ML classification engine will be used by newly trained gem expert to assist them in their job.

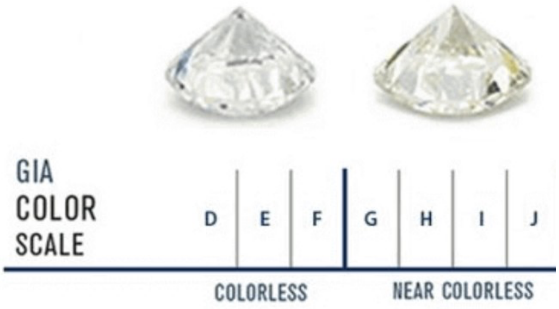
The dataset consists of 10,000 diamonds over various characteristics, in .csv format and the following data dictionary:

carat	weight of the diamond (1 carat = 0.2 g)
cut	quality of the cut rated by experts (Fair, Good, Very Good, Premium, Ideal)
colour	rated from D (best) to J (worst) for our dataset
clarity	a measure of how clear the diamond is as defined by GIA Clarity Chart ((I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best) – see later
length, width	the widest orthogonal dimensions looking from the top (also called x, y in mm. for round diamond, it is also called diameter and x = y)
height	the height (also called depth z in mm) of the diamond
depth ratio	depth percentage = height / mean (length, width) = 2 * z / (x + y) in %
table	percentage of width of the top facet of the diamond to the width

(Note: GIA = Gemological Institute of America, an institute that define many standards for assessing diamond)

To learn more about the above parameters of a diamond, you are encouraged to visit the following website: https://www.diamonds.pro/education. The following information may also be helpful:





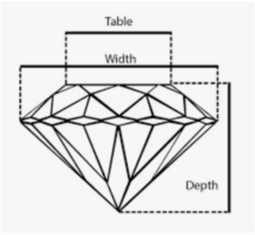


Table % = Table / Diameter
Depth % = Depth /Diameter

Deliverable:

1. Design and implement a ML model to classify any new diamond in the three categories (Low, Mid and High grade).
2. A report in PDF format.
3. The Colab notebook (Python programs with comments and notes) that you use to produce your results in `.ipynb` (Colab notebook) format.

You should include the following in your report:

1. An account of what you have done in investigating and transforming your data and why.
2. Use **Classification** models you have learned so far to build a classification engine as specified.
3. Any observation you may have in your project to the management of the company.

Datasets provided: DiamondDataset.csv