Nitte Meenakshi Institute of Technology



**Title: Diamond Classification**

Submitted in partial fulfilment of LA2

**Bachelor of Engineering**

**In**

**Computer Science and Engineering**

**Submitted by:**

**Ayushman Shivam Veeresh**

1NT19CS223 1NT19CS213

**Uday Kiran Chari Vaibhav Jamwal**

1NT19CS204 1NT19CS209

**Under the Guidance of**:

**Dr. Vani V**

**Department of Computer Science Engineering**

**Introduction**

Natural diamonds are one of the precious stones bought to wear as jewellery or as investment as well. Diamonds are not that glittery and beautiful in their raw form. The rough diamond stone is normal looking stone as others are. The miners filter the mined soil to find the rough diamonds and sell them to the manufactures. The manufactures do the creative work on those rough stones. There are many shapes of polished diamonds available in the market. Manufactures plan and polish the rough diamonds based on the maximum financial gain from the polished product. Here, the diamond's price depends upon hundreds of parameters but mainly on 4 C's (Carat, Cut, Clarity, Colour).

With this project we aim to classify diamonds into various categories using image data on the basis of various factors to demystify the enigma of these stones and perform price prediction based on relevant factors.

**Data Mining Tasks**

The planned approach involves most of the standard data mining steps which include:

* **Data Understanding**: Taking a closer look at the dataset available, particularly understanding the attributes available and the quality of the data. Based on the understanding, planning and modifying the approach to be taken for reaching the end goal.
* **Data Preparation**: Involving multiple actions to convert the existing raw data into final data that can be used for the analysis, which includes cleaning the data, data reduction based on the requirement, and data transformation. The data is also normalized in this process.
* **Training the model**: Based on the identified training dataset and the method adopted, the model is trained.
* **Evaluating the model**: The model trained is then used to predict the values using the test dataset.

**Data Set**

The data set consists of 6 CSV files based on their shape (cushion, emerald, heart, oval, radiant, round) with the following attributes: -

**Id**: unique identification number of diamond

**Shape**: shape of the diamond

**Weight**: weight of the diamond in Carats (the bigger the weight the expensive it is)

**Clarity**: clarity of the diamonds (FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, SI3, I1, I2, I3)

**Colour**: colour shade of the diamond (D, E, F, … Z)

**Cut**:  cutting level of the polished stone (Poor, Fair, Good, very Good, Excellent)

**Polish**: polish level of the stone

**Symmetry**: over all symmetry of the stone's shape

**Fluorescence**: Fluorescence is the ability of certain chemicals to give off visible light after absorbing radiation which is not normally visible, such as ultraviolet light

**Measurement**: The measurement of the diamonds.

**Methods and Models**

**Normalization**: Rescaling real-valued numeric properties into a 0 to 1 range is referred to as normalization. In machine learning and data mining, data normalization is used to make model training less sensitive to feature scale. As a result, our model can converge to better weights, resulting in a more accurate model. When characteristics are normalized, they become more consistent with one another, allowing the model to predict outputs more correctly. Python has a library called preprocessing which has functions that can be used to normalize the data. Categorise the images into dictionaries according to their shape.

**Model Building**: The resultant output of the normalisation is first split into the test and training data using the train\_test\_split function. Then sequential analysis is done using the functions provided under keras library under tensorflow

**Bibliography**

* Kaggle
* Tutorials Point
* One Stop Data Analysis
* Ask Python