Diamond clarity prediction

Submitted By :

Y Aravind Reddy (VU22CSEN0500188)

M V S Nitheesh (VU22CSEN0100830)

ABSTRACT

Formal Description of the Problem:

The objective of this problem is to develop a machine learning model that predicts the clarity of diamonds based on various characteristics such as price, carat weight, cut quality, color, depth, and other relevant parameters. Clarity is a critical factor in determining a diamond's overall quality and value, significantly influencing consumer choices in the jewelry market.

Well-Posed Problem:

- **Task (T):**
  The task is to build a machine learning model that predicts the clarity grade of diamonds based on attributes such as price, carat weight, cut, color, depth, and other relevant features. The model will classify diamonds into different clarity categories, providing a reliable method for assessing the quality and value of a diamond based on its features.
- **Experience (E):**
  The model gains experience by training on a labeled dataset consisting of diamond samples with their corresponding features (price, carat weight, cut, color, depth, and clarity grades). As the model processes more examples, it learns patterns between these features and clarity grades, improving its ability to predict clarity for unseen diamonds.
- **Performance (P):**
  The performance of the model will be measured by how accurately it predicts clarity grades using:

  Accuracy: Percentage of correct predictions.

  Precision and Recall: Quality of the predictions.

  F1-Score: Balances precision and recall.

Problem Statement:

"Given a dataset of diamonds with various attributes (carat weight, price, cut, color, depth, etc.), design a machine learning model that accurately predicts the clarity grade of each diamond. The model's performance will be evaluated using accuracy and other relevant metrics to ensure reliable assessments of diamond quality."

INTRODUCTION

Motivation:

As a student, solving the diamond clarity prediction problem offers the opportunity to apply machine learning techniques to a real-world challenge in the jewelry industry. This project enhances skills in predictive modeling and classification tasks. Additionally, accurate predictions of diamond clarity contribute to fair pricing in the market, helping consumers and jewelers make informed decisions.

Benefits of Solution:

Enhanced Diamond Quality Assessment: Provides quick, reliable predictions of diamond clarity based on features, improving transparency in the market.

Market Impact: Streamlines the grading process, leading to consistent pricing and better consumer trust.

Solution Use:

The solution can be integrated into an application where jewelers or consumers input diamond attributes to get clarity predictions. It can be used in both retail and online settings.

Operationalization:

The model is expected to have a long-term operational lifetime, requiring periodic retraining and data updates. It will serve as a foundation for future advancements in diamond evaluation technology.

Maintenance Considerations:

Regular updates to include new diamond data.

1. Continuous monitoring of model accuracy and retraining.
2. Ensuring software compatibility with evolving technology.
3. Adapting to new machine learning techniques.

Functional and Non-Functional Requirements:

Functional:
The solution must accurately predict diamond clarity and process real-time or batch inputs.

Non-Functional:
Scalability, reliability, and maintainability are critical to handle large datasets and ensure operational longevity.

ML ALGORITHMS

To address the task of predicting diamond clarity based on features like carat weight, price, cut, color, and depth, the following machine learning algorithms are suitable:

1. **Logistic Regression:**

   A linear model for binary or multi-class classification, serving as a baseline.

   Suitability: Works well for linear relationships between features and clarity grades.

2. **Multinomial Logistic Regression:**

   An extension of logistic regression for multi-class classification problems.

   Suitability: Useful for predicting multiple clarity grades.

3. **Support Vector Machines (SVM):**

   A robust algorithm that finds the optimal hyperplane separating different clarity classes.

   Suitability: Effective for high-dimensional data but computationally intensive for large datasets.

Performance Metrics:

The performance of these algorithms will be evaluated using:

1. **Accuracy:**
   Measures the overall correctness of the model.
2. **Precision:**
   Assesses how many of the predicted positives are actually positive.
3. **F1-Score:**
   Balances precision and recall, ideal for imbalanced data.

DATASET FINALIZATION

Here are the three datasets chosen for the project:

1. Kaggle Dataset
2. PyCaret Dataset
3. Diamond ML Dataset

Dataset Overview:

These datasets contain a comprehensive collection of diamond attributes such as carat weight, cut, color, clarity, depth, and price. The data is crucial for predicting diamond clarity and has been used in various applications for assessing diamond quality.

Features:

a. **Carat Weight:**
Measures the weight of the diamond.
Importance: Heavier diamonds are more valuable, but clarity also influences quality.

b. **Cut:**
Describes the diamond's polish and shape quality.
Importance: Enhances clarity perception.

c. **Color:**
Measures the presence of color in diamonds.
Importance: Colorless diamonds are more desirable, impacting clarity.

d. **Clarity:**
Grades range from Flawless (FL) to Included (I).
Importance: Determines a diamond's quality and value.

e. **Depth Percentage:**
Ratio of the diamond's depth to its average diameter.
Importance: Affects light performance and brilliance.

f. **Table Percentage:**
Ratio of the width of the diamond's table to its diameter.
Importance: Influences light entry and clarity.

g. **Price:**
Market price of the diamond.
Importance: Outcome variable influenced by features.

h. **Polish:**
Describes surface finish quality.
Importance: Affects light reflection and perceived clarity.

i. **Symmetry:**
Measures facet alignment.
Importance: Improves light performance, impacting clarity.

j. **Report:**
Certification from a gemological lab.
Importance: Provides an authoritative clarity grade.

These datasets have previously been utilized for predicting diamond prices, and we aim to enhance their application in our analysis.