ML For Diamond Cut Classification

LAB-P3 - 2
GABRIEL LEE JUN RONG, 2301906
QUAH YONG YAO, 2302009
SUDIPTA KANTI BISWAS, 2303435

ML in Gemology & Diamond Classification

Machine learning has been applied in gemology for **tasks such as origin, gemstone determination and grading** tasks.

- **GIA** found that **ML algorithms can complement traditional spectral analysis**, achieving classification error rates **as low as ~5%**.

ML in Gemology & Diamond Classification

AI-based systems are emerging for the “**4 Cs**” grading:

- Cut, Carat, Color, Clarity

Automated ML graders have been trained on tens of thousands of diamonds:

- Sarine Clarity

Traditionally done by human graders who measures the proportions and visually assessing light performance.

The 4 Cs of Grading:

- Clarity, Color, Cut, Carat

Problem Statement:

“Determine a given diamond’s **cut quality grade** from its attributes using ML.”

*We aim to **bring objectivity, speed and consistency** to this process with the use of ML.*

Dataset Overview

Sample Size: 53,940

Carat	Cut	Color	Depth	Table	Price	x	y	z	Clarity
0.23	Ideal	E	61.5	55	326	3.95	3.98	2.43	SI2

“Cut” Distribution:
Ideal (40%) Fair (3%) Premium (26%) Very Good (22%) Good (9%)

This imbalance suggests the model must be careful not to simply favor the majority class

Patterns in the data:
- **Fair-cut diamonds tend to be larger on average** (mean ~1.04 carats) whereas Ideal cuts are smaller (mean ~0.70 ct) - indicates a trade-off between retaining weight versus achieving top cut quality.

Challenges:
- Class imbalance and multicollinearity (e.g. the high correlation between x, y, z dimensions).

Model Training

5 Different Algorithms were trained.

Logistic Regression	KNN	Random Forest	XGBoost	SKLearn Gradient Boost

Approach:

- Each model was trained on an **80% training split** (with stratification) and evaluated on the **20% holdout test** set.
- Employed **5-fold cross-validation** on the training data for **hyperparameter tuning** and model selection.
- **Early stopping** was used to **prevent overfitting** in boosting iterations.

Experimentation Tweaking:

- Number of neighbours in KNN
- Regularization strength in Logistic Regression

Extra Experimental Models:

- Hyperparameter Fine-Tuned XGBoost
- Weighted XGBoost Approach
    1. Identify which classes are often misclassified
    2. Upweighting those classes by a factor of 1.1x
- Model Ensembling (Stacking & Voting)

Model Performance & Comparison

Worst - Best (Accuracy)	KNN	Logistic Regression	Random Forest	SKLearn Gradient Boost	XGBoost	Fine-Tuned XGBoost	Voting (WXGB + GB)	Weighted XGBoost	Stacking (WXGB + GB)
Accuracy	64.4%	65.3%	73.6%	80.1%	80.7%	81.2%	81.2%	81.4%	81.5%
Cohen's Kappa	0.659	0.546	0.787	0.826	0.832	0.834	0.830	0.844	0.833
F1	0.556	0.551	0.708	0.794	0.800	0.805	0.811	0.810	0.800

Given class imbalance, we computed Cohen's Kappa

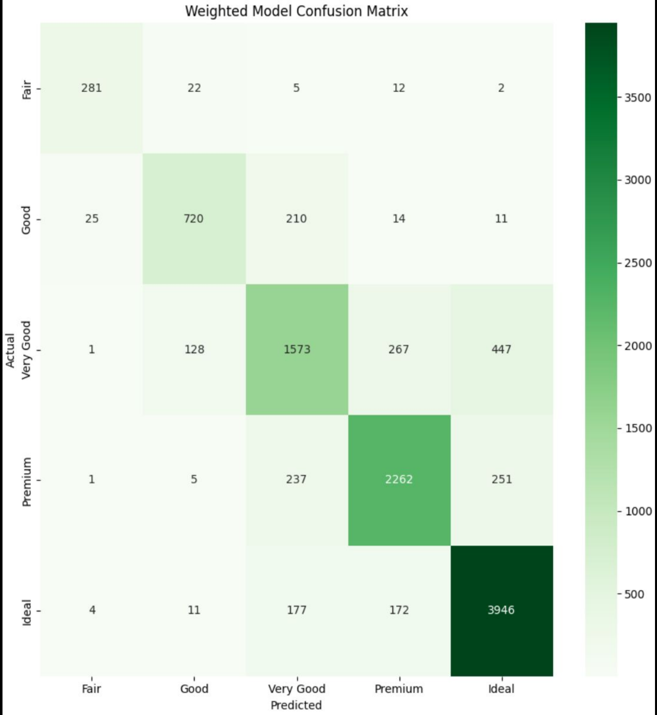
- Which measures agreement with the true labels adjusted for chance.

Confusion Matrix (For XGBoost):

- **Ideal and Premium cuts** had the highest individual precision and recall.
- Most misclassifications occurred between adjacent grades.

Misclassification Rates:

- The Weighted XGBoost **misclassifies 7.96% of the time**.



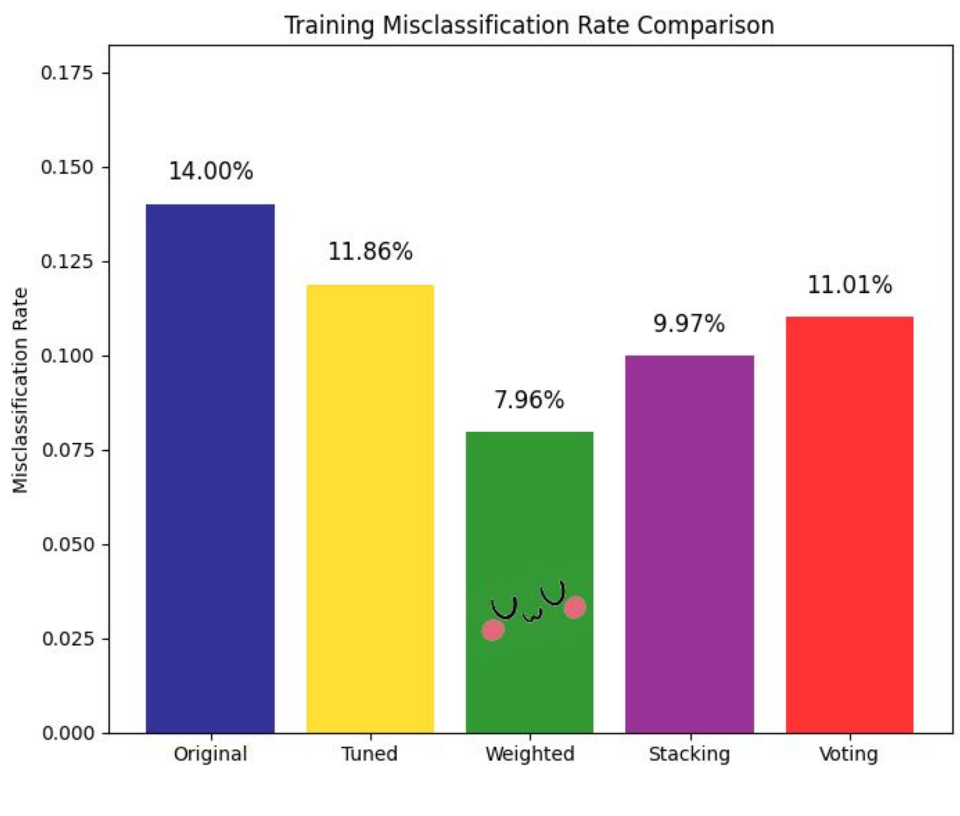
Final Model Selection

	XGBoost	Hyperparameter Fine-Tuned XGBoost	Weighted XGBoost	Stacking (Weighted XGBoost + GB)	Voting (Weighted XGBoost + GB)
Accuracy	80.7%	81.2%	**81.4%**	81.5%	81.2%
Misclassification Rate	14.00%	11.86%	**7.96%**	9.97%	11.01%

Chosen because it's **not worth sacrificing ~2% Misclassification rate for a 0.1% improvement** than the Stacking model.

The choice was not based solely on accuracy:

- XGBoost **provides rich feature importance feedback** and worked very well with SHAP analysis, allowing us to verify that its predictions were driven by logical factors.
- Training and Evaluation is faster than other models.



References

https://www.gia.edu/doc/fall-2024-machine-learning.pdf

https://nationaljeweler.com/articles/11975-state-of-the-diamond-industry-ai-and-the-future-of-diamond-grading

https://pmc.ncbi.nlm.nih.gov/articles/PMC10570374/

https://ieeexplore.ieee.org/document/10817052

https://www.kaggle.com/datasets/shivam2503/diamonds

Video Presentation Link

https://youtu.be/6LVVXqL1Le4