Clustering Diamonds for Quality Assessment

Course: Advanced Machine Learning
Lecturer: Dr. Hen Hagag
Team member's:Niv Levi & Sarah Hazziza

TABLE OF CONTENTS

PROBLEM OVERVIEW

While a gemological certificate for a diamond examines more than 20 different parameters, when a buyer wants to purchase a diamond and searches for information about it online, they primarily base their decision on the 4 main parameters, the 4Cs (Carat, Color, Cut, and Clarity).
Diamond sellers capitalize on this knowledge gap and omit certain information to the customers leading to inflated prices and limited choices for consumers.

Furthermore, parameters such as color, clarity, and fluorescence are determined by nature, while parameters like cut, symmetry, and polish are determined by the quality of the processing of that particular diamond. It is unclear whether diamonds with good natural parameters will be more closely associated with better artificial processing.

PREVIOUS METHODS

PRICE PREDICTION

Multiple methods were previously used on diamonds dataset, usually involving regression models for price prediction.
Those techniques don’t evaluate characteristics and clusters influencing the price and a diamond value.

EMPIRICAL

Diamond sellers use their knowledge and experience to assess quality of diamonds. They examine the 4Cs and employ gemological tools like microscopes and loops to assess the value of a diamond.

MOTIVATIONS & GOALS

Our goal is to improve the existing knowledge in the field and to know whether the industry's focus on the 4Cs is correct, and if so, what is the extent of the influence of the 4Cs in the division into clusters.

After dividing into clusters, we would like to examine the distribution of the clusters with the other features and check whether certain clusters are consistently associated with the same grades of parameters.

Furthermore, we seek to provide deeper understanding and more informations to stakeholders.

DATASET

	Cut	Polish	Symmetry	Depth%	Table%	Crownangle	Pavilion_Angle	Eyeclean	Certification	Price	Color	Clarity	Carat	Natural	Pointed Culet	Fluorescence	Thinest_Girdle	Thickest_Girdle	L/W%
0	Ideal	Excellent	Excellent	60.9	56.0	35.0	40.8	Yes	GIA	20795	J	VS1	2.13	1	0	None	thin	medium	0.995221
1	Ideal	Excellent	Excellent	61.6	56.0	35.0	40.8	Yes	GIA	19750	J	VS1	2.03	1	0	Faint	medium	medium	0.992656
2	Ideal	Excellent	Excellent	60.8	57.0	34.5	40.8	Yes	GIA	11975	J	VS2	1.64	1	0	None	medium	medium	0.997382
3	Ideal	Excellent	Excellent	61.9	57.0	35.0	40.8	Yes	GIA	51750	I	VS2	3.32	1	0	None	medium	slightly thick	1.002099
4	Ideal	Excellent	Excellent	61.5	56.0	34.5	40.8	Yes	GIA	9388	E	VS1	1.07	1	0	Faint	medium	medium	0.996974
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13498	Ideal	Excellent	Excellent	60.9	59.0	33.6	40.8	Yes	IGI	20297	G	VVS2	13.14	0	0	None	medium	slightly thick	1.004630
13499	Ideal	Excellent	Excellent	60.2	59.0	33.0	40.8	Yes	IGI	16161	G	VS1	13.15	0	0	None	medium	slightly thick	1.004593
13500	Ideal	Excellent	Excellent	61.4	58.0	34.8	40.9	Inquire	IGI	27798	F	VS2	13.54	0	1	None	medium	medium	0.996091
13501	Ideal	Excellent	Excellent	61.2	58.0	34.4	40.9	Yes	IGI	39815	G	VS1	15.34	0	1	None	medium	medium	0.994400
13502	Ideal	Excellent	Excellent	61.1	58.0	34.5	40.8	Yes	IGI	40282	G	VS1	15.52	0	1	None	medium	medium	0.996273

13503 rows × 20 columns

FEATURES

SOME OF PRINCIPAL FEATURES:

- CARAT
- COLOR
- CLARITY
- CUT
- POLISH
- SYMMETRY
- FLUORESCENCE
- MEASUREMENTS
- PROPORTIONS (CROWN & PAVILION ANGLES)

GIA®

GIA REPORT
1443371178

Verify this report at GIA.edu

FACSIMILE

This is a digital representation of the original GIA Report. This representation might not be accepted in lieu of the original GIA Report in certain circumstances. The original GIA Report includes certain security features which are not reproducible on this facsimile.

GIA NATURAL DIAMOND GRADING REPORT

December 07, 2022
GIA Report Number 1443371178
Shape and Cutting Style ..... Round Brilliant
Measurements ..... 7.62 - 7.64 x 4.64 mm

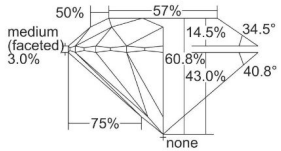
GRADING RESULTS

Carat Weight ..... 1.64 carat
Color Grade ..... J
Clarity Grade VS2
Cut Grade ..... Excellent

ADDITIONAL GRADING INFORMATION

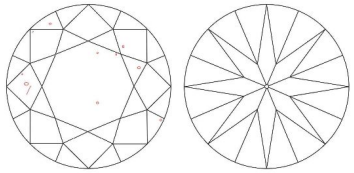
Polish ..... Excellent
Symmetry ..... Excellent
Fluorescence ..... None
Inscription(s): GIA 1443371178
Comments: Pinpoints, internal graining and surface graining are not shown.

PROPORTIONS



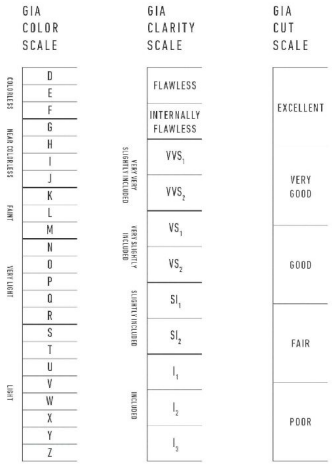
Profile to actual proportions

CLARITY CHARACTERISTICS



KEY TO SYMBOLS*
○ Crystal
Feather
Needle

GRADING SCALES



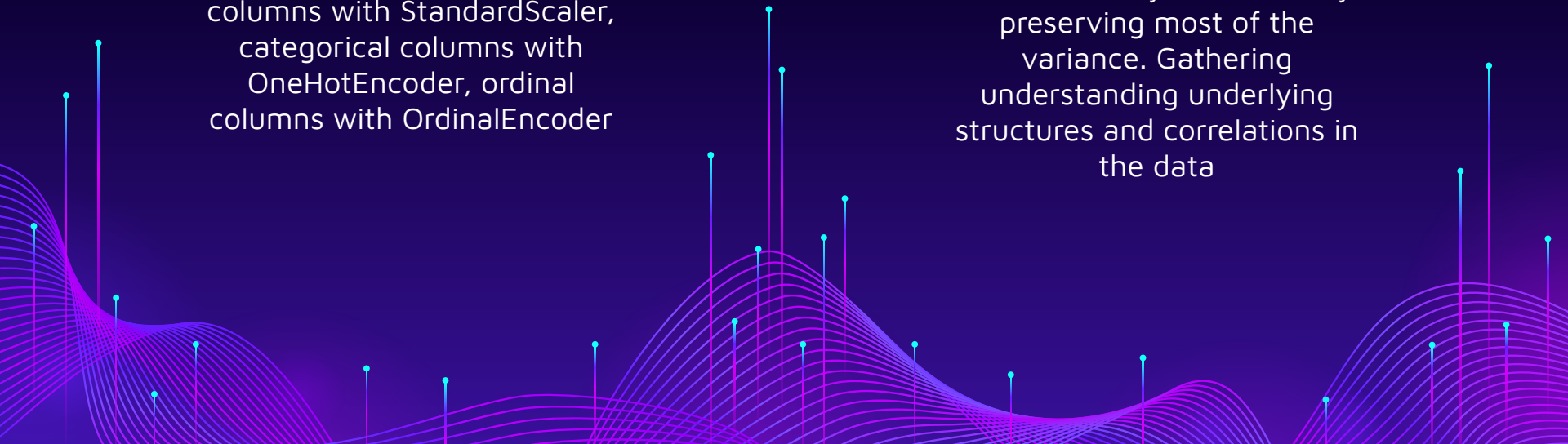
METHODOLOGY: TECHNIQUES

PIPELINE

Transforming numerical columns with StandardScaler, categorical columns with OneHotEncoder, ordinal columns with OrdinalEncoder

PCA

Dimensionality Reduction by preserving most of the variance. Gathering understanding underlying structures and correlations in the data



METHODOLOGY: EVALUATING DIFFERENT ALGORITHMS

KMEANS

Dataset well separated across columns could potentially be separated into roughly equal size clusters

AGGLOMERATIVE

Could be some kind of hierarchy between clusters for diamonds with similar characteristics

DBSCAN

Diamonds could form dense clusters based on particular features like carat or price

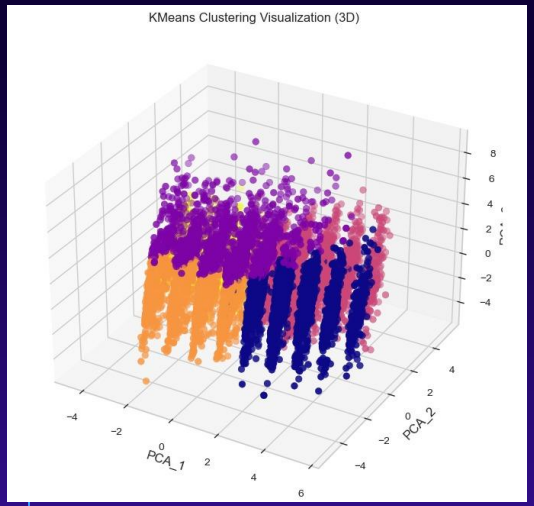
EXPERIMENTS AND EVALUATION

We experimented with different clustering algorithms, used functions and techniques such as Elbow method to find the best parameters of each algorithm.

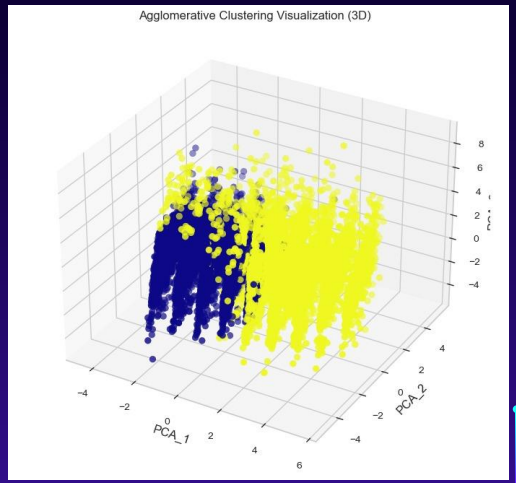
Thanks to internal evaluation like silhouette scores, within-cluster sum of squared error and visualization of the clusters, we understood that the best algorithm is K Means clustering.

COMPARISON OF ALGORITHMS

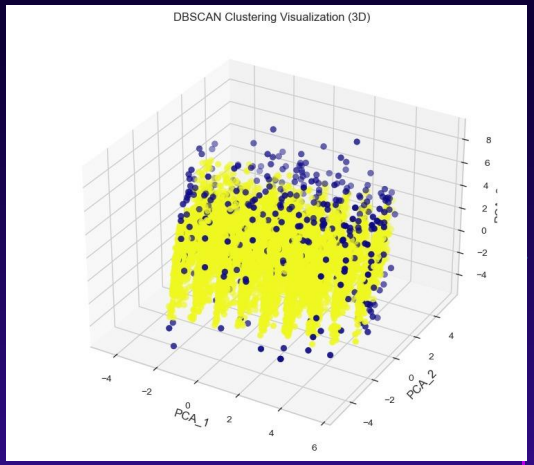
KMEANS



AGGLOMERATIVE



DBSCAN



RESULTS

- Cluster 4 predominantly comprised Lab diamonds with superior color, clarity, and cut, indicating high quality, likely due to controlled selection of color and clarity features when creating Lab diamonds.
- Cluster 3 exhibited good color but lower clarity, often featuring very good to excellent cut and less consistent symmetry.
- Cluster 2 showed exceptional clarity but poorer color quality, with predominantly excellent cut.
- Cluster 0 contained diamonds with good to excellent cut but lower color quality and clarity, likely enhanced through meticulous cutting.
- Cluster 1 displayed diverse characteristics across the 4Cs, presenting challenges in defining its distinct features, yet showing a trend of lower clarity and better colors alongside various cuts, polish, and symmetry.

CONCLUSION

- KMeans clustering effectively categorized our dataset and clustered diamonds based on attributes like cut, color, and clarity. The algorithm helped us assess different quality of diamonds.

- Findings offer actionable insights for industry stakeholders and empower consumers with comprehensive diamond information.

- Future research may explore grading systems, computer vision applications on microscopes pictures of the diamonds, and alternative diamond shapes to advance analysis and market understanding.