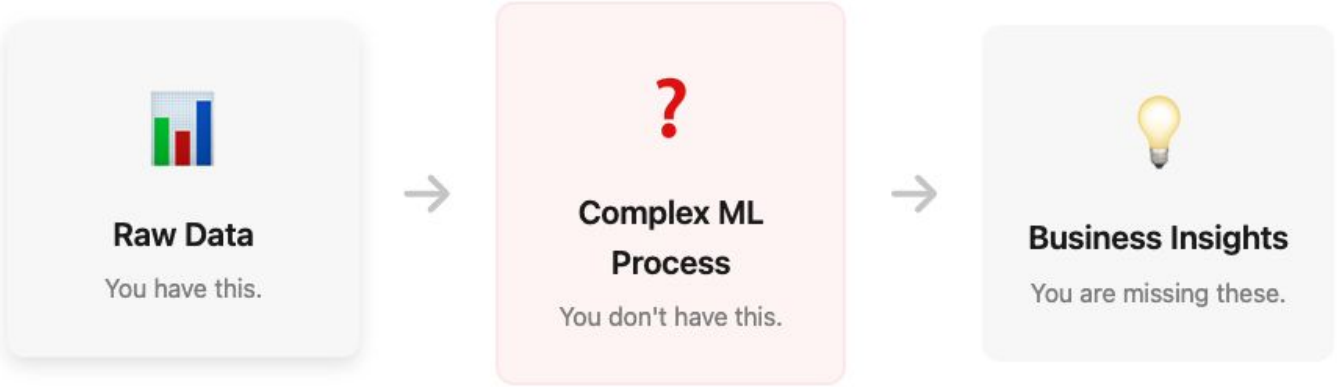
Aaron Panych
RoboGarden
Capstone Presentation

Universal ML Analysis Suite

From Data to Insights in Minutes

The Problem



Universal ML Analysis Suite

From Data to Insights in Minutes

Our Solution

CSV Upload

30 seconds

One-Click Analysis

5-10 minutes

Professional Results

Enterprise Quality

Two-Step Workflow

CSV Upload

Any dataset

EDA First

Understand data

ML Analysis Next

15 algorithms

Professional Results

Ready to present

Suite Components

Exploratory Data Analysis (EDA)

Automatically generates correlation matrices, histograms, and pair plots to reveal what your data contains before any modeling. Shows which features matter most and uncovers hidden relationships.

- Diamonds: Size drives price more than cut quality
- Sales: Which factors predict customer spending
- Medical: Patient data patterns and outliers

Classification

Predicts categories or classes (like quality grades, customer types, or risk levels) by testing 6 different algorithms. Tells you which approach works best for your specific data.

- Diamonds: Predict cut quality (Fair to Ideal)
- Customers: Segment as high/medium/low value
- Loans: Approve/review/reject applications

Regression

Predicts numerical values (like prices, sales forecasts, or scores) using 6 different models. Automatically finds the most accurate approach and shows you the prediction error.

- Diamonds: Predict price within ±$542 accuracy
- Real Estate: Estimate property values
- Sales: Forecast monthly revenue

Clustering

Discovers natural groups in your data without being told what to look for (like customer segments or product categories). Shows how many groups exist and what makes them different.

- Diamonds: Two market segments (43.5% vs 56.5%)
- Customers: Natural buying behavior groups
- Products: Identify similar item categories

Diamond EDA Process

Diamond Dataset

53,940 samples

9 features: carat, cut, color,
clarity, depth, table, price,
x, y, z

Automated Analysis

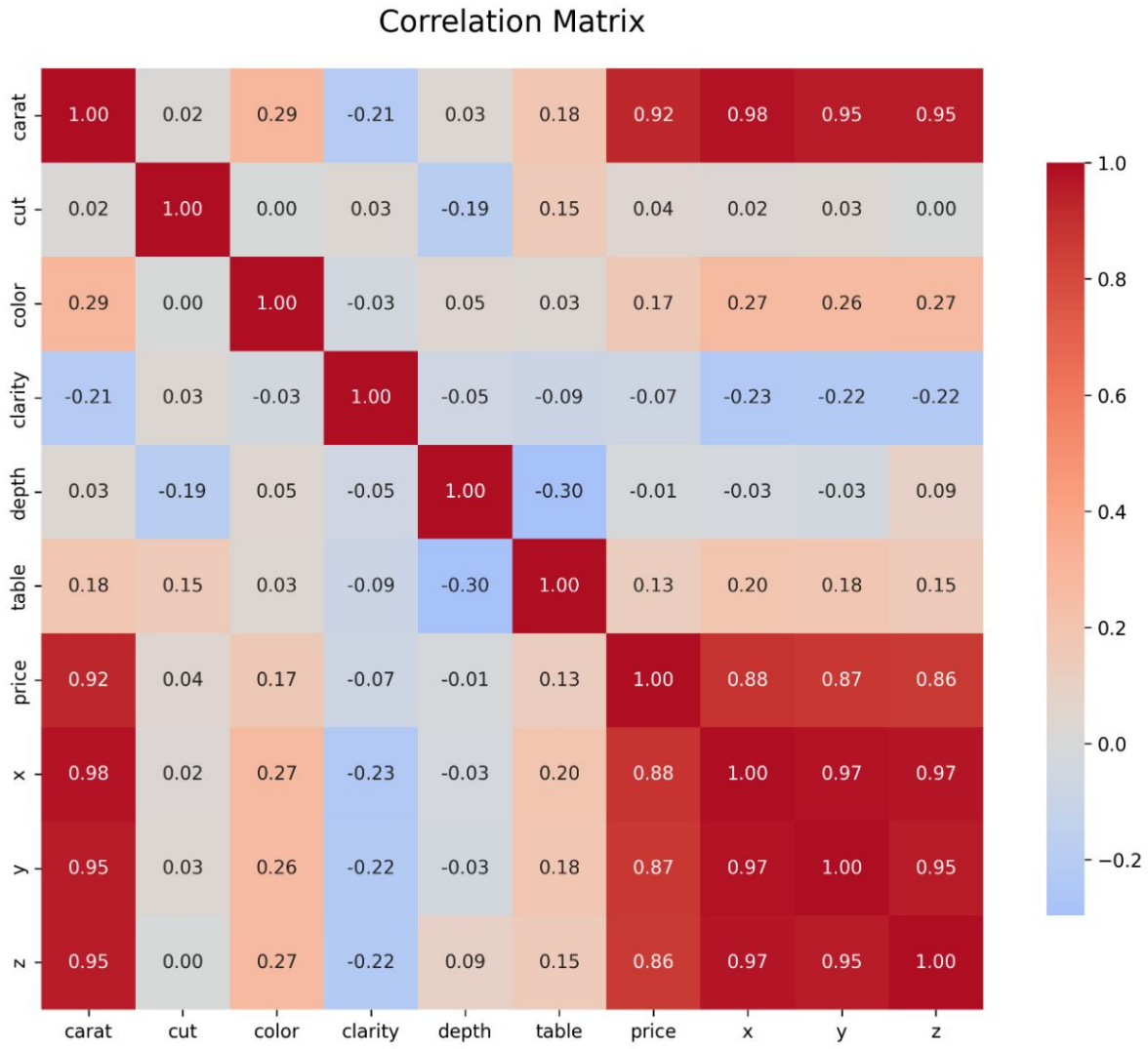
5 minutes

Correlation analysis
Distribution mapping
Pattern detection

Key Discovery

Size = Price

0.92 correlation reveals
carat weight drives pricing


Correlation Matrix
carat 1.00 0.02 0.29 -0.21 0.03 0.18 0.92 0.98 0.95 0.95
cut 0.02 1.00 0.00 0.03 -0.19 0.15 0.04 0.02 0.03 0.00
color 0.29 0.00 1.00 -0.03 0.05 0.03 0.17 0.27 0.26 0.27
clarity -0.21 0.03 -0.03 1.00 -0.05 -0.09 -0.07 -0.23 -0.22 -0.22
depth 0.03 -0.19 0.05 -0.05 1.00 -0.30 -0.01 -0.03 -0.03 0.09
table 0.18 0.15 0.03 -0.09 -0.30 1.00 0.13 0.20 0.18 0.15
price 0.92 0.04 0.17 -0.07 -0.01 0.13 1.00 0.88 0.87 0.86
x 0.98 0.02 0.27 -0.23 -0.03 0.20 0.88 1.00 0.97 0.97
y 0.95 0.03 0.26 -0.22 -0.03 0.18 0.87 0.97 1.00 0.95
z 0.95 0.00 0.27 -0.22 0.09 0.15 0.86 0.97 0.95 1.00
carat cut color clarity depth table price x y z
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

Diamond Cut Classification

Predicting Quality Grades Automatically

Classification Challenge

Diamond Features

Physical Properties

Carat, color, clarity, depth, table, price, dimensions

6 ML Algorithms

Automated Testing

Logistic, KNN, Tree, Forest, SVC, Neural Network

Cut Quality

5 Classes Predicted

Fair, Good, Very Good, Premium, Ideal

Diamond Cut Classification

Predicting Quality Grades Automatically

Algorithm Performance Results

Neural Network
79.70%
Best overall accuracy

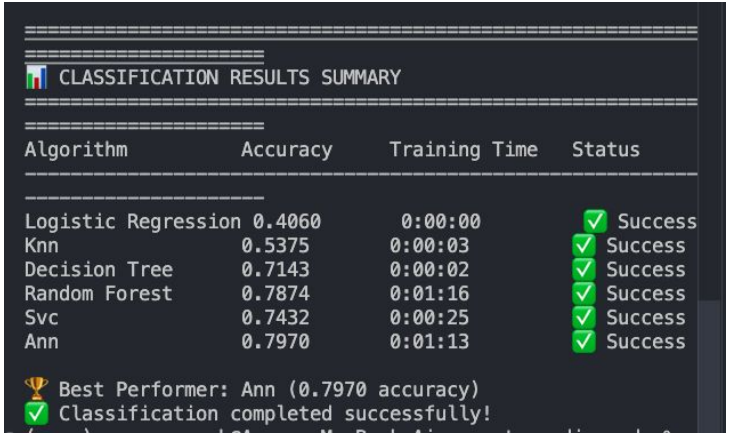
Random Forest
78.74%
Close second, stable

SVC
74.32%
Good performance

Decision Tree
71.43%
Fast training

KNN
53.75%
Moderate results

Logistic Regression
40.60%
Poor fit



```
📊 CLASSIFICATION RESULTS SUMMARY

Algorithm           Accuracy     Training Time   Status
Logistic Regression 0.4060       0:00:00         ✅ Success
Knn                 0.5375       0:00:03         ✅ Success
Decision Tree       0.7143       0:00:02         ✅ Success
Random Forest       0.7874       0:01:16         ✅ Success
Svc                 0.7432       0:00:25         ✅ Success
Ann                 0.7970       0:01:13         ✅ Success

🏆 Best Performer: Ann (0.7970 accuracy)
✅ Classification completed successfully!
```

Diamond Price Prediction

98.15% Accuracy • ±$542 Average Error

Regression Challenge

Diamond Features

Physical & Quality

Carat, cut, color, clarity, depth, table, dimensions

6 ML Models

Automated Testing

Linear, KNN, Tree, Forest, SVR, Neural Network

Price Prediction

$326 - $18,823

Mean: $3,933
53,940 diamonds

Diamond Price Prediction

98.15% Accuracy • ±$542 Average Error

Model Performance Results

Random Forest
98.15% R²
±$542 RMSE
Exceptional accuracy

Neural Network
96.85% R²
±$708 RMSE
Strong performance

Decision Tree
96.66% R²
±$729 RMSE
Fast & interpretable

KNN
94.98% R²
±$894 RMSE
Good local patterns

Linear Regression
88.51% R²
±$1,351 RMSE
Simple baseline

SVR
51.45% R²
±$2,778 RMSE
Poor fit

```
================================================================================
📊 REGRESSION RESULTS SUMMARY
================================================================================
Algorithm          R² Score   RMSE         Training Time   Status
--------------------------------------------------------------------------------
Linear Regression  0.8851     1351.26      0:00:00
✅ Success
Knn                0.9498     893.76       0:00:17
✅ Success
Decision Tree      0.9666     729.03       0:00:40
✅ Success
Random Forest      0.9815     542.02       0:00:16
✅ Success
Svr                0.5145     2778.03      0:00:48
✅ Success
Ann                0.9685     707.71       0:00:39
✅ Success

🏆 Best Performer: Random Forest (R² = 0.9815, RMSE = 542.02)
✅ Regression completed successfully!
```

Diamond Market Segmentation

Discovering Natural Customer Groups

Clustering Challenge

💎

Diamond Features

8 Dimensions

Carat, color, clarity, depth, table, dimensions (x, y, z)

→

🔍

3 Algorithms

Pattern Discovery

K-Means, Agglomerative, Mean Shift clustering

→

🎯

Market Segments

Natural Groups

Two-tier market structure 43.5% vs 56.5% split

Diamond Market Segmentation

Discovering Natural Customer Groups

Clustering Results & Market Structure

K-Means 0.3223

2 clusters • 100% coverage

Best silhouette score

Mean Shift -0.0596

12 clusters • 27.8% coverage

Over-segmentation issue

Agglomerative -0.1213

2 clusters • 18.5% coverage

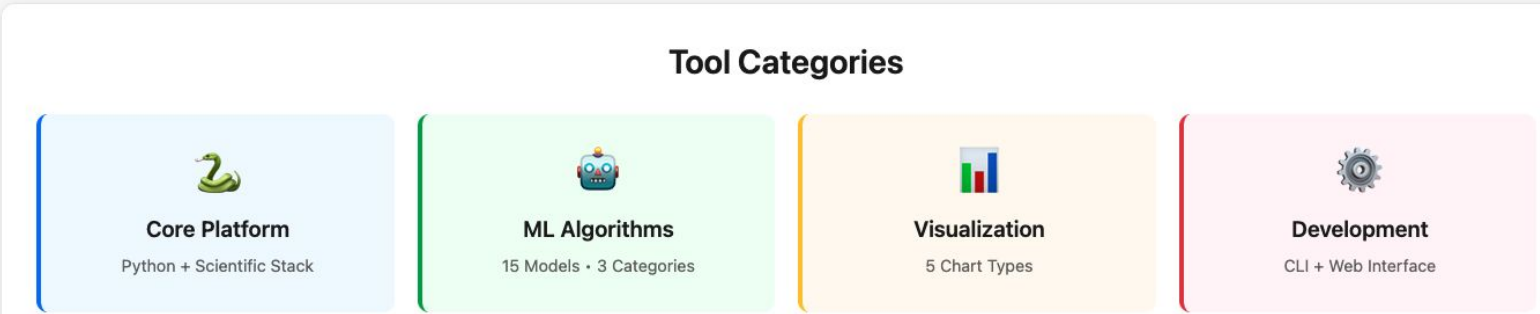
Limited by sampling

Two-Tier Market Structure

Segment 1	23,485 (43.5%)
Segment 2	30,455 (56.5%)
Total Market	53,940 (100%)

```
============================
CLUSTERING RESULTS SUMMARY
================================================================
============================
Algorithm       Clusters   Silhouette   Optimal
Param   Data Coverage Notes
----------------------------------------------------------------
----------------------------
Kmeans ⭐        2          0.3171       K=2
     100.0%       Full dataset
Agglomerative   2          -0.1183      N=2
    18.5%        Sampled data
Meanshift       12         -0.0596      BW=2.987
    27.8%        Sampled data
----------------------------------------------------------------
----------------------------
⭐ Best performing algorithm based on silhouette
score: Kmeans
```

Technology Stack



Python 3.12
pandas
numpy
scikit-learn

Logistic Regression
Linear Regression
K-Nearest Neighbors
Decision Tree
Random Forest
SVC
SVR
ANN

matplotlib
seaborn
Correlation Matrix
Pair Plots
Histograms
Box Plots

CLI interface
argparse
*Flask interface

Major Development Problems & Solutions

Repetitive Data Analysis

Same plots generated for every ML task. Output was 70% redundant EDA, 30% ML results.

Critical UX Issue

SOLVED

Two-Path Architecture

Created dedicated EDA script + clean ML outputs. Users choose exploration vs analysis.

Clean & Focused

Memory Limitations

53,940 samples crashed clustering algorithms. Agglomerative and Mean Shift became unusable.

Scalability Blocker

SOLVED

Smart Optimization

Intelligent sampling + garbage collection + progressive processing. Handles 53k+ samples efficiently.

Enterprise Scale

Marketing Strategies

🏢 SMB Market Focus

SMALL-MEDIUM BUSINESSES

Target the $2B SMB market with "We have data but no data scientists" messaging. Position as enterprise analytics at startup prices.

LinkedIn Ads | Industry Webinars | Free Trials | Case Studies

Volume-based pricing

💼 Consulting Partnerships

CONSULTING FIRMS

"Deliver client insights in hours, not weeks." Partner with analytics consultancies for white-label deployment and revenue sharing.

Partner Program | White Label | Revenue Share | Training

Custom pricing • Channel Growth

🌍 Vertical Expansion

INDUSTRY-SPECIFIC SOLUTIONS

Create templates for retail, healthcare, finance, real estate. "Diamond analysis" becomes "Customer analysis," "Property analysis," etc.

Industry Templates | Vertical Landing Pages | Trade Shows | Industry Partnerships

Custom Pricing • Premium Tiers