Who Is The Killer?

Piraeus Vice Pattern Recognition Project

Vasilaina Maria (Π23015)
Grigoraskou Teresa (Π22037)
Liveris Fotis (Π23104)

Department of Informatics
University of Piraeus

February 19, 2026

Outline

Project Overview

Objective

Identify the most likely killer for each crime incident in the “Piraeus Vice” dataset using machine learning techniques.

Supervised Methods:

- Gaussian Bayes Classifier
- Linear Classifier
- SVM (RBF kernel)
- Multi-Layer Perceptron

Unsupervised:

- PCA for visualization
- k-means clustering

Dataset Overview

- **Total incidents:** 4,800 crime cases
- **Target:** 8 killers (multiclass classification)
- **Split:** TRAIN / VAL / TEST (predefined)

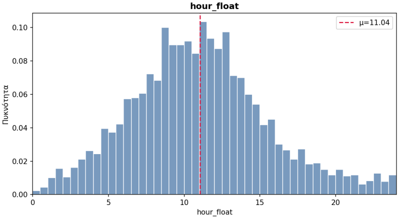
Continuous Features (8):

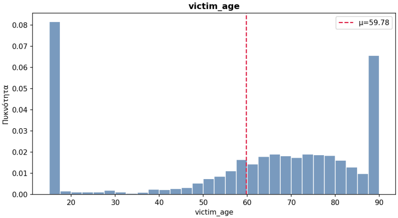
- hour_float
- latitude, longitude
- victim_age
- temp_c, humidity
- dist_precinct_km
- pop_density

Categorical Features (4):

- weapon_code
- scene_type
- weather
- vic_gender

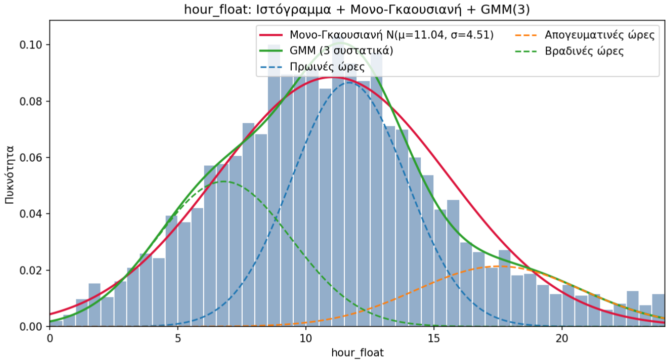
Q1: Key Variable Distributions





- **hour_float:** Peak incidents during daytime (unexpected)
- **victim_age:** Bimodal — very young or very old (vulnerable groups)

Q1: Gaussian Mixture vs Single Gaussian



Key Insight

Single Gaussian inadequate — GMM (3 components) captures multiple time-of-day patterns better.

Q2: Maximum Likelihood Estimation

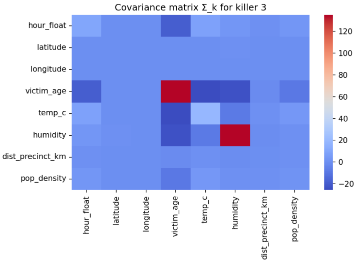
Approach

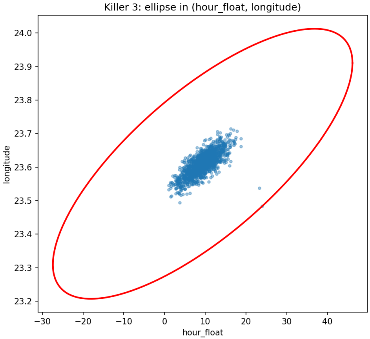
For each killer k, estimate mean μ_k and covariance Σ_k of continuous features using MLE on TRAIN split.

Assumption: Each killer's feature vector follows a multivariate Gaussian distribution:

$$p(\mathbf{x}|k) = \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$$

Q2: Covariance Heatmaps & Ellipses





- Each killer exhibits distinct spatial-temporal patterns
- 2D ellipses show hour_float vs longitude distributions

Q3: Multiclass Gaussian Bayes Classifier

Method

Use MLE parameters from Q2 to build a Bayesian classifier:

$$\hat{k} = \arg\max_{k} \; p(k|\mathbf{x}) \propto p(\mathbf{x}|k) \cdot p(k)$$

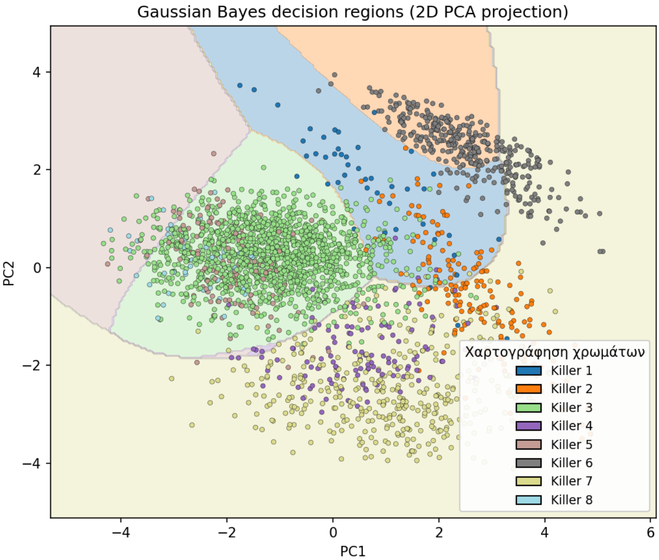
Results:

- TRAIN Acc: **89%**
- VAL Acc: **90%**

Observations:

- Confusion matrix mostly diagonal
- Killers 3, 6, 7 dominate

Q3: Decision Regions (PCA Projection)



Q4: Linear Classifier

Approach

Discriminative multiclass classification using all 12 features (8 continuous + 4 categorical via one-hot encoding).

Results:

- TRAIN Acc: **77%**
- VAL Acc: **78%**

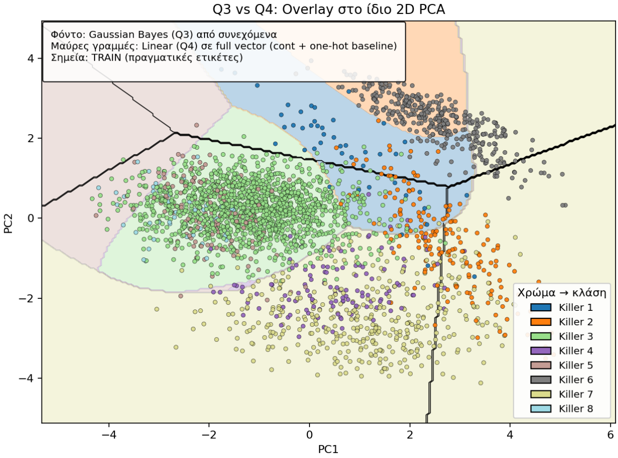
Analysis:

- Lower than Bayes (90%)
- Linear boundaries too restrictive

Limitation

Cannot capture nonlinear killer patterns in feature space.

Q4: Linear Decision Boundaries



Linear boundaries (straight lines) fail to separate complex killer profiles.

Q5: SVM with RBF Kernel

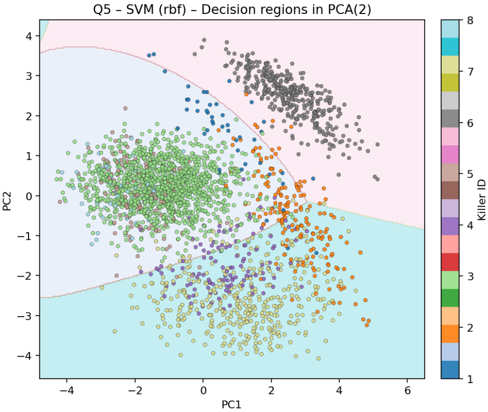
Configuration

- **Kernel:** RBF (allows nonlinear boundaries)
- **Strategy:** One-vs-Rest (8 binary classifiers)
- **Tuning:** Grid search over $C \in \{0.3, 1, 3\}$ and $\gamma \in \{\text{scale}, 0.1, 0.03\}$

Performance

VAL Accuracy: 94% — Best performance so far!

Q5: SVM Decision Regions & Support Vectors



Nonlinear boundaries effectively capture killer-specific patterns.

Q6: Neural Network (MLP)

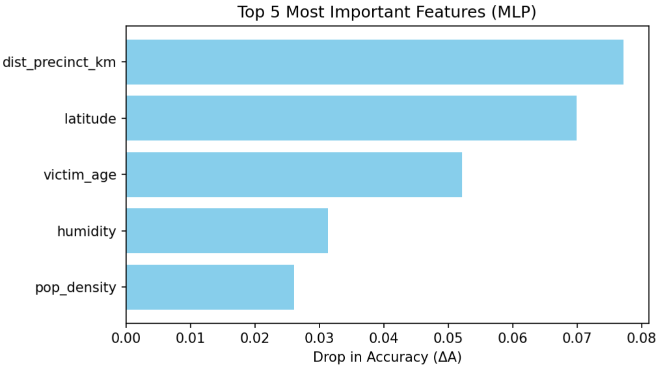
Architecture

- **Hidden layers:** 2 layers (64, 32 neurons)
- **Activation:** ReLU
- **Optimizer:** Adam
- **Regularization:** Early stopping

Performance

VAL Accuracy: 94% — Matches SVM performance!

Q6: Feature Importance (Permutation)



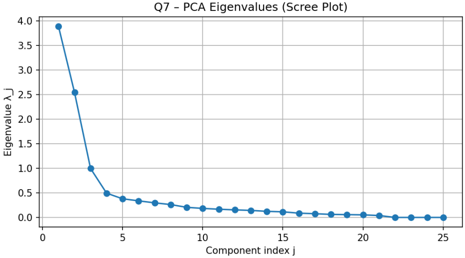
Top 5 Features:

1. dist_precinct_km ($\Delta A = 0.07$)
2. latitude ($\Delta A = 0.06$)
3. victim_age ($\Delta A = 0.05$)
4. humidity ($\Delta A = 0.03$)
5. pop_density ($\Delta A = 0.02$)

Q7: PCA for Dimensionality Reduction

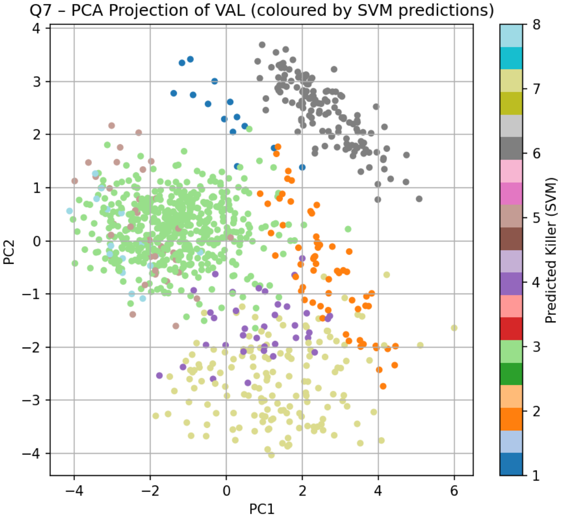
Purpose

- Visualize high-dimensional data in 2D
- Reduce noise and computational cost
- Preserve maximum variance



First few components capture most variance.

Q7: PC1–PC2 Scatter (Colored by Killer)



Q8: Unsupervised k-means in PCA Space

Approach

- Apply PCA (5 components) on TRAIN
- Run k-means with $k = 8$ (number of killers)
- Map clusters to killers via majority voting
- Evaluate on VAL

Result

Lower accuracy than supervised methods.
Natural clusters in feature space do not perfectly align with killer identities.

Q8: Interpretation

- k-means finds geometric clusters, not class labels
- Supervised methods leverage labeled data more effectively
- Useful for exploratory analysis, not final prediction

Key Takeaway

When labels are available, supervised learning outperforms unsupervised clustering for classification tasks.

Performance Comparison

Table: VAL Accuracy across all models

Model	**VAL Accuracy**
SVM (RBF kernel)	**94.8%**
MLP (2 hidden layers)	**94.4%**
Gaussian Bayes	90.5%
Linear Classifier	78.2%
PCA + k-means	Lower

Ranking: SVM $\approx$ MLP $>$ Bayes $>$ Linear $>$ k-means

Visual Insights Summary

Exploratory:

- GMM better than single Gaussian
- Victims: vulnerable age groups
- Spatial clustering around centers

Decision Boundaries:

- Gaussian: curved, smooth
- Linear: straight, restrictive
- SVM/MLP: complex, accurate

Most Important Features

Geographic (lat/long), victim info (age, gender), scene conditions (weather, type), weapon type.

Key Findings

1. **Nonlinear models excel:** SVM and MLP achieve 94% accuracy
2. **Feature importance:** Geographic + victim characteristics dominate
3. **Gaussian assumption works:** Bayes classifier at 90% is strong baseline
4. **Linear limits:** Only 78% — insufficient for complex patterns
5. **Unsupervised falls short:** k-means cannot match supervised performance

Main Contributions

What We Did

- Comprehensive comparison of 5 methods (generative, discriminative, nonlinear, unsupervised)
- Proper preprocessing pipeline (standardization, one-hot encoding)
- No data leakage (fit on TRAIN, apply to VAL/TEST)
- Systematic hyperparameter tuning
- Rich visualizations (heatmaps, ellipses, decision boundaries, PCA)

Limitations & Future Work

Limitations

- Limited to provided features (no time series, network analysis)
- Imbalanced classes (killers 3, 6, 7 dominate)
- No ensemble methods explored

Future Directions

- Ensemble models (voting, stacking)
- Deep learning (CNN, attention mechanisms)
- External data (social networks, historical patterns)
- Explainability (SHAP, LIME)

Thank You!

Questions?

Who Is The Killer? — Piraeus Vice

Β Μ , Γ Τ , Λ Φ
University of Piraeus