

PIRAEUS VICE: Ποιος είναι ο Δολοφόνος;

Μια Έρευνα Αναγνώρισης Προτύπων & Μηχανικής Μάθησης

Αλέξανδρος Σπατούλας (Π23175) Κλειώ Συρίγου (Π23180)

Τμήμα Πληροφορικής
Πανεπιστήμιο Πειραιώς

20 Φεβρουαρίου 2026

- 1 Εισαγωγή
- 2 Ανάλυση Δεδομένων
- 3 Επιβλεπόμενη Μάθηση
- 4 Μη Επιβλεπόμενη Μάθηση
- 5 Συμπεράσματα

- **Πλαίσιο:** Το Τμήμα Ανθρωποκτονιών "Piraeus Vice" ερευνά μια σειρά από εγκλήματα.
- **Στόχος:** Ταυτοποίηση 8 κατά συρροή δολοφόνων ($S = 8$) σε ένα σύνολο δεδομένων 4.000+ εγκλημάτων.
- **Κεντρικό Ερώτημα:**

"Για κάθε περιστατικό, ποιος είναι ο πιθανότερος δράστης;"

Διαθέτουμε τα εξής χαρακτηριστικά για κάθε έγκλημα:

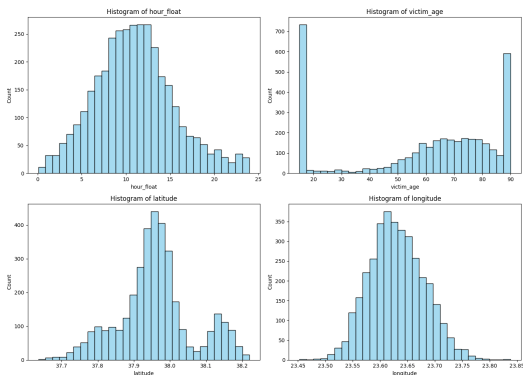
- **Χωρικά:** Γεωγραφικό Πλάτος (Latitude), Μήκος (Longitude), Απόσταση από Αστυνομικό Τμήμα.
- **Χρονικά:** Ώρα ημέρας (float), Ημερομηνία.
- **Περιβαλλοντικά:** Καιρός, Θερμοκρασία, Υγρασία.
- **Χαρακτηριστικά Θύματος:** Ηλικία, Φύλο.
- **Άλλα:** Όπλο, Τύπος Σκηνής Εγκλήματος.

Για την επίλυση του προβλήματος, ακολουθήσαμε τα εξής βήματα:

- 1 **Διερευνητική Ανάλυση (EDA):** Κατανόηση των κατανομών.
- 2 **Γενετικά Μοντέλα:** MLE και Gaussian Bayes.
- 3 **Διαχωριστικά Μοντέλα:** Γραμμικοί Ταξινομητές, SVM, Νευρωνικά Δίκτυα.
- 4 **Μη Επιβλεπόμενη Μάθηση:** PCA και K-Means Clustering.
- 5 **Αξιολόγηση:** Σύγκριση ακρίβειας (Accuracy) στο Validation Set.

Q1: Κατανομή της Ώρας

- Η εγκληματικότητα δεν κατανέμεται ομοιόμορφα μέσα στην ημέρα.
- Η κατανομή είναι **πολυτροπική** (multimodal), με σαφείς εξάρσεις δραστηριότητας.

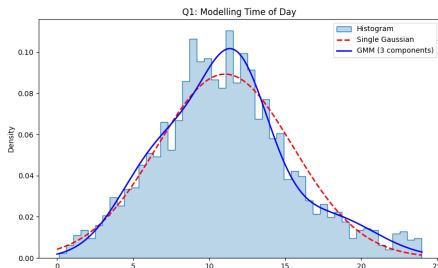


Σχήμα: Ιστογράμματα Χαρακτηριστικών

Q1: Μοντελοποίηση Ώρας (GMM)

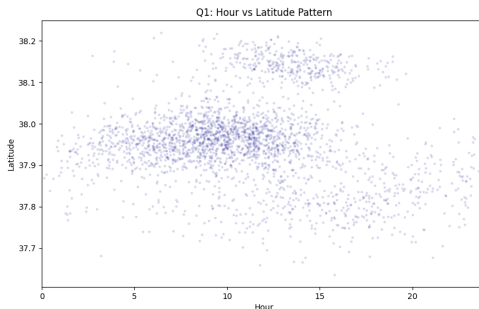
Συγκρίναμε δύο προσεγγίσεις για την κατανομή της ώρας:

- **Απλή Γκαουσιανή:** Αποτυγχάνει (κόκκινη γραμμή). Είναι πολύ "επίπεδη".
- **GMM (3 components):** Εντοπίζει επιτυχώς τις 3 ζώνες αιχμής: Πρωί, Μεσημέρι, Βράδυ (μπλε γραμμή).



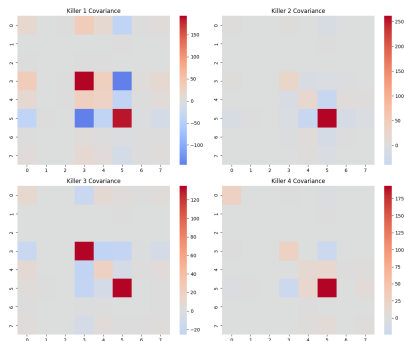
Q1: Χωροχρονικά Μοτίβα

- Εξετάζοντας τη σχέση Ώρας και Γεωγραφικού Πλάτους (Latitude).
- Παρατηρούμε **οριζόντιες λωρίδες** (bands).
- **Συμπέρασμα:** Συγκεκριμένοι δολοφόνοι δρουν σε συγκεκριμένα γεωγραφικά πλάτη και ώρες.



Q2: Εκτίμηση Παραμέτρων (MLE)

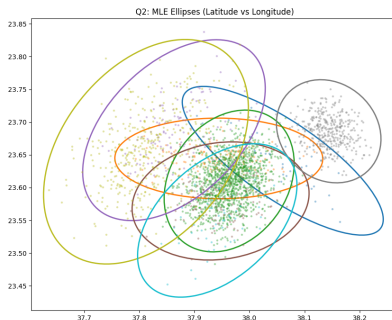
- Υπόθεση: Τα χαρακτηριστικά κάθε δολοφόνου ακολουθούν πολυμεταβλητή Γκαουσιανή Κατανομή $\mathcal{N}(\mu_k, \Sigma_k)$.
- Υπολογίσαμε τους Πίνακες Συνδιακύμανσης (Σ_k) για κάθε δράστη.



Σχήμα: Heatmaps Συνδιακύμανσης

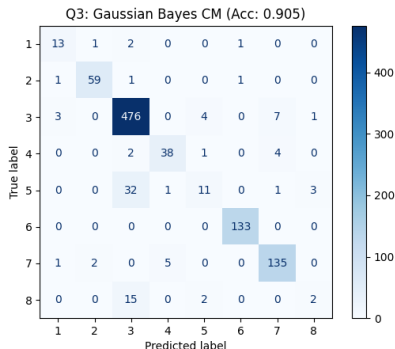
Q2: Γεωγραφική Δράση (Ελλείψεις)

- Οι ελλείψεις οριοθετούν την περιοχή δράσης (99% confidence).
- **Πρόβλημα:** Υπάρχει επικάλυψη (overlap) μεταξύ των περιοχών.
- Η τοποθεσία μόνη της δεν αρκεί πάντα για διαχωρισμό.



Q3: Gaussian Bayes Classifier

- Κατασκευάσαμε έναν ταξινομητή Bayes χρησιμοποιώντας τις παραμέτρους MLE και τις a priori πιθανότητες.
- Ακρίβεια (Validation):** $\approx 90.5\%$



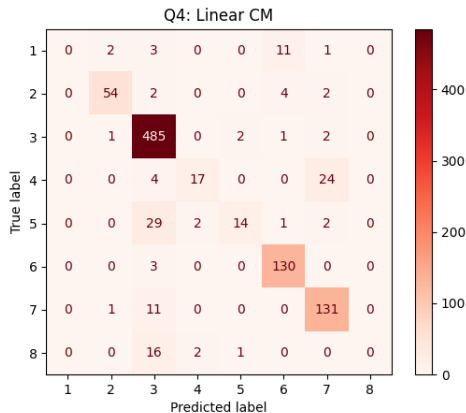
Σχήμα: Πίνακας Σύγχυσης (Bayes)

Q4: Γραμμικός Ταξινομητής

- Μοντέλο Ridge Classifier με όλα τα χαρακτηριστικά (One-Hot Encoded).
- **Ακρίβεια (Validation):** $\approx 87.6\%$
- **Παρατήρηση:** Χαμηλότερη επίδοση από τον Bayes.

Αυτό υποδηλώνει ότι τα όρια διαχωρισμού μεταξύ των δολοφόνων **δεν είναι γραμμικά**.

Q4: Αποτελέσματα Γραμμικού Μοντέλου



Σχήμα: Περισσότερα λάθη εκτός διαγωνίου.

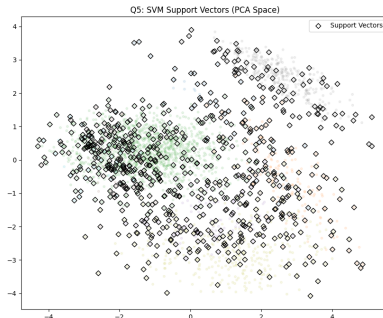
Q5: Support Vector Machines (SVM)

- Δοκιμάσαμε SVM με πυρήνα **RBF** (Radial Basis Function).
- Βελτιστοποίηση Υπερπαραμέτρων (Grid Search):
 - $C = 1$
 - $\gamma = 0.1$
- **Ακρίβεια (Validation):** $\approx 94.8\%$

Αποτέλεσμα: Αυτό είναι το βέλτιστο μοντέλο της μελέτης μας.

Q5: Οπτικοποίηση Περιοχών Απόφασης

- Οι ρόμβοι δείχνουν τα Support Vectors.
- Ο πυρήνας RBF δημιουργεί καμπύλα, "φυσικά" όρια που περικλείουν τις ομάδες πολύ πιο αποτελεσματικά από τις ευθείες γραμμές.



Q6: Πολυεπίπεδο Νευρωνικό Δίκτυο (MLP)

- **Αρχιτεκτονική:**

- Είσοδος: Όλα τα χαρακτηριστικά
- Hidden Layer 1: 64 νευρώνες (ReLU)
- Hidden Layer 2: 32 νευρώνες (ReLU)
- Έξοδος: 8 κλάσεις (Softmax)

- **Ακρίβεια (Validation):** $\approx 94.0\%$

- Εξαιρετική επίδοση, συγκρίσιμη με το SVM.

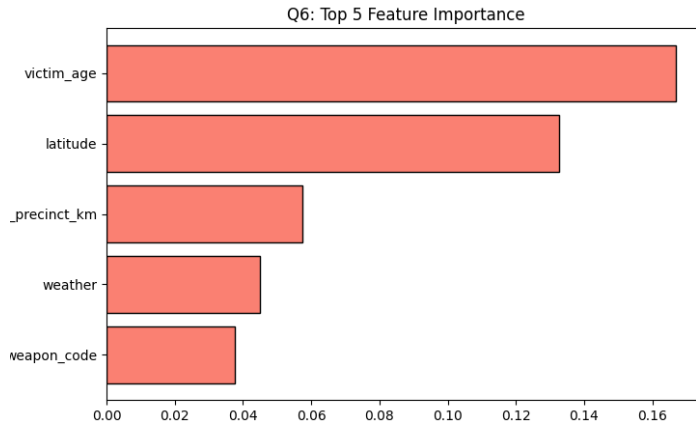
Q6: Τι "προδίδει" τον δολοφόνο;

Εφαρμόσαμε την τεχνική **Permutation Importance** για να βρούμε ποια χαρακτηριστικά επηρεάζουν περισσότερο την απόφαση του μοντέλου.

Τα Κρισιμότερα Χαρακτηριστικά

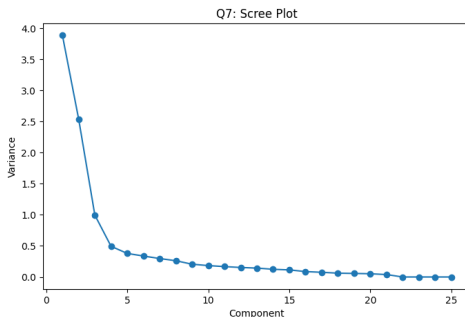
- 1 **Distance to Precinct** (Απόσταση από Αστ. Τμήμα) - *Το σημαντικότερο!*
- 2 **Victim Age** (Ηλικία Θύματος)
- 3 **Weather** (Καιρικές Συνθήκες)

Q6: Διάγραμμα Σπουδαιότητας Χαρακτηριστικών



Q7: Μείωση Διαστάσεων (PCA)

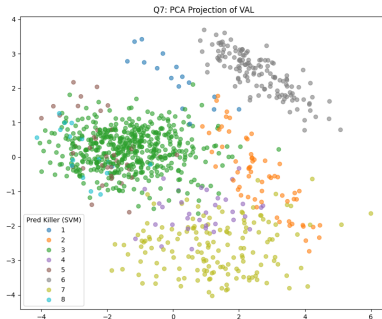
- Εφαρμόσαμε PCA για να προβάλουμε τα δεδομένα σε 2 διαστάσεις.
- Το Scree Plot δείχνει "αγκώνα" στις 2 πρώτες συνιστώσες.



Σχήμα: Scree Plot (Ιδιοτιμές)

Q7: Οπτικοποίηση στο Χώρο PCA

- Η προβολή στο 2D αποκαλύπτει 8 ξεκάθαρες, διαχωρισμένες ομάδες.
- Αυτός ο σαφής διαχωρισμός εξηγεί γιατί τα μοντέλα μας πέτυχαν τόσο υψηλή ακρίβεια.

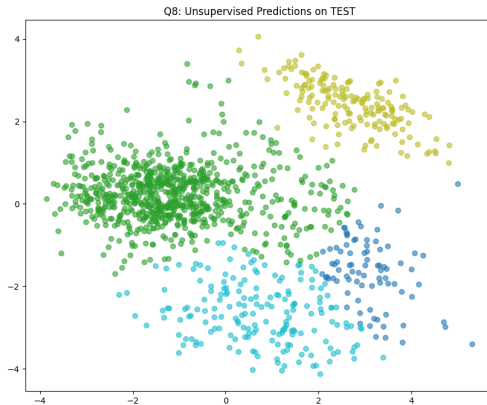


Q8: K-Means Clustering (Unsupervised)

- **Σενάριο:** Τι θα γινόταν αν δεν είχαμε τις ετικέτες (labels);
- Εφαρμόσαμε K-Means με $K = 8$ στα δεδομένα PCA.
- **Ακρίβεια:** $\approx 78.3\%$

Η υψηλή ακρίβεια επιβεβαιώνει ότι οι δολοφόνοι σχηματίζουν φυσικές συστάδες (clusters).

Q8: Προβλέψεις στο TEST Set



Σχήμα: Τελικές Προβλέψεις K-Means

Συγκεντρωτικά Αποτελέσματα

Μοντέλο	Ακρίβεια (Val)	Σχόλιο
Linear Classifier	87.6%	Αδυναμία σε μη-γραμμικά όρια
Gaussian Bayes	90.5%	Καλή γενίκευση
MLP (Neural Net)	94.0%	Ισχυρό μοντέλο
SVM (RBF)	94.8%	Βέλτιστη Επίδοση
K-Means (Unsup.)	78.3%	Εντυπωσιακό χωρίς ετικέτες

- 1 **Επιτυχής Ταυτοποίηση:** Μπορούμε να εντοπίσουμε τον δράστη με $\approx 95\%$ βεβαιότητα χρησιμοποιώντας SVM.
- 2 **Προφίλ Δραστών:** Οι δολοφόνοι έχουν πολύ συγκεκριμένο "Modus Operandi" (τόπος, χρόνος, επιλογή στόχου).
- 3 **Κλειδί:** Η απόσταση από το Αστυνομικό Τμήμα και η ηλικία του θύματος είναι τα πιο διακριτικά στοιχεία.
- 4 **Φυσική Δομή:** Οι δράστες σχηματίζουν φυσικές συστάδες, επιτρέποντας ακόμα και μη-επιβλεπόμενο εντοπισμό.



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

Ευχαριστούμε!

Ερωτήσεις;