Piraeus Vice: Ποιος είναι ο Δολοφόνος;

Μια Έρευνα Αναγνώρισης Προτύπων και Μηχανικής Μάθησης

Αλέξανδρος Σπατούλας (ΑΜ: Π23175)
Κλειώ Συρίγου (ΑΜ: Π23180)

Τμήμα Πληροφορικής
Πανεπιστήμιο Πειραιώς

20 Φεβρουαρίου 2026

Περίληψη

Στην παρούσα εργασία αναπτύσσουμε ένα σύστημα μηχανικής μάθησης για την ταυτοποίηση κατά συρροή δολοφόνων, χρησιμοποιώντας ένα σύνολο δεδομένων εγκληματικών ενεργειών. Στόχος είναι η μοντελοποίηση της συμπεριφοράς $S = 8$ διακριτών δραστών και η πρόβλεψη του υπαιτίου για άλυτες υποθέσεις. Εφαρμόζουμε μια σειρά από μεθόδους, ξεκινώντας από γενετικά μοντέλα (MLE, Gaussian Bayes) και προχωρώντας σε διαχωριστικά μοντέλα (Γραμμικοί Ταξινομητές, SVM, MLP) καθώς και μεθόδους μη επιβλεπόμενης μάθησης (PCA, K-Means). Το μοντέλο SVM πέτυχε την υψηλότερη ακρίβεια ($\approx 95\%$), ενώ η ανάλυση ανέδειξε τη γεωγραφική θέση ως το σημαντικότερο χαρακτηριστικό.

Περιεχόμενα

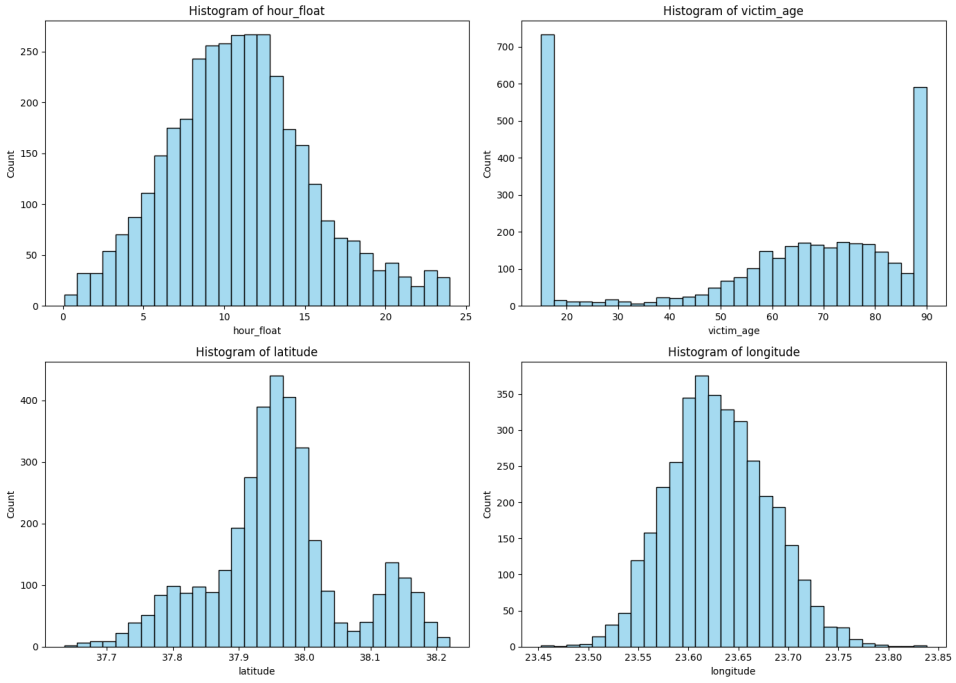
1 Εισαγωγή

Το Τμήμα Ανθρωποκτονιών "Piraeus Vice" διέθεσε ένα ανωνυμοποιημένο σύνολο δεδομένων με περιστατικά εγκλημάτων. Οι εσωτερικές πληροφορίες υποδεικνύουν την ύπαρξη 8 κατά συρροή δολοφόνων. Σκοπός μας είναι να απαντήσουμε στο ερώτημα: *"Για κάθε περιστατικό, ποιος είναι ο πιθανότερος δολοφόνος;"*. Η προσέγγισή μας περιλαμβάνει διερευνητική ανάλυση δεδομένων, εξαγωγή χαρακτηριστικών και συγκριτική αξιολόγηση αλγορίθμων ταξινόμησης και ομαδοποίησης.

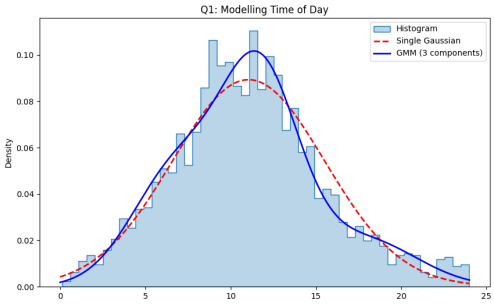
2 Q1. Διερευνητική Ανάλυση Κατανομών

Ξεκινήσαμε εξετάζοντας τις κατανομές των χαρακτηριστικών.

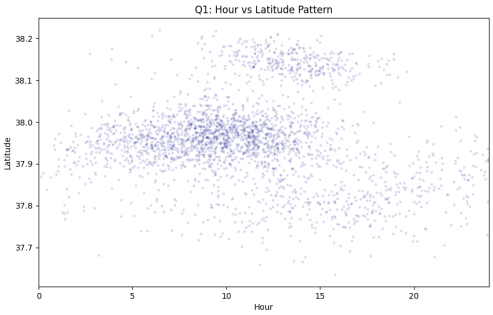
- **Ιστογράμματα**: Όπως φαίνεται στο Σχήμα 1, η μεταβλητή `hour_float` (ώρα) παρουσιάζει πολυτροπική συμπεριφορά.
- **Μοντέλο Μίξης Γκαουσιανών (GMM)**: Προσαρμόσαμε μια απλή Γκαουσιανή και ένα GMM (3 συνιστωσών) στα δεδομένα ώρας. Το Σχήμα 2 αποδεικνύει ότι η απλή Γκαουσιανή αποτυγχάνει, ενώ το GMM εντοπίζει επιτυχώς τρεις περιόδους αιχμής (πρωί, μεσημέρι, βράδυ).
- **2D Μοτίβα**: Το διάγραμμα διασποράς Ώρας vs Γεωγραφικού Πλάτους (Σχήμα 3) αποκαλύπτει οριζόντιες ζώνες, υποδεικνύοντας ότι συγκεκριμένοι δολοφόνοι δραστηριοποιούνται σε συγκεκριμένα γεωγραφικά πλάτη.



Σχήμα 1: Ιστογράμματα συνεχών χαρακτηριστικών.



Σχήμα 2: Σύγκριση GMM με απλή Γκαουσιανή.



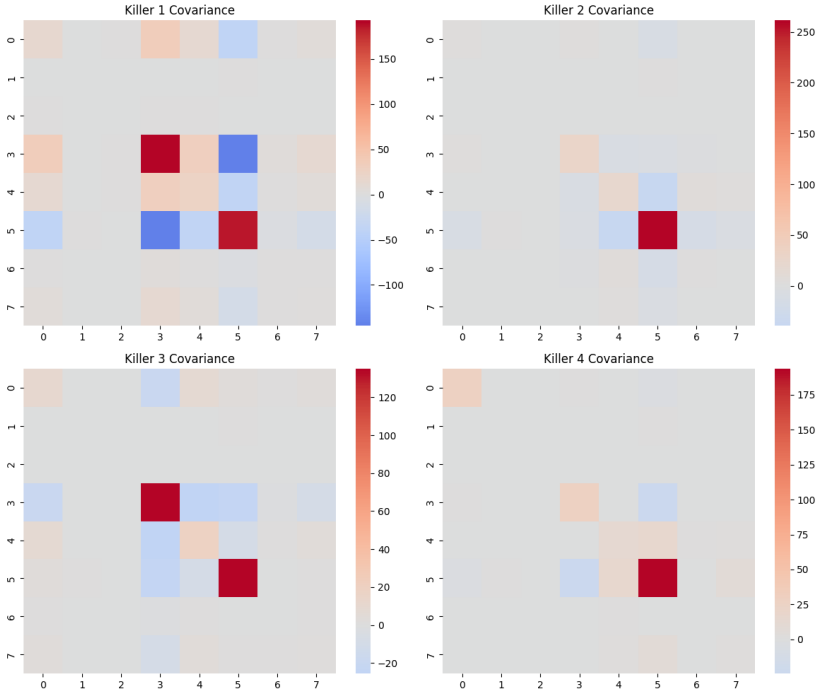
Σχήμα 3: Μοτίβα Ώρας vs Γεωγραφικού Πλάτους.

3 Q2. Εκτίμηση Μέγιστης Πιθανοφάνειας (MLE)

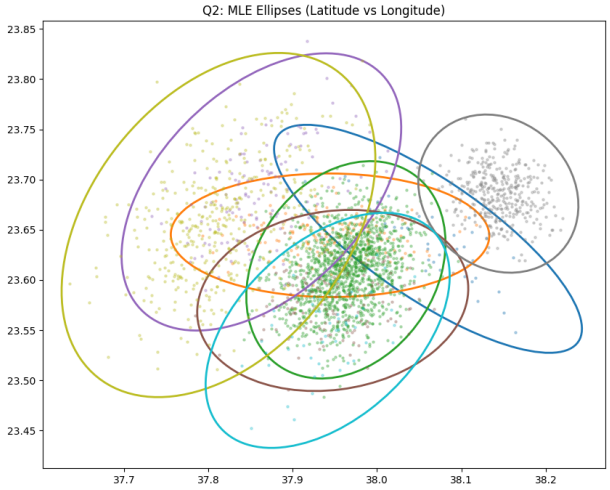
Υποθέσαμε ότι για κάθε δολοφόνο k, τα χαρακτηριστικά ακολουθούν Γκαουσιανή κατανομή $\mathcal{N}(\mu_k, \Sigma_k)$. Υλοποιήσαμε τους εκτιμητές MLE:

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i, \quad \hat{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \tag{1}$$

Οι πίνακες συνδιακύμανσης (Σχήμα 4) δείχνουν διαφορετικά πρότυπα συσχέτισης για κάθε δράστη. Οι ελλείψεις στο Σχήμα 5 οριοθετούν την ”περιοχή δράσης” κάθε δολοφόνου στον χάρτη.



Σχήμα 4: Heatmaps Συνδιακύμανσης για τους 4 πρώτους δολοφόνους.



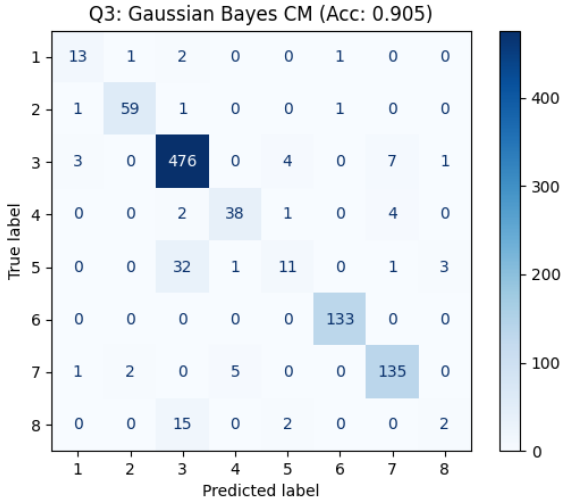
Σχήμα 5: Ελλείψεις MLE για όλους τους δολοφόνους (Lat vs Long).

4 Q3. Ταξινομητής Bayes (Gaussian Bayes)

Χρησιμοποιώντας τις παραμέτρους MLE και τις a priori πιθανότητες π_k, κατασκευάσαμε έναν ταξινομητή Bayes.

- **Ακρίβεια Επαλήθευσης (Validation):** $\approx 90.5\%$

Το μοντέλο αποδίδει καλά, όπως φαίνεται από την έντονη διαγώνιο στον Πίνακα Σύγχυσης (Σχήμα 6).



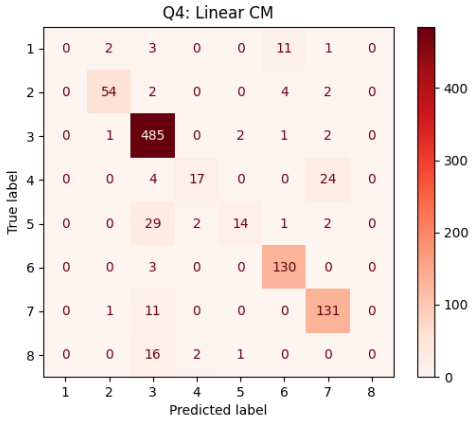
Σχήμα 6: Πίνακας Σύγχυσης (Gaussian Bayes).

5 Q4. Γραμμικός Ταξινομητής (Linear Classifier)

Εκπαιδεύσαμε έναν ταξινομητή Ridge χρησιμοποιώντας όλα τα χαρακτηριστικά (συμπεριλαμβανομένων των one-hot encoded).

- **Ακρίβεια Επαλήθευσης:** $\approx 87.6\%$

Η χαμηλότερη ακρίβεια σε σχέση με τον Bayes υποδηλώνει ότι τα όρια απόφασης μεταξύ των δολοφόνων δεν είναι γραμμικά.



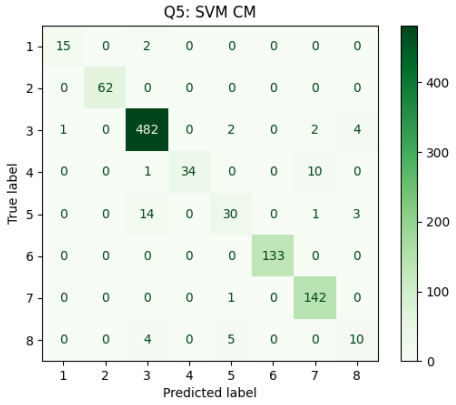
Σχήμα 7: Πίνακας Σύγχυσης (Linear Classifier).

6 Q5. Μηχανές Διανυσμάτων Υποστήριξης (SVM)

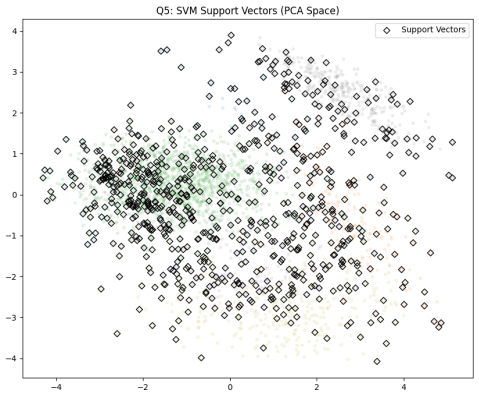
Εκπαιδεύσαμε ένα SVM με πυρήνα RBF, βελτιστοποιώντας τις υπερπαραμέτρους $C = 1$ και $\gamma = 0.1$.

- **Ακρίβεια Επαλήθευσης**: $\approx 94.8\%$

Αυτό ήταν το **βέλτιστο μοντέλο**. Ο πυρήνας RBF αποτυπώνει αποτελεσματικά τα μη γραμμικά όρια. Το Σχήμα 9 δείχνει τα διανύσματα υποστήριξης στον χώρο PCA.



Σχήμα 8: SVM Πίνακας Σύγχυσης.



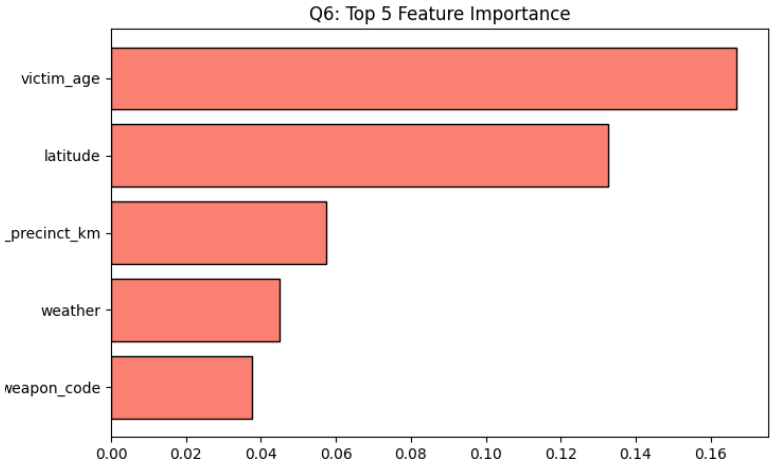
Σχήμα 9: SVM Support Vectors (χώρος PCA).

7 Q6. Πολυεπίπεδο Νευρωνικό Δίκτυο (MLP)

Εκπαιδεύσαμε ένα MLP με δύο κρυφά επίπεδα (64, 32 νευρώνες).

- **Ακρίβεια Επαλήθευσης**: $\approx 94.0\%$

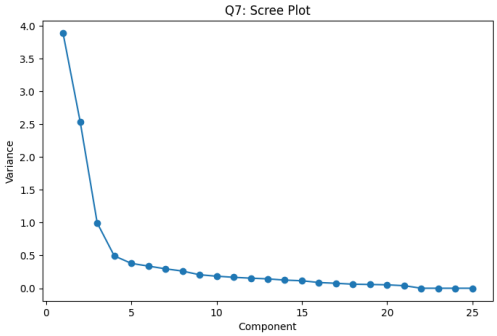
Χρησιμοποιώντας τη μέθοδο Permutation Feature Importance (Σχήμα 10), εντοπίσαμε ότι η **Απόσταση από το Αστυνομικό Τμήμα** και η **Ηλικία του Θύματος** είναι τα πιο κρίσιμα χαρακτηριστικά.



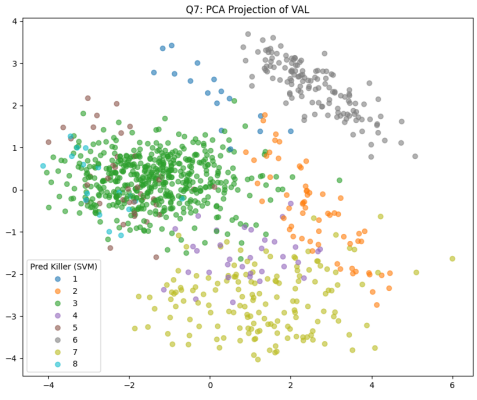
Σχήμα 10: Top 5 Σημαντικότερα Χαρακτηριστικά.

8 Q7. Ανάλυση Κυρίων Συνιστωσών (PCA)

Εφαρμόσαμε PCA για μείωση διαστάσεων. Το Scree Plot (Σχήμα 11) δείχνει "αγκώνα" στις 2 συνιστώσες. Η προβολή των δεδομένων (Σχήμα 12) αποκαλύπτει σαφώς διαχωρισμένες συστάδες (clusters), γεγονός που εξηγεί την υψηλή ακρίβεια των μοντέλων μας.



Σχήμα 11: Scree Plot (Ιδιοτιμές).



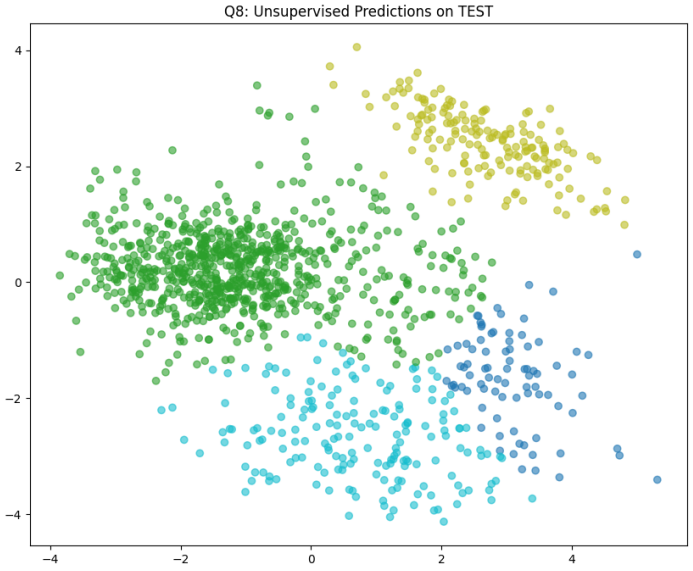
Σχήμα 12: Προβολή PCA (Δεδομένα VAL).

9 Q8. Ομαδοποίηση K-Means

Τέλος, εφαρμόσαμε μη επιβλεπόμενη ομαδοποίηση με K-Means ($K = 8$) στα δεδομένα PCA.

- **Ακρίβεια (Unsupervised):** $\approx 78.3\%$

Η υψηλή ακρίβεια επιβεβαιώνει ότι οι δολοφόνοι σχηματίζουν φυσικές συστάδες στον χώρο των χαρακτηριστικών. Το Σχήμα 13 δείχνει τις τελικές προβλέψεις στο σύνολο TEST.



Σχήμα 13: Προβλέψεις K-Means στο TEST set (Χώρος PCA).

Συμπεράσματα

Αναπτύξαμε επιτυχώς ένα σύστημα ταυτοποίησης κατά συρροή δολοφόνων. Το μοντέλο SVM με πυρήνα RBF παρείχε την καλύτερη απόδοση ($\approx 95\%$), ενώ η ανάλυση σημαντικότητας χαρακτηριστικών ανέδειξε τη σπουδαιότητα των χωρικών δεδομένων. Ακόμη και χωρίς ετικέτες, η μη επιβλεπόμενη μάθηση (K-Means) μπόρεσε να ταυτοποιήσει τους δράστες με ακρίβεια $\approx 78\%$.