ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
Πρόγραμμα Μεταπτυχιακών Σπουδών
“Πληροφοριακά Συστήματα & Υπηρεσίες”
Ειδίκευση: “Προηγμένα Πληροφοριακά Συστήματα”

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ: 2024-2025

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ DATA MINING ΚΑΙ ΑΣΑΦΟΥΣ ΛΟΓΙΚΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΣΗΜΑΤΩΝ ΑΣΤΕΡΩΝ ΝΕΤΡΟΝΙΩΝ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΣΤΟΙΧΕΙΑ ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΦΟΙΤΗΤΗ:

ΙΩΑΝΝΗΣ ΜΑΝΩΛΗΣ

ΜΕ2344

ΠΕΙΡΑΙΑΣ 2025

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

ΔΗΜΟΣΘΕΝΗΣ ΚΥΡΙΑΖΗΣ

ΜΙΧΑΗΛ ΦΙΛΙΠΠΑΚΗΣ

ΜΑΡΙΑ ΧΑΛΚΙΔΗ

Περίληψη

Η παρούσα μελέτη διερευνά την απόδοση τόσο κλασικών όσο και ασαφών αλγορίθμων ταξινόμησης στην πρόβλεψη της αυθεντικότητας σημάτων pulsar, χρησιμοποιώντας ένα ευρέως αποδεκτό σύνολο χαρακτηριστικών που έχει προταθεί ως πρότυπη αναφορά για τις μελλοντικές έρευνες ανίχνευσης. Εκπαιδεύτηκαν και αξιολογήθηκαν έξι μοντέλα μηχανικής μάθησης, συμπεριλαμβανομένων τεσσάρων κλασικών αλγορίθμων ταξινόμησης (Λογιστική Παλινδρόμηση, Random Forest, SVM και XGBoost), καθώς και δύο προσαρμοσμένων ταξινομητών ασαφούς λογικής (Fuzzy k-NN και Fuzzy Decision Tree). Τα ασαφή μοντέλα ενσωματώνουν μηχανισμούς ταξινομικής ασαφοποίησης και διαδικασίες εξαγωγής κανόνων για τη βελτίωση της ερμηνευσιμότητας. Τα αποτελέσματα έδειξαν ότι τα κλασικά μοντέλα επιτυγχάνουν υψηλή ακρίβεια και ανθεκτικότητα, με τα Random Forest και SVM να παρουσιάζουν ιδιαίτερα υψηλές αποδόσεις. Ωστόσο, οι ασαφείς προσεγγίσεις προσφέρουν σημαντικά πλεονεκτήματα στη διαφάνεια των αποφάσεων που τις καθιστούν χρήσιμες σε πεδία όπου η ερμηνευσιμότητα είναι κρίσιμη. Η μελέτη επιβεβαιώνει επίσης την καταλληλότητα των επιλεγμένων χαρακτηριστικών για την αξιόπιστη ανίχνευση σημάτων pulsar και προτείνει κατευθύνσεις για μελλοντική έρευνα ώστε να αξιολογηθεί περαιτέρω η επάρκειά τους. Η έρευνα υλοποιήθηκε χρησιμοποιώντας τις κατάλληλες βιβλιοθήκες της γλώσσας προγραμματισμού Python.

Λέξεις-κλειδιά:
Ταξινόμηση σημάτων, Ασαφής λογική, Μηχανική μάθηση, Εξαγωγή ασαφών κανόνων, Ανίχνευση pulsar

Abstract

This study investigates the performance of both classical and fuzzy classification algorithms in predicting the authenticity of pulsar signals, using a widely accepted feature set proposed as a standard reference for future screening research. Six models were trained and evaluated, including four classical machine learning algorithms (Logistic Regression, Random Forest, SVM and XGBoost), as well as two custom fuzzy logic classifiers (Fuzzy k-NN and Fuzzy Decision Tree). The fuzzy models incorporate fuzzification mechanisms and rule extraction procedures to enhance interpretability. Results demonstrate that the classical models achieve high accuracy and robustness, with Random Forest and SVM performing particularly well. However, the fuzzy approaches offer meaningful advantages in decision transparency, making them useful in domains where interpretability is critical. The study also confirms the suitability of the selected features for reliable pulsar signal screening and suggests future research directions to further assess their adequacy. The study was implemented using the appropriate libraries of the Python programming language.

Keywords:

Signal classification, Fuzzy logic, Machine learning, Rule extraction, Pulsar detection

Περιεχόμενα

Εισαγωγή

Κύριο αντικείμενο της παρούσας εργασίας είναι η συγκριτική μελέτη των επιδόσεων κάποιων κλασικών αλγορίθμων ταξινόμησης καθώς και ορισμένων λιγότερων συνηθισμένων, οι οποίοι ενσωματώνουν στοιχεία ασαφούς λογικής (fuzzy logic). Η σύγκριση των υπό μελέτη μοντέλων επιχειρήθηκε στη βάση της αξιολόγησης των αποτελεσμάτων της εκπαίδευσής τους με ένα μεγάλο σύνολο δεδομένων από αστρονομικές καταγραφές ηλεκτρομαγνητικών σημάτων, ορισμένα από τα οποία αντιστοιχούν σε σήματα εκπομπών αστέρων νετρονίων, ενώ η μεγάλη πλειοψηφία τους αφορά σε διαστρικό θόρυβο ή σε σήματα ραδιο-εκπομπών προερχόμενων από άλλα είδη διαστημικών πηγών. Ουσιαστικά πρόκειται για ένα πρόβλημα δυαδικής (binary) ταξινόμησης, όπου το μοντέλο καλείται να τοποθετήσει τις υποψήφιες εγγραφές σε μία από τις δύο δυνατές κλάσεις, στη θετική (positive) με σήμανση 1 που αφορά στα σήματα των πραγματικών εκπομπών από αστέρες νετρονίων ή στην αρνητική (negative) με σήμανση 0, στην οποία τοποθετούνται τα σήματα ραδιο-εκπομπών που προέρχονται από άλλου είδους πηγές ή/και διαστημικό θόρυβο.

Η ταξινόμηση (classification) αποτελεί μία από τις πλέον κεντρικές μεθόδους εξόρυξης γνώσης από μεγάλα σύνολα δεδομένων στο πεδίο της μηχανικής μάθησης. Πρόκειται για ένα από τα βασικότερα είδη επιβλεπόμενης μάθησης (supervised learning), όπου ο στόχος του υπολογιστικού συστήματος συνίσταται στο να μάθει να αντιστοιχεί σύνολα παρατηρήσεων (δεδομένα εισόδου) σε προκαθορισμένες κατηγορίες ή ετικέτες (labels). Η διαδικασία αυτή βρίσκει εφαρμογή σε πληθώρα πρακτικών πεδίων, όπως η ιατρική διάγνωση, η ανίχνευση απάτης, η επεξεργασία γλώσσας, αλλά και σε εκείνο της “καθαρής” επιστημονικής έρευνας, όπως στην παρούσα μελέτη, όπου αξιοποιείται για την αυτόματη αναγνώριση ουράνιων σωμάτων όπως οι pulsars. Κατά την εκπαίδευση ενός ταξινομητή, παρέχεται ένα σύνολο δεδομένων εκπαίδευσης το οποίο αποτελείται από εγγραφές – δείγματα (π.χ. διανύσματα χαρακτηριστικών) και τις αντίστοιχες ετικέτες τους. Ο στόχος συνίσταται στη δημιουργία ενός μοντέλου που να μπορεί να γενικεύσει και να προβλέπει σωστά την κατηγορία – κλάση νέων άγνωστων δειγμάτων. Η απόδοση του μοντέλου, μετά την εκπαίδευση, αξιολογείται στη βάση καλά ορισμένων τυπικών μετρικών όπως η ακρίβεια (accuracy), η ανάκληση (recall), η ακρίβεια θετικών προβλέψεων (precision), η καμπύλη ROC-AUC κ.α..

Το πρόβλημα ταξινόμησης που μελετά η παρούσα έρευνα προέρχεται από τον χώρο της αστροφυσικής. Η ραγδαία αναπτυσσόμενη τεχνολογία παραγωγής ισχυρών τηλεσκοπίων και ραδιο-τηλεσκοπίων έχει οδηγήσει τις τελευταίες δύο δεκαετίες σε μια “έκρηξη” του όγκου των παρατηρησιακών δεδομένων, ενώ τα προγράμματα των τεράστιων δικτύων από κεραίες ραδιο-παρατήρησης που πρόκειται να μπουν σύντομα σε εφαρμογή όπως Square Kilometre Array (SKA), αναμένεται να προκαλέσουν μια κυριολεκτική *αλλαγή Παραδείγματος* (Paradigm shift) στα πλαίσια του κλάδου. Από την πρώτη ήδη δεκαετία του 21^{ου} αιώνα ο αυξανόμενος όγκος των παρατηρησιακών δεδομένων οδήγησε στην ταχεία ανάπτυξη αυτοματοποιημένων μεθόδων διαλογής και στην υιοθέτηση τεχνικών του machine learning. Ωστόσο, οι όγκοι των δεδομένων που παράγει η νέα παρατηρησιακή τεχνολογία απαιτούν περαιτέρω εκλέπτυνση των αυτοματοποιημένων διαδικασιών διαλογής και επεξεργασίας των ραδιο-σημάτων, ώστε αυτές να καθίστανται αποτελεσματικές στη real time διαχείριση της πληροφορίας, μιας και οι δυνατότητες αποθήκευσης καθίστανται περιορισμένες, λόγω του πολύ μεγάλου όγκου.

Η κρίσιμη κλάση των δεδομένων που αναλύονται εδώ αφορά σε μια ιδιαίτερα σημαντική κατηγορία αστρονομικών αντικειμένων, τις ραδιοπηγές που είναι γνωστές με το όνομα pulsars. Πρόκειται για περιστρεφόμενα άστρα νετρονίων, τα οποία προέρχονται από τα

εξαιρετικά πυκνά απομεινάρια των υπερκαινοφανών εκρήξεων (supernova) που έπονται της βαρυτικής κατάρρευσης μεγάλων άστρων και τα οποία εκπέμπουν στενά ραδιοσήματα από τους μαγνητικούς τους πόλους. Καθώς το άστρο περιστρέφεται, οι δέσμες ακτινοβολίας "σαρώνουν" τον μεσοαστρικό χώρο και ανιχνεύονται στη Γη ως *περιοδικοί παλμοί*, όμοιοι με τις εκπομπές ενός "κοσμικού φάρου". Η σταθερότητα της περιστροφής ορισμένων εξ αυτών είναι τόσο ακριβής που συγκρίνεται με την ακρίβεια των ατομικών ρολογιών. Η μελέτη τους από τους αστροφυσικούς έχει τεράστια σημασία για την κατανόηση ακραίων φυσικών ορίων, όπως η συμπεριφορά της ύλης σε πολύ ισχυρά βαρυτικά και μαγνητικά πεδία, μιας και το εσωτερικό ενός τέτοιου αστέρα προσομοιάζει, από σωματιδιακή άποψη, με τον πυρήνα ενός γιγάντιου ατόμου. Επιπλέον, οι ακριβείς παλμοί τους χρησιμοποιούνται ως εργαλεία ελέγχου της γενικής θεωρίας της σχετικότητας, ενώ συμβάλλουν ισχυρά και στο πεδίο της έρευνας για τον εντοπισμό βαρυτικών κυμάτων, καθώς και σε εκείνο που αφορά στην χαρτογράφηση των δυναμικών αλλαγών στον μεσοαστρικό χώρο.

Οι διαδικασίες διαλογής των παλμών των pulsars στις μεγάλες αστρονομικές έρευνες της περασμένης δεκαετίας, όπως το πρόγραμμα LOTAAS (LOFAR Tied-Array All-Sky Survey), από τις χειροκίνητες (μη-αυτοματοποιημένες) παραδοσιακές μεθόδους πέρασαν πολύ γρήγορα σε αυτοματοποιημένες μορφές με χρήση λογισμικού και μεθόδων μηχανικής μάθησης. Οι διαδικασίες αυτές έχουν στον πυρήνα τους την εφαρμογή στατιστικών κυρίως ελέγχων επί των χαρακτηριστικών των σημάτων που παρέχει η παρατήρηση. Η βασική μονάδα των εν λόγω χαρακτηριστικών είναι το απλό, αλλά εξαιρετικά κομβικό, μέγεθος του λόγου *σήματος προς θόρυβο* (Signal-to-Noise Ratio, S/N ή SNR) και αποτελούν συνήθως σύνθετες στατιστικές ποσότητες, οι οποίες περιγράφουν αναλυτικά κρίσιμες πτυχές του σήματος, αναπτύσσοντας περαιτέρω την πληροφορία του λόγου SNR. Παρά την τεράστια πρόοδο που έχει σημειωθεί μέσω της εισαγωγής των αυτοματοποιημένων μεθόδων διαλογής, υφίστανται ακόμη ισχυρές προκλήσεις που οφείλουν να αντιμετωπιστούν και οι οποίες πρόκειται να ενταθούν ακόμη περισσότερο με τη νέα "έκρηξη δεδομένων" που θα επιφέρει σύντομα η νέα παρατηρησιακή τεχνολογία. Ένα από τα κύρια προβλήματα είναι η φύση των γνωρισμάτων που περιγράφουν τα παρατηρησιακά στιγμιότυπα με τα οποία εκπαιδεύονται οι αλγόριθμοι. Δεν έχει καταστεί ακόμη κατορθωτό να υιοθετηθεί και να τυποποιηθεί ένα ενιαίο σύνολο γνωρισμάτων που να είναι ανεξάρτητο από την εκάστοτε επιμέρους έρευνα (survey-independent), ώστε να διευκολύνεται η διαλειτουργικότητα των δεδομένων, να ελαχιστοποιούνται οι μεροληψίες (biases) και τα φαινόμενα επιλογής, να αξιολογείται με τη χρήση ενός ενιαίου στατιστικού πλαισίου που να επιτρέπει τη σύγκριση και την αναπαραγωγιμότητα, καθώς και να είναι θωρακισμένο απέναντι στο ενδεχόμενο υψηλής διαστατικότητας (high dimensionality).

Το dataset που χρησιμοποιήθηκε για την εκπαίδευση των αλγόριθμων ταξινόμησης στο πλαίσιο της παρούσας εργασίας συνιστά μέρος της επιστημονικής προσπάθειας καθορισμού ενός συνόλου γνωρισμάτων (features) που να υπερβαίνει τα προηγούμενα προβληματικά σημεία. Ακόμη και το κεντρικό μέγεθος του λόγου S/N εμφανίζει αναξιοπιστία στην περιοχή των χαμηλών τιμών του, δηλαδή στις περιπτώσεις των πολύ αδύναμων σημάτων (μικρή τιμή του S) καθώς και σε εκείνες της παρουσίας ισχυρού RFI (μεγάλη τιμή του N). Τα γνωρίσματα που εισάγονται στο dataset που χρησιμοποιήθηκε αφορούν σε στατιστικές ποσότητες που σχετίζονται με δύο ευρέως χρησιμοποιούμενα μαθηματικά εργαλεία στην έρευνα γύρω από τον εντοπισμό των pulsars, το *ολοκληρωμένο προφίλ* και την *καμπύλη DM-SNR*.

Το ολοκληρωμένο προφίλ (integrated or folded profile) αποτελεί μια χαρακτηριστική αναπαράσταση του παλμικού του σήματος, η οποία προκύπτει από τη στοίχιση και τον μέσο όρο πολλών διαδοχικών παλμών με βάση την περίοδο περιστροφής του αστέρα. Μέσω αυτής

της διαδικασίας, γνωστής ως *folding*, ενισχύεται το σταθερά επαναλαμβανόμενο σήμα του pulsar ενώ καταστέλλεται ο τυχαίος θόρυβος, αποκαλύπτοντας τη δομή και τη μορφολογία του παλμού σε μία πλήρη περίοδο. Το προφίλ αυτό είναι το αποτέλεσμα ολοκλήρωσης τόσο στον χρόνο όσο και στη συχνότητα, και αποδίδει την ενέργεια του σήματος ως προς το φαινομενικό μήκος του παλμού (longitude). Από την άλλη, η καμπύλη DM-SNR (Dispersion Measure – Signal-to-Noise Ratio) απεικονίζει τη μεταβολή της αναλογίας σήματος προς θόρυβο (SNR) σε συνάρτηση με διαφορετικές τιμές του μέτρου διασποράς (DM), το οποίο εκφράζει την ολική ποσότητα ελεύθερων ηλεκτρονίων ανά μονάδα επιφάνειας που διασχίζει το σήμα στο διαστρικό μέσο. Η αλληλεπίδραση του σήματος με τα ελεύθερα ηλεκτρόνια του μεσοαστρικού χώρου αποτελεί την κύρια αιτία διασποράς και αποδυνάμωσής του. Μέσω της DM-SNR curve είναι δυνατός ο προσδιορισμός της βέλτιστης τιμής του DM, για την οποία το σήμα ευθυγραμμίζεται ορθότερα στις διαφορετικές συχνότητες και μεγιστοποιείται ο λόγος SNR. Η ερευνητική προσπάθεια καθορισμού ενός καθολικού συνόλου γνωρισμάτων, προϊόν της οποίας είναι και το dataset που χρησιμοποιήθηκε εδώ, επικεντρώνει σε οκτώ στατιστικές ποσότητες που αποτελούν χαρακτηριστικά τόσο του folded profile όσο και της DM-SNR curve.

Οι εγγραφές του dataset που χρησιμοποιήθηκε αποτελούν παρατηρησιακά προϊόντα της προηγμένης ραδιο-αστρονομικής έρευνας HTRU-2 (High Time Resolution Universe Survey 2), η οποία επικεντρώνει στην ανίχνευση νέων pulsar, με έμφαση στους millisecond pulsars και στα συστήματα υψηλής επιτάχυνσης, μέσω παρατηρήσεων υψηλής χρονικής και συχνικής ανάλυσης και αξιοποιώντας τεχνικές πολυδέσμης και λεπτομερούς επεξεργασίας σήματος. Με τα δεδομένα αυτά εκπαιδεύτηκαν και αξιολογήθηκαν τέσσερις κλασικοί ταξινομητές και δύο ακόμη οι οποίοι ενσωματώνουν στοιχεία ασαφούς λογικής. Συγκεκριμένα, οι κλασικοί αλγόριθμοι που επιλέχθηκαν είναι οι *Logistic Regression*, *Random Forest*, *Support Vector Machine (SVM)* και *XGBoost*. Η επιλογή των μοντέλων αυτών έγινε λόγω της διαφορετικής τους φύσης και συμπεριφοράς τους σε προβλήματα δυαδικής ταξινόμησης, με στόχο την όσο το δυνατόν πληρέστερη συγκριτική αποτίμηση των αποτελεσμάτων. Όσον αφορά στους fuzzy ταξινομητές, επιλέχθηκαν οι *Fuzzy k-Nearest Neighbors* (*Fuzzy k-NN*) και *Fuzzy Decision Tree*. Η επιλογή να εξεταστούν οι ταξινομητικές ικανότητες και η απόδοση των fuzzy μοντέλων έχει να κάνει κυρίως με τη δυνατότητά τους να παρέχουν ερμηνεύσιμα αποτελέσματα υπό μορφή ασαφών κανόνων, ενισχύοντας έτσι τη διαφάνεια και την κατανόηση των αποφάσεών τους - στοιχείο ιδιαίτερα σημαντικό στο πεδίο της ανάλυσης επιστημονικών δεδομένων. Πέρα από την αξιολόγηση της επάρκειας των υπό μελέτη μοντέλων, οι αποδόσεις της εκπαίδευσής τους με το συγκεκριμένο dataset μπορούν να καταστούν κριτήριο της ποιότητας των γνωρισμάτων που προτείνονται προς καθολική χρήση στο πεδίο της έρευνας για την ανίχνευση pulsars.

Η αξιολόγηση και η συγκριτική αποτίμηση των μοντέλων πραγματοποιήθηκε στη βάση ενός συνόλου πολλαπλών μετρικών απόδοσης, όπως η *ακρίβεια* (accuracy), η *ισορροπημένη ακρίβεια* (balanced accuracy), η *precision*, η *recall*, ο *f1-score* και ορισμένες ακόμη, ενώ η ανάλυση συμπληρώθηκε με τα αποτελέσματα που παρέχουν οι καμπύλες *ROC* και *Precision-Recall*. Στους fuzzy ταξινομητές πραγματοποιήθηκε επιπλέον η ενδεικτική *εξαγωγή ασαφών κανόνων* (*fuzzy rule extraction*), η ερμηνεία των οποίων υποδεικνύει μοτίβα που οδηγούν στην αναγνώριση των σημάτων που προέρχονται από pulsars. Η επιλογή των μετρικών για την αξιολόγηση των μοντέλων καθορίστηκε, ως έναν βαθμό και από τη φύση του dataset, οι λεπτομέρειες της οποίας αναδείχθηκαν μέσω στατιστικών ελέγχων που προηγήθηκαν της εκπαίδευσης. Μέσω των ελέγχων αυτών καθορίστηκαν και οι απαιτήσεις προεπεξεργασίας, ώστε τα αποτελέσματα της εκπαίδευσης να είναι όσο το δυνατό πιο αξιόπιστα. Το κεντρικό προβληματικό στοιχείο για την εκπαίδευση των μοντέλων που ανέδειξαν οι συγκεκριμένοι

έλεγχοι αφορά στην ισχυρή ανισορροπία των κλάσεων του dataset, η οποία αντιμετωπίστηκε μέσω της τεχνικής προεπεξεργασίας SMOTE (Synthetic Minority Over-sampling Technique), καθώς και της ρύθμισης balanced στους αλγόριθμους όπου αυτή είναι ικανή να εξαλείψει τις συνέπειες της ανισορροπίας. Τόσο οι προκαταρκτικοί έλεγχοι της στατιστικής φύσης των δεδομένων όσο και οι κινήσεις προεπεξεργασίας, καθώς και εκπαίδευση – αξιολόγηση των μοντέλων, πραγματοποιήθηκαν με χρήση των δυνατότητων που προσφέρουν στο πεδίο της μηχανικής μάθησης οι κατάλληλες βιβλιοθήκες της γλώσσας προγραμματισμού Python.

Περιγραφή Dataset

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων αντλήθηκε από την ακόλουθη διεύθυνση https://archive.ics.uci.edu/dataset/372/htru2 του UC Irvine Machine Learning repository και αφορά σε ένα μεγάλο CSV αρχείο με 17898 εγγραφές με 8 γνωρίσματα (features) και μια επιπλέον στήλη που αντιστοιχεί στην κλάση. Οι εγγραφές αφορούν σε σήματα υποψήφιων pulsars, τα οποία κατέγραψαν τα αστρονομικά όργανα στο πλαίσιο της σχετικά πρόσφατης μεγάλης κλίμακας παρατηρησιακής έρευνας High Time Resolution Universe Pulsar Survey (HTRU) και τα οποία ελέγχθηκαν και επεξεργάστηκαν από ανθρώπινους αναλυτές.

Περιλαμβάνονται συνολικά 16.259 ψευδείς εγγραφές (spurious examples) που αντιστοιχούν σε ραδιοφωνικές παρεμβολές ανθρώπινης προέλευσης (RFI), θόρυβο, καθώς και εκπομπές διαστημικών πηγών διαφορετικής φύσης από τα pulsars και 1.639 πραγματικές εγγραφές (real pulsar examples) που αντιστοιχούν σε εκπομπή σήματος από αστέρα νετρονίων. Τα 8 γνωρίσματα των εγγραφών αυτών αφορούν σε στατιστικά μεγέθη που προέκυψαν από την ανάλυση των ενισχυμένων μορφών των σημάτων των καταγεγραμμένων παλμών.

Τα 4 πρώτα αντιστοιχούν σε ορισμένες κρίσιμες μεταβλητές του λεγόμενου ολοκληρωμένου προφίλ του παλμού (integrated pulse profile). Πρόκειται για μια σειρά συνεχής μεταβλητές, οι οποίες περιγράφουν το ενισχυμένο σήμα ως προς τον χρόνο και ως προς τη συχνότητα. Οι μεταβλητές που εμφανίζονται εδώ είναι:

1. **Mean of the integrated profile (μέση τιμή του ολοκληρωμένου προφίλ):** Υπολογίζει τη μέση τιμή του προφίλ του παλμού. Αντανακλά το γενικό επίπεδο του σήματος και βοηθά στην κατανόηση της ισχύος του σήματος.
2. **Standard deviation of the integrated profile (τυπική απόκλιση του ολοκληρωμένου προφίλ):** Μετρά τη διασπορά των τιμών γύρω από τη μέση τιμή του ολοκληρωμένου προφίλ. Όσο μεγαλύτερη είναι η τιμή της, τόσο μεγαλύτερη είναι η μεταβλητότητα του σήματος.
3. **Excess kurtosis of the integrated profile (υπέρβαση κύρτωσης του ολοκληρωμένου προφίλ):** Μετρά την "αιχμηρότητα" της κατανομής του ολοκληρωμένου προφίλ σε σχέση με την κανονική κατανομή.
   - Θετική τιμή: το σήμα παρουσιάζει πιο αιχμηρές κορυφές.
   - Αρνητική τιμή: το σήμα είναι πιο "απλωμένο" με λιγότερο έντονες κορυφές.
4. **Skewness of the integrated profile (παραμόρφωση του ολοκληρωμένου προφίλ):** Μετρά τη συμμετρία της κατανομής του ολοκληρωμένου προφίλ γύρω από τη μέση τιμή.
   - Θετική τιμή: περισσότερες τιμές στα αριστερά της μέσης τιμής.
   - Αρνητική τιμή: περισσότερες τιμές στα δεξιά της μέσης τιμής.

Οι υπόλοιπες 4 μεταβλητές αντιστοιχούν σε στατιστικές ποσότητες που αντλούνται από την καμπύλη **DM-SNR** (Dispersion Measure – Signal to Noise Ratio), η οποία αναπαριστά τη συσχέτιση μεταξύ του μέτρου διασποράς (DM) και του λόγου σήματος προς θόρυβο (SNR) του σήματος. Οι μεταβλητές αυτές είναι:

5. **Mean of the DM-SNR curve (μέση τιμή της καμπύλης DM-SNR):** Υπολογίζει τη μέση τιμή της καμπύλης DM-SNR, η οποία αντιπροσωπεύει τη συνολική ισχύ του σήματος σε σχέση με το θόρυβο.

6. **Standard deviation of the DM-SNR curve (τυπική απόκλιση της καμπύλης DM-SNR):** Μετρά τη διασπορά των τιμών γύρω από τη μέση τιμή της καμπύλης DM-SNR και αποτελεί το μέτρο της διακύμανσης της ισχύος του σήματος.
7. **Excess kurtosis of the DM-SNR curve (υπέρβαση κύρτωσης της καμπύλης DM-SNR):** Μετρά την "αιχμηρότητα" της κατανομής των τιμών της καμπύλης DM-SNR σε σχέση την κανονική κατανομή.
    - Θετική τιμή: η καμπύλη έχει έντονες κορυφές.
    - Αρνητική τιμή: η καμπύλη είναι πιο επίπεδη.
8. **Skewness of the DM-SNR curve (παραμόρφωση της καμπύλης DM-SNR):** Μετρά τη συμμετρία της κατανομής των τιμών της καμπύλης DM-SNR.
    - Θετική τιμή: ασυμμετρία προς τις υψηλότερες τιμές.
    - Αρνητική τιμή: ασυμμετρία προς τις χαμηλότερες τιμές.

Η τελευταία στήλη όπως προαναφέρθηκε αντιστοιχεί στην κλάση. Αν η εγγραφή αντιστοιχεί σε πραγματική εκπομπή pulsar λαμβάνει στη στήλη αυτή την τιμή **1**, ενώ αν πρόκειται για ραδιο-παρεμβολές (RFI), άλλου είδους εκπομπές ή θόρυβο λαμβάνει την τιμή **0**.

Στατιστικοί έλεγχοι & προ-επεξεργασία δεδομένων

Πριν την εκπαίδευση των αλγορίθμων, με χρήση κυρίως των βιβλιοθηκών pandas, matplotlib και seaborn της Python, πραγματοποιήθηκαν κάποιοι βασικοί έλεγχοι της φύσης του dataset, οι οποίοι επέτρεψαν να εκτιμηθούν με ασφάλεια οι απαιτήσεις για περαιτέρω ενέργειες προ-επεξεργασίας των δεδομένων.

1. Επισκόπηση των πρώτων γραμμών και εμφάνιση γενικών πληροφοριών

Με τη χρήση της μεθόδου .read_csv της pandas δημιουργείται το κατάλληλο dataframe και στη συνέχεια με τις .head() και .info() εμφανίζονται οι πρώτες γραμμές του csv (Εικόνα 1) και κάποιες συνοπτικές πληροφορίες σχετικά με τη δομή του dataframe (Εικόνα 2).

	Profile_mean	Profile_stdev	Profile_skewness	Profile_kurtosis	DM_mean	DM_stdev	DM_skewness	DM_kurtosis	Class
0	140.562500	55.683782	-0.234571	-0.699648	3.199833	19.110426	7.975532	74.242225	0
1	102.507812	58.882430	0.465318	-0.515088	1.677258	14.860146	10.576487	127.393580	0
2	103.015625	39.341649	0.323328	1.051164	3.121237	21.744669	7.735822	63.171909	0
3	136.750000	57.178449	-0.068415	-0.636238	3.642977	20.959280	6.896499	53.593661	0
4	88.726562	40.672225	0.600866	1.123492	1.178930	11.468720	14.269573	252.567306	0

Εικόνα 1. Οι πέντε πρώτες γραμμές του DataFrame

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17898 entries, 0 to 17897
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Profile_mean     17898 non-null  float64
 1   Profile_stdev    17898 non-null  float64
 2   Profile_skewness 17898 non-null  float64
 3   Profile_kurtosis 17898 non-null  float64
 4   DM_mean          17898 non-null  float64
 5   DM_stdev         17898 non-null  float64
 6   DM_skewness      17898 non-null  float64
 7   DM_kurtosis      17898 non-null  float64
 8   Class            17898 non-null  int64  
dtypes: float64(8), int64(1)
memory usage: 1.2 MB
```

Εικόνα 2. Συνοπτικές πληροφορίες της δομής του DataFrame

Η pandas επιβεβαιώνει πως πρόκειται για ένα dataset 17898 εγγραφών, του οποίου οι εννέα στήλες αντιστοιχούν στα οκτώ γνωρίσματα του ηλεκτρομαγνητικού σήματος και στην κλάση (target variable). Όλες οι στήλες είναι πλήρης (non-null) κάτι που υποδεικνύει ότι δεν υπάρχουν ελλείπουσες τιμές και επίσης περιέχουν μόνο αριθμητικά δεδομένα. Οι στήλες των κατηγορημάτων περιέχουν συνεχή αριθμητικά δεδομένα δεκαδικής μορφής (float64), ενώ η κλάση ακέραιους αριθμούς (int64), δηλαδή τις τιμές 1 και 0 που υποδεικνύουν αν η εγγραφή αντιστοιχεί (ή όχι) σε σήμα pulsar.

2. Έλεγχος ελλειπουσών τιμών (missing values) και διπλότυπων (duplicate rows)

Με εφαρμογή των μεθόδων .isnull() και .duplicated() επιβεβαιώνεται ότι δεν υφίστανται ελλείπουσες τιμές, καθώς επίσης και ότι δεν υπάρχουν διπλοεγγραφές (Εικόνα 3). Η καθαρή και εξαιρετικά δομημένη φύση του dataset συνάδει πλήρως με το γεγονός ότι οι εγγραφές του αποτελούν προϊόντα παρατήρησης οργάνων εξαιρετικά υψηλής ακρίβειας.

```
Ελλείπουσες τιμές ανά χαρακτηριστικό:
Profile_mean        0
Profile_stdev       0
Profile_skewness    0
Profile_kurtosis    0
DM_mean             0
DM_stdev            0
DM_skewness         0
DM_kurtosis         0
Class               0
dtype: int64
Πλήθος διπλότυπων: 0
```

Εικόνα 3. Ελλείπουσες τιμές και διπλότυπα

3. Συνοπτική στατιστική περιγραφή

Η μέθοδος .describe() εμφανίζει τη βασική στατιστική περιγραφή του dataset, η οποία περιέχει σημαντική πληροφορία για τη στατιστική συμπεριφορά των οκτώ μεταβλητών του συνόλου δεδομένων, επιστρέφοντας τη μέση τιμή, την τυπική απόκλιση, τη διάμεσο, τη μέγιστη και ελάχιστη τιμή, καθώς και ενδοτεταρτημοριακό τους εύρος (Εικόνα 4).

Βασική Στατιστική DataSet:

	Profile_mean	Profile_stdev	Profile_skewness	Profile_kurtosis	DM_mean	DM_stdev	DM_skewness	DM_kurtosis	Class
count	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000	17898.000000
mean	111.079968	46.549532	0.477857	1.770279	12.614400	26.326515	8.303556	104.857709	0.091574
std	25.652935	6.843189	1.064040	6.167913	29.472897	19.470572	4.506092	106.514540	0.288432
min	5.812500	24.772042	-1.876011	-1.791886	0.213211	7.370432	-3.139270	-1.976976	0.000000
25%	100.929688	42.376018	0.027098	-0.188572	1.923077	14.437332	5.781506	34.960504	0.000000
50%	115.078125	46.947479	0.223240	0.198710	2.801839	18.461316	8.433515	83.064556	0.000000
75%	127.085938	51.023202	0.473325	0.927783	5.464256	28.428104	10.702959	139.309330	0.000000
max	192.617188	98.778911	8.069522	68.101622	223.392141	110.642211	34.539844	1191.000837	1.000000

Εικόνα 4. Συνοπτική στατιστική περιγραφή του DataSet

Τα κεντρικά σημεία της παρεχόμενης πληροφορίας για καθένα από τα γνωρίσματα του dataset είναι τα ακόλουθα:

- Το χαρακτηριστικό **Profile_mean** εμφανίζει μέσο όρο 111.08 και τυπική απόκλιση 25.65, με τιμές που κυμαίνονται από 5.81 έως 192.62, γεγονός που υποδεικνύει σημαντική διασπορά και παρουσία ακραίων τιμών (outliers).

- Το **Profile_stdev**, με μέση τιμή 46.55 και χαμηλότερη διασπορά (τυπική απόκλιση 6.84), κυμαίνεται από 24.77 έως 98.78, ακολουθώντας μάλλον πιο συμπαγή κατανομή.

- Η **Profile_skewness** έχει μέση τιμή 0.47 και τυπική απόκλιση 1.06, εμφανίζοντας μια ελαφρά δεξιά ασυμμετρία, με ελάχιστη τιμή -1.88 και μέγιστη 8.07.

- Η **Profile_kurtosis** με μέσο όρο 1.77 και τυπική απόκλιση 6.17 παρουσιάζει σημαντική μεταβλητότητα, φθάνοντας μέχρι και τις 68.10 μονάδες, στοιχείο που ενισχύει την υπόθεση έντονης παρουσίας ακραίων παρατηρήσεων.

Οι τέσσερις επιπλέον μεταβλητές που σχετίζονται με την διασπορά μέτρησης (DM-SNR) χαρακτηρίζονται από έντονη ασυμμετρία, καθώς και από την ισχυρή πιθανότητα παρουσίας εξαιρετικά ακραίων τιμών. Συγκεκριμένα:

- Η **DM_mean**, η οποία αποτυπώνει τον μέσο όρο του DM-SNR καναλιού, έχει πολύ μικρή διάμεσο (2.80) και τιμή πρώτου τεταρτημόριου (1.92), ενώ η μέγιστη τιμή φτάνει τα 223.39, στοιχείο που υποδηλώνει την παρουσία μεγάλης δεξιάς ουράς και την εμφάνιση πολλών

εξαιρέσεων. Η τυπική της απόκλιση (29.47) είναι πολλαπλάσια της μέσης τιμής, ενισχύοντας την ένδειξη έντονης μεταβλητότητας.

- Η **DM_stdev** που καταγράφει τη διασπορά της DM-SNR μέτρησης, κυμαίνεται από 7.37 έως 110.64, με μέσο όρο 26.33, γεγονός που αναδεικνύει ότι οι περισσότερες παρατηρήσεις καταλαμβάνουν την περιοχή των χαμηλότερων τιμών, ενώ είναι βέβαιη και εδώ η ύπαρξη ακραίων τιμών.

- Η **DM_skewness** (μέση τιμή 8.30 και τυπική απόκλιση 4.56) και η **DM_kurtosis** (μέση τιμή 104.86 και τυπική απόκλιση 106.51) ξεχωρίζουν για τις εξαιρετικά υψηλές τιμές τους. Η skewness κυμαίνεται από -3.13 έως 34.53, και η kurtosis από -1.97 έως 1191.00, κάτι που υποδηλώνει πολύ μακριές δεξιές ουρές και βαριές κατανομές. Πρόκειται για χαρακτηριστικά που συνδέονται με την ισχυρή παρουσία αιχμηρών κορυφών (peaks) και με την ύπαρξη εξαιρετικά απόμακρων τιμών.

- Τέλος, η **Class** (κλάση), που αντιπροσωπεύει τη δυαδική μεταβλητή στόχο (0: θόρυβος, 1: pulsar), εμφανίζει σαφή ανισορροπία, καθώς μόνο το 9.16% των παρατηρήσεων ανήκουν στη θετική εκδοχή της (σήμα pulsar).

Συνολικά, η στατιστική περιγραφή φανερώνει ένα σύνολο δεδομένων με σημαντικές ασυμμετρίες, ισχυρή παρουσία ακραίων τιμών (outliers) και έντονη ανισορροπία κλάσης.

4. Εντοπισμός outliers βάσει του κριτηρίου IQR

Η αναζήτηση ακραίων τιμών (outliers) με βάση το κριτήριο IQR (Interquartile Range) ανέδειξε σημαντικό αριθμό εξαιρέσεων σε όλα τα γνωρίσματα του συνόλου δεδομένων (Εικόνα 5), γεγονός που συνάδει πλήρως με τη στατιστική εικόνα ασυμμετρίας και υψηλής διακύμανσης που περιεγράφηκε στα προηγούμενα.

```
Πλήθος outliers ανά χαρακτηριστικό:
Profile_mean        1030
Profile_stdev        262
Profile_skewness    1596
Profile_kurtosis    1901
DM_mean             2927
DM_stdev            2346
DM_skewness          487
DM_kurtosis          901
dtype: int64
```

Εικόνα 5. Ακραίες τιμές (outliers)

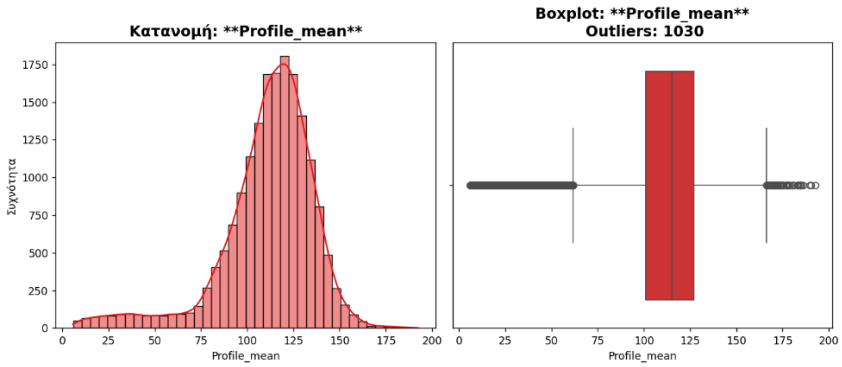
Συγκεκριμένα, το **DM_mean** παρουσιάζει τον μεγαλύτερο αριθμό outliers, με 2927 παρατηρήσεις να βρίσκονται εκτός του αποδεκτού εύρους τιμών. Ακολουθεί το **DM_stdev** με 2346 outliers, επιβεβαιώνοντας τη σημαντική διασπορά και την ύπαρξη σημείων με εξαιρετικά υψηλή ή χαμηλή διακύμανση. Παρόμοια είναι η εικόνα που παρατηρείται και στα **Profile_kurtosis** με 1901 εξαιρέσεις και στο **Profile_skewness** με 1596, υποδηλώνοντας έντονη απόκλιση από την κανονικότητα στις παρατηρήσεις που σχετίζονται με τη μορφή και την ασυμμετρία της κατανομής των Profile χαρακτηριστικών. Το **Profile_mean** από την άλλη μεριά καταγράφει 1030 outliers, ένδειξη σημαντικής διασποράς και στις μέσες τιμές του προφίλ, ενώ το **DM_kurtosis** εμφανίζει 901 ακραίες τιμές, ενισχύοντας την εικόνα "βαριάς ουράς" που είχε αναδειχθεί και στη στατιστική περιγραφή. Αξιοσημείωτο είναι ότι το

DM_skewness, παρά την πολύ υψηλή τιμή της μέσης και μέγιστης ασυμμετρίας, εμφανίζει σχετικά περιορισμένο αριθμό εξαιρέσεων (487 outliers). Τέλος, το **Profile_stdev** με 262 outliers, είναι το γνώρισμα με τις λιγότερες ακραίες τιμές, ενδεχομένως λόγω της χαμηλής διασποράς του σε σχέση με τα υπόλοιπα.

5. Διάγραμμα κατανομής συχνότητας και boxplot ανά χαρακτηριστικό

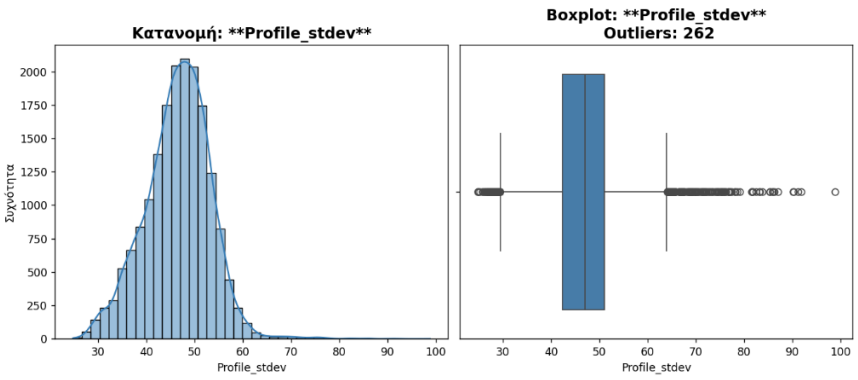
Με χρήση των δυνατοτήτων των βιβλιοθηκών seaborn και matplotlib κατασκευάστηκαν, για κάθε ανεξάρτητη μεταβλητή του dataset, τα διαγράμματα κατανομής συχνοτήτων και τα αντίστοιχα boxplots, τα οποία επιτρέπουν μια αρκετά ακριβή οπτικοποίηση της στατιστικής φύσης του κάθε γνωρίσματος. Κάθε ιστόγραμμα συχνοτήτων συνοδεύεται και από την KDE (Kernel Density Estimate) καμπύλη, η οποία προσφέρει μια πιο ομαλή εκδοχή της εικόνας της κατανομής, ακόμη πιο συμβατή με τη συνεχή υφή των υπό μελέτη μεταβλητών. Τα κεντρικά σημεία της εξαγόμενης πληροφορίας από τις συγκεκριμένες οπτικοποιήσεις είναι:

Η μεταβλητή ***Profile_mean*** παρουσιάζει σχετικά κανονική κατανομή, με τιμές που κυμαίνονται από 5.81 έως 192.62 και μέσο όρο 111.08. Το boxplot επαληθεύει την ύπαρξη αρκετών outliers και στα δύο άκρα της κατανομής. Ωστόσο, η μικρή αρνητική ασυμμετρία δεν αλλοιώνει τη γενικότερη εικόνα ισορροπίας των δεδομένων (Εικόνα 6).



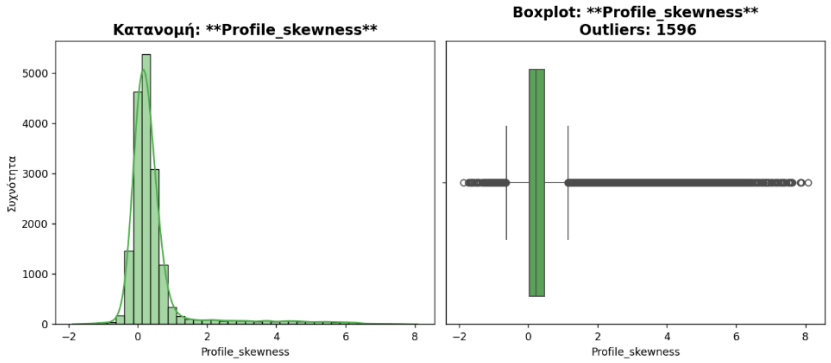
Εικόνα 6. Profile_mean

Η κατανομή της ***Profile_stdev*** εμφανίζεται αρκετά συμμετρική με μέσο όρο 46.55 και ελάχιστη διακύμανση. Η πλειοψηφία των τιμών εντοπίζεται στην περιοχή 42 έως 51 ενώ το boxplot εμφανίζει μικρό αριθμό outliers. Αποτυπώνεται έτσι η περιορισμένη διασπορά και η σταθερή συμπεριφορά της συγκεκριμένης μεταβλητής (Εικόνα 7).



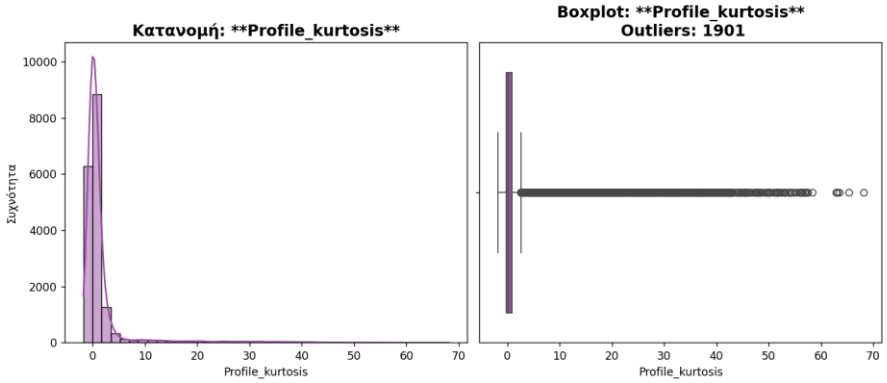
Εικόνα 7. Profile_stdev

Η μεταβλητή ***Profile_skewness*** χαρακτηρίζεται από θετική ασυμμετρία, με αρκετές τιμές να συγκεντρώνονται κοντά στο μηδέν. Το boxplot αποκαλύπτει την ύπαρξη πλήθους outliers στα υψηλότερα επίπεδα. Ενώ η παρουσία και αρνητικών τιμών ενισχύει περαιτέρω την εικόνα ανομοιομορφίας (Εικόνα 8).



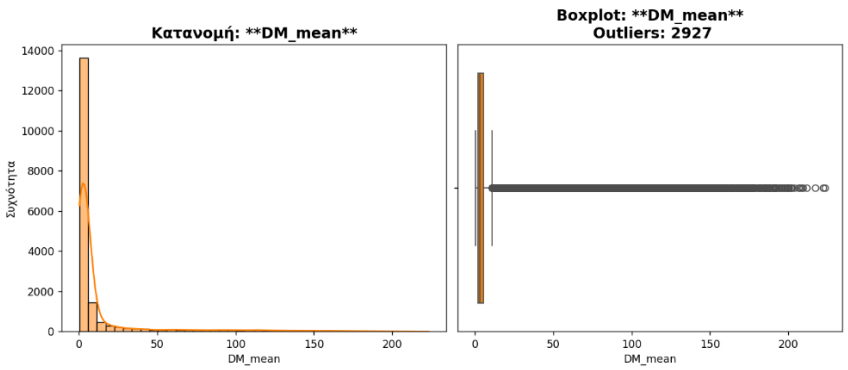
Εικόνα 8. Profile_skewness

Η ***Profile_kurtosis*** εμφανίζει μεγάλη διασπορά και σημαντική θετική ασυμμετρία, με μέγιστη τιμή 68.10. Η διάμεσος πλησιάζει το μηδέν, ενώ το boxplot εμφανίζει πολυάριθμους outliers στο άνω άκρο. Παρά το χαμηλή μέση τιμή (1.77), η υψηλή τυπική απόκλιση μαρτυρά έντονη ετερογένεια στις μορφές της κατανομής των προφίλ.



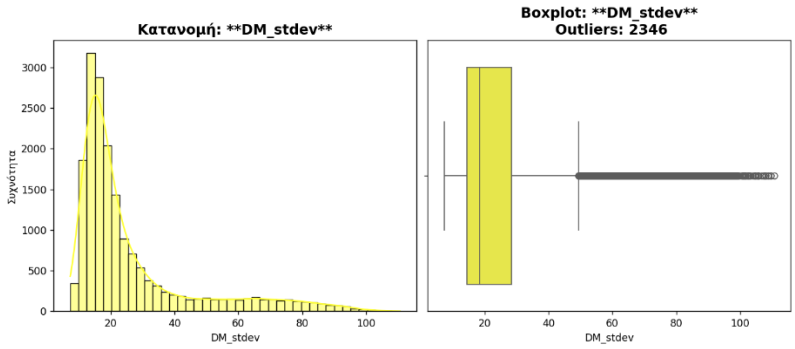
Εικόνα 9. Profile_kurtosis

Η κατανομή του ***DM_mean*** εμφανίζει έντονη δεξιά ασυμμετρία. Οι περισσότερες τιμές βρίσκονται κάτω του 10, ενώ ακραίες παρατηρήσεις φτάνουν έως 223.39, όπως δείχνει και το boxplot. Η διακύμανση είναι πολύ μεγάλη, όπως και το πλήθος των outliers (Εικόνα 10).



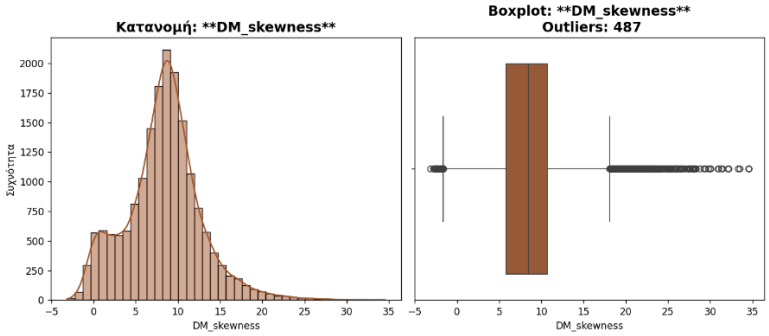
Εικόνα 10. DM_mean

Η μεταβλητή ***DM_stdev*** εμφανίζει μεγάλη μεταβλητότητα και ισχυρή θετική ασυμμετρία. Η μέση τιμή (26.32) και η διάμεσος (18.46) διαφέρουν σημαντικά, ενώ οι τιμές κυμαίνονται από 7.37 έως 110.64. Το boxplot αποκαλύπτει αρκετούς outliers, κυρίως προς το άνω άκρο, σκιαγραφώντας μια εικόνα ετερογένειας στις αποκλίσεις εντός του DM σήματος (Εικόνα 11).



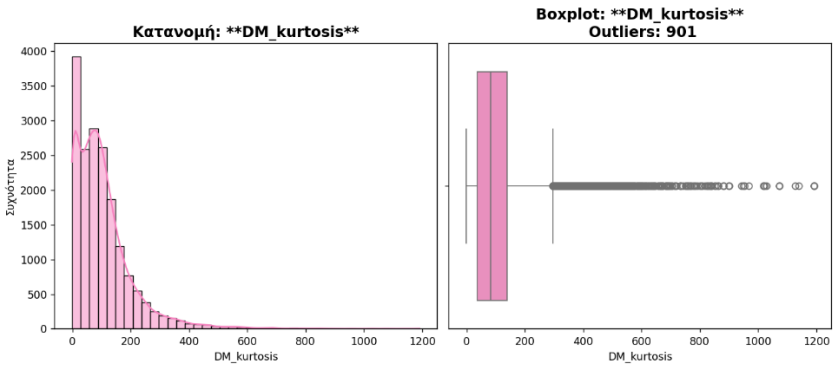
Εικόνα 11. DM_stdev

Η ***DM_skewness*** χαρακτηρίζεται από πολύ μεγάλη θετική ασυμμετρία, με τιμές από -3.13 έως 34.54. Παρότι η διάμεσος (8.43) και ο μέσος όρος (8.30) είναι κοντά, η τυπική απόκλιση είναι υψηλή, γεγονός που αποδίδεται στην παρουσία εξαιρετικά υψηλών outliers, η οποία αποτυπώνεται με ξεκάθαρο τρόπο στο boxplot (Εικόνα 12).



Εικόνα 12. DM_skewness

Η ***DM_kurtosis*** είναι η πιο "εκρηκτική" μεταβλητή του συνόλου, με μέγιστη τιμή που φτάνει το 1191.00. Η μέση τιμή ξεπερνά το 104 ενώ η διάμεσος είναι μόλις 8.06. Η θετική ασυμμετρία και τεράστιο πλήθος outliers που αποτυπώνεται στο boxplot συναρτώνται με μεγάλες αποκλίσεις από την τυπική συμπεριφορά και δεικνύουν σοβαρή παραμόρφωση της κατανομής (Εικόνα 13).



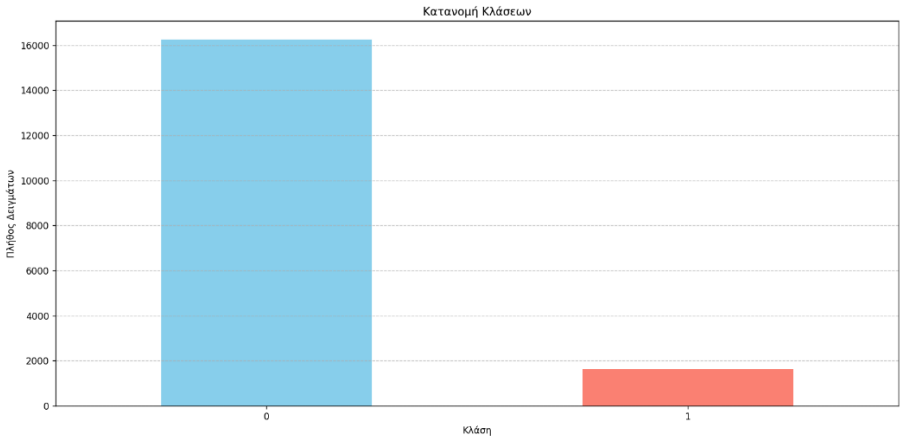
Εικόνα 13. DM_kurtosis

6. Έλεγχος ανισορροπίας κλάσης

Η κατανομή των κλάσεων στο εξεταζόμενο σύνολο δεδομένων παρουσιάζει σημαντική ανισορροπία, όπως αντανακλάται και στο αποτέλεσμα που επιστρέφει η pandas (Εικόνα 14) και απεικονίζεται και στο αντίστοιχο διάγραμμα (Εικόνα 15). Συγκεκριμένα, η κλάση 0 που αντιστοιχεί στα σήματα θορύβου περιλαμβάνει 16.259 δείγματα, ενώ η κλάση 1, που αφορά σήματα εκπομπών pulsars, μόλις 1.639 δείγματα. Αυτό συνεπάγεται ότι περίπου το 91% των παρατηρήσεων ανήκει στην κλάση 0 και μόλις το 9% στην κλάση 1.

```
Κατανομή Κλάσεων:
Κλάση 0: 16259 δείγματα
Κλάση 1: 1639 δείγματα
```

Εικόνα 14. Ανισορροπία κλάσης



Εικόνα 15. Διάγραμμα ανισορροπίας κλάσης

Το φαινόμενο αυτό μπορεί να οδηγήσει σε προσανατολισμό των ταξινομητών προς την πλειοψηφούσα κλάση κατά την εκπαίδευση των μοντέλων, γεγονός που οφείλει να ληφθεί σοβαρά υπόψιν και να αναζητηθούν αξιόπιστοι τρόποι αντιμετώπισής του, ιδιαίτερα στο πλαίσιο της παρούσας μελέτης που το ενδιαφέρον επικεντρώνεται στην ανίχνευση σπάνιων συμβάντων (εκπομπές pulsars).

7. Έλεγχος κατά Pearson συσχέτισης και δημιουργία Heatmap

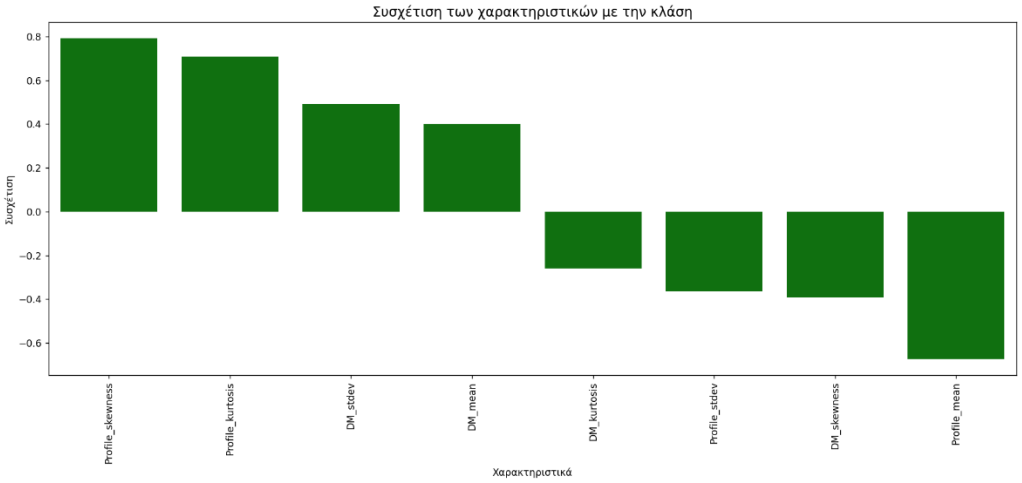
Ο συντελεστής συσχέτισης Pearson αποτελεί ένα στατιστικό μέτρο που αποτυπώνει τον βαθμό και την κατεύθυνση της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Λαμβάνει τιμές στο διάστημα [-1, 1], όπου τιμές κοντά στο 1 υποδηλώνουν ισχυρή θετική συσχέτιση, τιμές κοντά στο -1 ισχυρή αρνητική συσχέτιση, ενώ τιμές κοντά στο 0 υποδεικνύουν απουσία γραμμικής σχέσης. Ο υπολογισμός της συγκεκριμένης στατιστικής ποσότητας επιτρέπει την κατανόηση των συσχετίσεων τόσο μεταξύ των γνωρίσματων του δείγματος όσο και μεταξύ κάθε γνωρίσματος και του στόχου πρόβλεψης, καθοδηγώντας έτσι τη διαδικασία επιλογής μεταβλητών και αυξάνοντας την κατανόηση της συμπεριφοράς του υπό μελέτη φαινομένου.

Στην παρούσα έρευνα, αρχικά υπολογίστηκε ο συντελεστής συσχέτισης Pearson μεταξύ κάθε χαρακτηριστικού των δεδομένων και της μεταβλητής στόχου (κλάσης) (Εικόνα 16), η οποία συνοδεύτηκε και από το αντίστοιχο διάγραμμα για την καλύτερη εποπτική κατανόηση της συμπεριφοράς κάθε μεταβλητής (Εικόνα 17). Τα αποτελέσματα υποδεικνύουν ότι τα χαρακτηριστικά Profile_skewness (0.791591) και Profile_kurtosis (0.709528) εμφανίζουν ισχυρή θετική συσχέτιση με την κλάση, γεγονός που σημαίνει ότι οι υψηλότερες τιμές τους σχετίζονται με μεγαλύτερη πιθανότητα ανήκειν στη θετική κλάση. Ενδιαφέρον παρουσιάζει

και η μέτρια θετική συσχέτιση των DM_stdev (0.491535) και DM_mean (0.400876), τα οποία επίσης φαίνεται να συνεισφέρουν στην πρόβλεψη. Από την άλλη πλευρά, χαρακτηριστικά όπως το Profile_mean (-0.673181), το DM_skewness (-0.390816) αλλά και το Profile_stdev (-0.363708) εμφανίζουν αρνητική συσχέτιση με την κλάση, πράγμα που σημαίνει ότι όσο μειώνονται οι τιμές τους τόσο αυξάνεται η πιθανότητα του ανήκειν στη θετική κλάση. Ενώ ασθενέστερη αρνητική συσχέτιση (-0.259171) με την κλάση παρουσιάζει το γνώρισμα η DM_kurtosis, του οποίου η στατιστική βαρύτητα μοιάζει λιγότερο σημαντική αναφορικά με την πρόβλεψη.

```
Πίνακας κατά Pearson συσχέτισης:
Profile_skewness    0.791591
Profile_kurtosis    0.709528
DM_stdev            0.491535
DM_mean             0.400876
DM_kurtosis        -0.259117
Profile_stdev      -0.363708
DM_skewness        -0.390816
Profile_mean       -0.673181
dtype: float64
```

Εικόνα 16. Συσχέτιση Pearson γνωρίσματος-κλάσης



Εικόνα 17. Ιστόγραμμα συσχέτισης Pearson γνωρίσματος-κλάσης

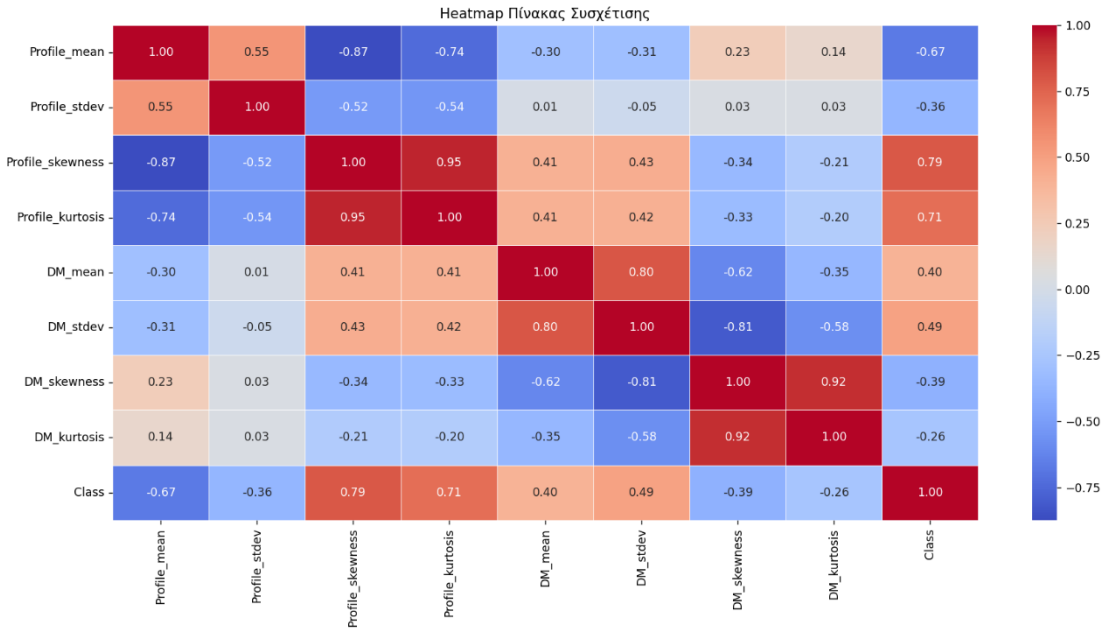
Στη συνέχεια αφού υπολογίστηκαν οι συντελεστές Pearson για όλους τους δυνατούς συνδυασμούς των μεταβλητών του dataset (Εικόνα 18), με χρήση των δυνατοτήτων που προσφέρουν οι βιβλιοθήκες seaborn και matplotlib της Python κατασκευάστηκε ο Heatmap πίνακας συσχέτισης (Εικόνα 19).

```
Πίνακας ολικής συσχέτισης:
                 Profile_mean  Profile_stdev  Profile_skewness  Profile_kurtosis   DM_mean   DM_stdev  DM_skewness  DM_kurtosis     Class
Profile_mean         1.000000       0.547137         -0.873898         -0.738775 -0.298841 -0.307016     0.234331     0.144033 -0.673181
Profile_stdev        0.547137       1.000000         -0.521435         -0.539793  0.006869 -0.047632     0.029429     0.027691 -0.363708
Profile_skewness    -0.873898      -0.521435          1.000000          0.945729  0.414368  0.432880    -0.341209    -0.214491  0.791591
Profile_kurtosis    -0.738775      -0.539793          0.945729          1.000000  0.412056  0.415140    -0.328843    -0.204782  0.709528
DM_mean             -0.298841       0.006869          0.414368          0.412056  1.000000  0.796555    -0.615971    -0.354269  0.400876
DM_stdev            -0.307016      -0.047632          0.432880          0.415140  0.796555  1.000000    -0.809786    -0.575800  0.491535
DM_skewness          0.234331       0.029429         -0.341209         -0.328843 -0.615971 -0.809786     1.000000     0.923743 -0.390816
DM_kurtosis          0.144033       0.027691         -0.214491         -0.204782 -0.354269 -0.575800     0.923743     1.000000 -0.259117
Class               -0.673181      -0.363708          0.791591          0.709528  0.400876  0.491535    -0.390816    -0.259117  1.000000
```

Εικόνα 18. Συντελεστές Pearson για όλα τα ζεύγη των γνωρίσματων του DataSet

Ο Heatmap πίνακας συνιστά ένα ισχυρό εργαλείο οπτικοποίησης για την εξερεύνηση της σχέσης μεταξύ πολλών μεταβλητών σε ένα σύνολο δεδομένων. Στην ουσία πρόκειται για την γραφική αποτύπωση των συντελεστών συσχέτισης Pearson με τη χρήση κλίμακας χρωμάτων: τα θερμότερα χρώματα (κόκκινο) υποδηλώνουν ισχυρή θετική συσχέτιση, ενώ τα πιο ψυχρά χρώματα (μπλε) ισχυρή αρνητική συσχέτιση. Με τον τρόπο αυτό διευκολύνεται η γρήγορη

αναγνώριση συσχετίσεων υψηλού ή χαμηλού βαθμού που με τη σειρά της συμβάλλει στον εντοπισμό προτύπων και σχέσεων μεταξύ των μεταβλητών.



Εικόνα 19. Heatmap πίνακας συσχέτισης

Ορισμένα ασφαλή ευρήματα που προκύπτουν από τον Heatmap πίνακα του υπό μελέτη dataset, τα οποία αφορούν στη φύση των σχέσεων μεταξύ των μεταβλητών και δύνανται να αναδείξουν κατευθύνσεις σωστής προετοιμασίας των δεδομένων είναι τα ακόλουθα: Αρχικά, εμφανίζεται εξαιρετικά υψηλή θετική συσχέτιση μεταξύ των μεταβλητών Profile_skewness και Profile_kurtosis (0.95), υποδεικνύοντας ότι τα δύο αυτά γνωρίσματα κινούνται σχεδόν απόλυτα μαζί. Ομοίως, τα DM_skewness και DM_kurtosis εμφανίζουν επίσης ισχυρή θετική συσχέτιση (0.92). Αυτές οι εξαιρετικά υψηλές τιμές υποδηλώνουν ότι τα ζεύγη αυτών των χαρακτηριστικών περιγράφουν παρόμοια στατιστικά μοτίβα και η ταυτόχρονη χρήση τους στην εκπαίδευση των μοντέλων μπορεί να εισάγει πολυσυγγραμμικότητα (multicollinearity), επηρεάζοντας αρνητικά τη σταθερότητα και τη γενίκευση των μοντέλων πρόβλεψης. Το χαρακτηριστικό Profile_mean εμφανίζει μέτρια θετική συσχέτιση με το Profile_stdev (0.55), γεγονός που υποδηλώνει ότι οι αυξήσεις στον μέσο όρο του προφίλ συνοδεύονται συνήθως (αλλά όχι πάντα) από αυξήσεις της διασποράς του. Επίσης, παρατηρείται μια ασθενέστερη θετική συσχέτιση μεταξύ DM_mean και DM_stdev (0.80).

Αξιοσημείωτες είναι κι οι αρνητικές συσχετίσεις που εντοπίζονται, μεταξύ Profile_mean και Profile_skewness (-0.87), αλλά και ανάμεσα σε Profile_mean και Profile_kurtosis (-0.74). Αυτές οι τιμές δείχνουν ότι όσο αυξάνεται ο μέσος όρος του προφίλ, τείνουν να μειώνονται τα επίπεδα ασυμμετρίας (incommensurability) και επιπεδότητας (flatness) στην κατανομή των δεδομένων. Αντίστοιχα, αν και σε μικρότερο βαθμό, παρατηρείται αρνητική συσχέτιση μεταξύ των Profile_stdev και Profile_skewness (-0.52), καθώς και μεταξύ Profile_stdev και Profile_kurtosis (-0.54), στοιχείο που δείχνει ότι μεγαλύτερη μεταβλητότητα σχετίζεται με λιγότερη ασυμμετρία και επιπεδότητα.

Στα χαρακτηριστικά που αφορούν τις μεταβλητές DM, η συσχέτιση μεταξύ DM_mean και DM_skewness είναι θετική (0.41), όπως και μεταξύ DM_mean και DM_kurtosis (0.41), αν και σε μέτριο επίπεδο. Οι συσχετίσεις του DM_stdev με το DM_skewness (0.43), καθώς και με το DM_kurtosis (0.42) δείχνουν επίσης ότι η μεταβλητότητα των τιμών DM συνδέεται μέτρια

θετικά με τη μορφή των αντίστοιχων κατανομών. Τέλος, μεταξύ χαρακτηριστικών προφίλ και των αντίστοιχων DM, οι συσχετίσεις είναι γενικά χαμηλές ή και αμελητέες, γεγονός που υποδηλώνει ότι τα δύο αυτά σύνολα χαρακτηριστικών καταγράφουν διαφορετικές όψεις της πληροφορίας, συμβάλλοντας συμπληρωματικά στη διαδικασία πρόβλεψης.

Τόσο ο υπολογισμός των συντελεστών Pearson όσο και η Heatmap οπτικοποίησή τους αποκαλύπτει μια πλούσια δομή σχέσεων μεταξύ των χαρακτηριστικών του υπό μελέτη dataset. Σε συνδυασμό και με τα αποτελέσματα των στατιστικών ελέγχων που προηγήθηκαν έγιναν οι εκτιμήσεις για κάποιες επιπλέον κινήσεις προεπεξεργασίας, ικανές να συμβάλλουν στην ενίσχυση της αποδοτικότητας των μοντέλων ταξινόμησης. Τα κεντρικότερα ζητήματα που αναδείχθηκαν αφορούν στην ύπαρξη περιοχών έντονης πλεονασματικότητας, απότοκης της ισχυρής συσχέτισης ανάμεσα σε ορισμένα από τα γνωρίσματα του συνόλου, καθώς και η έντονη ανισορροπία ανάμεσα στις κλάσεις.

8. Προεπεξεργασία δεδομένων

Δεδομένου ότι πτυχή της προβληματικής κατάστασης που εξετάζεται συνιστά κι η μελέτη του κατά πόσο τα γνωρίσματα του dataset θα μπορούσαν να τυποποιηθούν σε ένα καθολικό σύνολο χαρακτηριστικών για κάθε μελλοντική έρευνα αυτοματοποιημένης διαλογής άστρων νετρονίων, επιλέχθηκε να μην πραγματοποιηθεί κανενός είδους συγχώνευση γνωρισμάτων που αν και ενδεχομένως θα μείωνε τον κίνδυνο της πλεονασματικότητας, λόγω των ισχυρών συσχετίσεων, θα αύξανε τον σοβαρότερο κίνδυνο της απώλειας σημαντικής πληροφορίας. Οι κινήσεις προεπεξεργασίας που κρίθηκαν επιβεβλημένες αφορούν στην αντιμετώπιση των μεγάλων διαφορών στην κλίμακα των επιμέρους χαρακτηριστικών και στην εξομάλυνση της σημαντικής ανισορροπίας ανάμεσα στις δύο κλάσεις, η οποία είναι δυνατόν να προκαλέσει τη μεροληπτική ευαισθησία των μοντέλων υπέρ της πλειοψηφικής κλάσης.

Το πρόβλημα των διαφορών στην κλίμακα των χαρακτηριστικών αντιμετωπίστηκε μέσω της διαδικασίας της *τυποποίησης* (standardization). Πρόκειται για μια μαθηματική τεχνική προεπεξεργασίας δεδομένων κατά την οποία κάθε χαρακτηριστικό μετασχηματίζεται ώστε να έχει *μέση τιμή* (mean) 0 και *τυπική απόκλιση* (standard deviation) 1. Ο στόχος της είναι να τοποθετηθούν όλα τα χαρακτηριστικά στην ίδια κλίμακα, διατηρώντας όμως την κατανομή τους, ώστε αλγόριθμοι ευαίσθητοι σε διαφορετικές κλίμακες (κυρίως εκείνοι που βασίζονται σε αποστάσεις και γραμμικές σχέσεις όπως οι SVM, k-NN, Logistic Regression) να λειτουργούν αποδοτικότερα. Η υλοποίησή της πραγματοποιήθηκε με χρήση της μεθόδου *StandardScaler()* της βιβλιοθήκης scikit-learn της Python.

Από την άλλη πλευρά, ο κίνδυνος της μεροληπτικής συμπεριφοράς των μοντέλων υπέρ της πλειοψηφικής κλάσης, λόγω της έντονης ανισορροπίας, αντιμετωπίστηκε με εφαρμογή στο σύνολο εκπαίδευσης (training set) της τεχνικής SMOTE (Synthetic Minority Over-sampling Technique). Η εκτέλεση του SMOTE δημιουργεί *συνθετικά δείγματα* της μειοψηφικής κλάσης μέσω παρεμβολής μεταξύ των υπαρχόντων στον χώρο των χαρακτηριστικών. Με αυτόν τον τρόπο, το σύνολο εκπαίδευσης εξισορροπείται χωρίς απώλεια πληροφορίας, αυξάνοντας την ευαισθησία των μοντέλων στην ανίχνευση σπάνιων περιπτώσεων. Η τεχνική υλοποιήθηκε με χρήση των δυνατοτήτων που προσφέρει η βιβλιοθήκη imblearn της Python. Ωστόσο, επειδή η συγκεκριμένη τεχνική επηρεάζει τη γεωμετρία του χώρου των δεδομένων, επιλέχθηκε να μην εφαρμοστεί στους κατεξοχήν αλγόριθμους αποστάσεων, όπως η Logistic Regression και το SVM. Στα μοντέλα αυτά η διαχείριση της ανισορροπίας επιτεύχθηκε χρησιμοποιώντας την παράμετρο class_weight='balanced', κατά την εκπαίδευσή τους. Η ρύθμιση αυτή οδηγεί το μοντέλο να *σταθμίσει τις κλάσεις αντιστρόφως ανάλογα με τη συχνότητά τους*, ώστε τα

σπάνια δείγματα να μην παραβλέπονται. Ο συνδυασμός των τεχνικών αυτών συνέβαλλε καθοριστικά στη βελτίωση της ικανότητας γενίκευσης και στην *αύξηση της απόδοσης* των αλγορίθμων, περιορίζοντας σημαντικά τις συνέπειες της ανισορροπίας που χαρακτηρίζει τις κλάσεις του υπό μελέτη dataset.

9. Διαχωρισμός δεδομένων σε training και test set

Της εκπαίδευσης των μοντέλων προηγήθηκε η διαδικασία διαχωρισμού του συνόλου των δεδομένων σε *εκπαιδευτικό* (training) και *δοκιμαστικό* (test) υποσύνολο. Πρόκειται για ένα κρίσιμο βήμα για την αξιόπιστη αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης. Στο πλαίσιο της παρούσας μελέτης, εφαρμόστηκε η συνάρτηση train_test_split της βιβλιοθήκης scikit-learn με ορισμό της παραμέτρου test_size στην τιμή 0.3, ώστε το 30% των δειγμάτων, δηλαδή 5370 εγγραφές, να διατεθούν για τη φάση της δοκιμής (test set), ενώ το υπόλοιπο 70% (12528 εγγραφές) να χρησιμοποιηθεί ως training set (Εικόνα 20).

```
Μέγεθος Συνόλου Εκπαίδευσης: (12528, 8)
Μέγεθος Συνόλου Δοκιμής: (5370, 8)
```

Εικόνα 20. Σύνολο εκπαίδευσης και σύνολο δοκιμής

Επιπλέον, η παράμετρος stratify επιλέχθηκε στην τιμή y, η οποία εξασφαλίζει ότι η αναλογία των κλάσεων διατηρείται ίδια και στα δύο υποσύνολα διαχωρισμού, τακτική που καθίσταται επιβεβλημένη σε περιπτώσεις έντονης ανισορροπίας. Τέλος, η επιλογή της τιμής random_state=42 διασφαλίζει την *αναπαραγωγιμότητα* των αποτελεσμάτων, επιβάλλοντας τον ίδιο τυχαίο διαχωρισμό σε κάθε επαναλαμβανόμενη εκτέλεση του μοντέλου.

Εκπαίδευση αλγορίθμων ταξινόμησης (classification)

Κατηγορίες ταξινομητών

Οι αλγόριθμοι ταξινόμησης χωρίζονται σε διάφορες κατηγορίες ανάλογα με τη λογική και τη μεθοδολογία που ακολουθούν. Ορισμένες από τις βασικότερες είναι οι ακόλουθες:

- Στατιστικά Μοντέλα (Statistic Models)

Τα στατιστικά μοντέλα αποτελούν μία από τις παλαιότερες και με ισχυρή θεωρητική θεμελίωση προσεγγίσεις στο πεδίο της ταξινόμησης. Πρόκειται για μεθόδους, οι οποίες βασίζονται σε πιθανότητες και θεωρίες εκτίμησης παραμέτρων, με στόχο την πρόβλεψη της πιθανότητας μια παρατήρηση να ανήκει σε μια συγκεκριμένη κατηγορία. Ένα βασικό πλεονέκτημα των στατιστικών μοντέλων είναι η διαφάνεια και η δυνατότητα ερμηνείας των παραμέτρων τους, στοιχείο που τα καθιστά ιδιαίτερα χρήσιμα σε εφαρμογές όπου απαιτείται κατανόηση του τρόπου με τον οποίο τα δεδομένα οδηγούν σε συγκεκριμένες ταξινομήσεις. Χαρακτηριστικό παράδειγμα αυτής της κατηγορίας αποτελεί η *Λογιστική Παλινδρόμηση (Logistic Regression)*.

- Δέντρα Αποφάσεων (Tree-based Models)

Τα δέντρα απόφασης (*Decision Trees*) ανήκουν στους πιο ευέλικτους και εύχρηστους αλγόριθμους ταξινόμησης και βασίζονται στη λογική της σταδιακής διάσπασης του χώρου των χαρακτηριστικών σε διακριτές περιοχές. Δημιουργούν ιεραρχικές δομές, όπου ο κάθε εσωτερικός κόμβος αντιστοιχεί σε ένα κριτήριο διαχωρισμού ενώ τα φύλλα αντιστοιχούν στις τελικές κατηγορίες. Έχουν το πλεονέκτημα της ερμηνευσιμότητας και της ικανότητας χειρισμού μη γραμμικών συσχετίσεων. Μία από τις πιο ισχυρές επεκτάσεις τους είναι το *Random Forest*. Ένα σύνολο από πολλά δέντρα, τα οποία εκπαιδεύονται με διαφορετικά υποσύνολα των δεδομένων και των χαρακτηριστικών τους. Το τελικό αποτέλεσμα ανάγεται στην πλειοψηφούσα τιμή του συνόλου των επιμέρους αποφάσεων.

- Μέθοδοι Ενίσχυσης (Boosting Methods)

Οι μέθοδοι ενίσχυσης (boosting) συνιστούν ισχυρές τεχνικές συναθροιστικής μάθησης που στοχεύουν στη δημιουργία ενός ισχυρού ταξινομητή μέσω του συνδυασμού πολλών απλών και ασθενέστερων μοντέλων. Η βασική ιδέα έγκειται στην διαδοχική εκπαίδευση μοντέλων. Κάθε νέο μοντέλο εστιάζει στα σφάλματα των προηγούμενων, ενισχύοντας σταδιακά την ακρίβεια του συνολικού συστήματος. Μια αρκετά δημοφιλής υλοποίηση των μεθόδων αυτών είναι ο αλγόριθμος *XGBoost (Extreme Gradient Boosting)*, ο οποίος δημιουργεί δέντρα με διαδοχική ενίσχυση. Κάθε νέο δέντρο επιχειρεί να διορθώσει τα σφάλματα των προηγούμενων, βελτιστοποιώντας την απόδοση του μοντέλου μέσω της ελαχιστοποίησης μιας συναρτησιακής απώλειας.

- Μέθοδοι Υποχώρου και Περιθωρίων (Margin-based Methods)

Οι συγκεκριμένες μέθοδοι, με κυριότερο εκπρόσωπο τον αλγόριθμο Support Vector Machine (SVM), στηρίζονται στη γεωμετρική θεώρηση της ταξινόμησης. Ο βασικός στόχος εδώ είναι να εντοπιστεί το *υπερεπίπεδο που διαχωρίζει τις κατηγορίες με το μέγιστο δυνατό περιθώριο*, δηλαδή τη μεγαλύτερη δυνατή απόσταση από τα πλησιέστερα σημεία κάθε κατηγορίας (support vectors). Η προσέγγιση αυτή εξασφαλίζει

υψηλή γενίκευση στα δεδομένα, ενώ με τη χρήση κατάλληλων *πυρηνικών συναρτήσεων (kernels)* μπορεί να επεκταθεί και σε περιπτώσεις μη γραμμικά διαχωρίσιμων συνόλων. Πρόκειται στην ουσία για ισχυρή επέκταση των γεωμετρικών μεθόδων προσδιορισμού αποστάσεων στον χώρο των δεδομένων, στις οποίες βασίζονται αρκετά παραδοσιακά μοντέλα ταξινόμησης όπως ο k-Nearest Neighbors (k-NN). Το βασικό τους μειονέκτημα εντοπίζεται στο υπολογιστικό τους κόστος για πολύ μεγάλα σύνολα δεδομένων και στη δυσκολία ερμηνείας των μοντέλων σε μη γραμμικές περιπτώσεις.

- Μέθοδοι Ασαφούς Λογικής (Fuzzy Methods)

Εδώ δεν πρόκειται για μία κλασική κατηγορία μοντέλων ταξινόμησης. Οι αλγόριθμοι που ενσωματώνουν ασαφή λογική (fuzzy logic) μπορούν να διαχειριστούν καλύτερα από τους αντίστοιχους κλασικούς τους γεννήτορες περιπτώσεις ταξινόμησης όπου τα δεδομένα δεν ανήκουν αυστηρά σε μία μόνο κατηγορία, αλλά χαρακτηρίζονται από αβεβαιότητα. Σε αυτού του είδους την ταξινόμηση η ένταξη ενός δείγματος σε μια κατηγορία δεν είναι απόλυτη αλλά *βαθμιαία*, με τη μορφή βαθμών συμμετοχής (membership degrees). Δύο χαρακτηριστικοί αλγόριθμοι αυτού του είδους, που χρησιμοποιήθηκαν και στο πλαίσιο της παρούσας μελέτης, είναι ο *Fuzzy k-Nearest Neighbors (Fuzzy k-NN)* που επεκτείνει τον παραδοσιακό αλγόριθμο των πλησιέστερων γειτόνων (k-NN) και τα *Fuzzy Decision Trees* που ενσωματώνουν ασαφή κριτήρια στους κόμβους των παραδοσιακών δέντρων απόφασης, επιτρέποντας πιο ευέλικτες διακλαδώσεις και διαχείριση θορυβωδών ή αβέβαιων δεδομένων. Τέτοιες μέθοδοι είναι ιδιαίτερα ευέλικτες και μπορούν σχετικά εύκολα να ενσωματωθούν σε υβριδικά συστήματα, ωστόσο απαιτούν προσεκτική σχεδίαση των συναρτήσεων συμμετοχής και είναι συχνά υπολογιστικά απαιτητικοί.

Μετρικές και οπτικοποιήσεις αξιολόγησης μοντέλων

Η αξιολόγηση της απόδοσης των μοντέλων που εξετάζονται στην παρούσα μελέτη έγινε στη βάση ορισμένων τυπικών μετρικών και οπτικοποιήσεων (ROC Curve και Precision-Recall Curve) ικανών να παράσχουν ασφαλή συμπεράσματα για την ταξινομητική και προβλεπτική ικανότητά τους. Οι εν λόγω αξιολογικοί δείκτες αντλούν κρίσιμη πληροφορία από τις τιμές της *μήτρας σύγχυσης* (*confusion matrix*) του εκάστοτε αλγορίθμου.

- Μήτρα σύγχυσης

Η μήτρα σύγχυσης παρουσιάζει την απόδοση του μοντέλου με τη μορφή πίνακα, όπου οι γραμμές αντιπροσωπεύουν τις πραγματικές κατηγορίες κι οι στήλες τις προβλεπόμενες βάσει των ετικετών της μεταβλητής στόχου (κλάσης). Στην περίπτωση της δυαδικής ταξινόμησης πρόκειται για έναν 2 × 2 πίνακα (Εικόνα 21).

Confusion Matrix	**Predicted Negative (0)**	**Predicted Positive (1)**
Actual Negative (0)	*True Negative (TN)*	*False Positive (FP)*
Actual Positive (1)	*False Negative (FN)*	*True Positive (TP)*

Εικόνα 21. Δομή μήτρας σύγχυσης

Η ερμηνεία των στοιχείων του πίνακα είναι η ακόλουθη:

- **True Negative (TN):** Τα σήματα που ήταν πραγματικά *θόρυβος (0)* και προβλέφθηκαν σωστά ως *θόρυβος (0)*.

- **False Positive (FP):** Τα σήματα που ήταν *θόρυβος (0)* αλλά προβλέφθηκαν λανθασμένα ως *αστέρας νετρονίων (1)*.
- **False Negative (FN):** Τα σήματα που ήταν *αστέρας νετρονίων (1)* αλλά προβλέφθηκαν λανθασμένα ως *θόρυβος (0)*.
- **True Positive (TP)**: Τα σήματα που ήταν πραγματικά *αστέρας νετρονίων (1)* και προβλέφθηκαν σωστά ως *αστέρας νετρονίων (1)*.

Η πληροφορία που παρέχει η μήτρα σύγχυσης παράγει τις ακόλουθες μετρικές αξιολόγησης της ταξινομητικής ικανότητας του αλγόριθμου:

- Accuracy (Ακρίβεια)

Η accuracy είναι το ποσοστό των σωστών προβλέψεων (θετικών και αρνητικών) προς τον συνολικό αριθμό των περιπτώσεων. Πρόκειται ίσως για την πιο ευρέως χρησιμοποιούμενη μετρική μιας και παρέχει μια ολιστική εκτίμηση της απόδοσης του μοντέλου.

Υπολογίζεται ως: **(TP+TN)/(TP+TN+FP+FN)**

Ωστόσο, δεν θα ήταν σε καμία περίπτωση ασφαλές η αξιολόγηση του μοντέλου να βασιστεί μόνο σε αυτήν. Σε datasets με σημαντική ανισορροπία μεταξύ των κλάσεων (όπως στο παρόν, όπου τα "θετικά" – σήματα pulsars είναι πολύ λιγότερα από τα "αρνητικά" – θόρυβος), η accuracy μπορεί να είναι παραπλανητική.

- Precision (Ακρίβεια θετικών προβλέψεων)

Η precision μετρά την ακρίβεια των θετικών προβλέψεων του μοντέλου. Πιο συγκεκριμένα, πρόκειται για το ποσοστό των σωστών προβλέψεων θετικών περιπτώσεων προς το σύνολο των προβλέψεων θετικών περιπτώσεων.

Υπολογίζεται ως: **TP/(TP+FP)**

Υψηλή precision σημαίνει ότι από τις θετικές περιπτώσεις (σήματα pulsars) που προβλέπει το μοντέλο οι ψευδώς θετικές (FP) είναι περιορισμένες. Με άλλα λόγια, το μοντέλο είναι “προσεκτικό” στις θετικές προβλέψεις του. Είναι προφανές πως στο πλαίσιο της παρούσας μελέτης, όπου η εκπαίδευση των μοντέλων αποβλέπει στη σωστή πρόβλεψη των θετικών περιπτώσεων (σήματα pulsars), η συγκεκριμένη μετρική είναι αρκετά σημαντική.

- Recall (Ανάκληση /Ευαισθησία)

Η recall (true positive rate) μετρά την ικανότητα του μοντέλου να εντοπίζει το σύνολο των θετικών περιπτώσεων (να μην χάνει θετικά στιγμιότυπα).

Υπολογίζεται ως: **TP/(TP+FN)**

Υψηλή recall σημαίνει ότι το μοντέλο “πιάνει” τα περισσότερα πραγματικά θετικά, ακόμη κι αν έχει περισσότερους ψευδώς θετικούς. Σε έρευνες όπως η παρούσα, όπου η θετική ετικέτα της κλάσης (σήματα αστέρων νετρονίων) είναι πολύ πιο σπάνια από την αρνητική (θόρυβος) και το να χαθεί κάποιο θετικό στιγμιότυπο είναι πολύ πιο σοβαρό από τα να καταχωρηθούν κάποια αρνητικά ως θετικά (FP), η recall αποτελεί ίσως την κρισιμότερη μετρική. Ακόμη κι αν πέσει λίγο η ακρίβεια του μοντέλου η απαίτηση για υψηλή recall είναι σημαντικότερη.

- F1-Score

Ο F1-score είναι ο αρμονικός μέσος της precision και της recall και παρέχει μία κάπως πιο εξισορροπημένη εικόνα της απόδοσης του ταξινομητή, όταν πρέπει να ληφθούν υπόψη τόσο τα FP όσο και τα FN.

Υπολογίζεται ως: $2*(Precision * Recall)/(Precision + Recall)$

Ενδείκνυται όταν χρειάζεται *ισορροπία μεταξύ ανίχνευσης και ακρίβειας* της θετικής κλάσης και είναι πολύ χρήσιμη μετρική για τον έλεγχο της ικανότητας του μοντέλου στο πεδίο των πρακτικών προβλέψεων.

- Balanced Accuracy

Η balanced accuracy είναι ο *μέσος όρος των recall (ευαισθησιών) για κάθε κλάση*. Δηλαδή, υπολογίζει πόσο καλά ταξινομούνται οι ορθές παρατηρήσεις σε κάθε κλάση, δίνοντας ίσο βάρος και στις δύο, ανεξάρτητα από το πόσες παρατηρήσεις έχει η καθεμία.

Υπολογίζεται ως: $½ * (Recall\ (θετικής) + Recall\ (αρνητικής))$

Λόγω της στατιστικής δομής της η balanced accuracy εμφανίζει μεγάλη ανθεκτικότητα στην ανισορροπία των κλάσεων, αποτελώντας έτσι βελτιωμένη εκδοχή της κλασικής accuracy και ως τέτοια συνιστά κρίσιμο δείκτη αξιολόγησης των μοντέλων για την παρούσα μελέτη.

- Macro και Weighted Averages

Ο παράγοντας *macro avg* υπολογίζει τον απλό μέσο όρο των μετρικών precision, recall και F1 για κάθε κλάση, χωρίς να λαμβάνει υπόψη το μέγεθος της. Ως τέτοιος δεν είναι και τόσο αξιόπιστος για σύνολα δεδομένων με ανισόρροπη κατανομή κλάσεων.

Ο *weighted avg*, από την άλλη, λαμβάνει υπόψη το μέγεθος κάθε κλάσης, υπολογίζοντας το σταθμισμένο μέσο όρο των προαναφερόμενων μετρικών. Συνεπώς εκφράζει καλύτερα την πραγματική εικόνα του μοντέλου σε περιπτώσεις ανισορροπίας κλάσης.

- ROC AUC

Η συγκεκριμένη μετρική προσδιορίζει τη διαχωριστική ικανότητα του μοντέλου μεταξύ των κλάσεων και έχει άμεση σχέση με την καμπύλη ROC (Receiver Operating Characteristic), η οποία οπτικοποιεί τη σχέση της True Positive Rate (Recall) έναντι της False Positive Rate ($FP/(FP+TN)$). Η τιμή της ROC AUC αντιστοιχεί στο εμβαδόν κάτω από την καμπύλη (Area Under Curve), συνοψίζοντας την πληροφορία της οπτικοποίησης σε έναν αριθμό που μπορεί να πάρει τιμές στο διάστημα [0.5, 1.0], όπου η τιμή 0,5 αντιστοιχεί σε τυχαίο ταξινομητή ενώ το 1 στην απόλυτη διαχωριστικότητα. Σε datasets με ανισορροπία κλάσεων ωστόσο, η ROC AUC μπορεί να παρουσιάσει παραπλανητικά υψηλές τιμές.

- Precision-Recall AUC

Σε αντίθεση με την ROC AUC, η Precision-Recall AUC εστιάζει αποκλειστικά στην απόδοση του ταξινομητή ως προς τη θετική κλάση, αγνοώντας τα True Negatives. Η καμπύλη Precision-Recall οπτικοποιεί της Precision και της Recall, για διαφορετικά thresholds. Το εμβαδόν της

επιφάνειας κάτω από την καμπύλη, το οποίο αντιστοιχεί στην τιμή της PR AUC, παρέχει μια αξιόπιστη συνολική εκτίμηση της ικανότητας του μοντέλου να εντοπίζει σπάνια αλλά κρίσιμα γεγονότα.

- ROC Curve (Receiver Operating Characteristic Curve)

Όπως προαναφέρθηκε η καμπύλη ROC είναι ένα διαγνωστικό εργαλείο που απεικονίζει την απόδοση ενός ταξινομητή καθώς μεταβάλλεται το κατώφλι ταξινόμησης, οπτικοποιώντας τη σχέση ανάμεσα στην True Positive Rate (TPR-Recall) στην False Positive Rate (FPR).

- **Άξονας Y:** Εμφανίζει την True Positive Rate (άξονας ευαισθησίας).
- **Άξονας X:** Εμφανίζει το False Positive Rate (άξονας ειδικότητας).

Ιδιότητες:

- Η τέλεια καμπύλη ROC φτάνει στην πάνω αριστερή γωνία (TPR = 1, FPR = 0).
- Η διαγώνιος (γραμμή y = x) αντιστοιχεί σε *τυχαίο ταξινομητή*.
- Όσο μεγαλύτερη η απόσταση από τη διαγώνιο τόσο καλύτερος ο ταξινομητής.
- Το *εμβαδόν κάτω από την καμπύλη (AUC-ROC)* συνοψίζει την απόδοση του μοντέλου.

Η καμπύλη ROC και ο αντίστοιχος δείκτης AUC ROC παρέχουν μια κάπως γενική εκτίμηση της διαχωριστικής ικανότητας του μοντέλου. Σε datasets με πολυπληθή αρνητική κλάση η παρουσία πολλών True Negatives (TN) μπορεί να οδηγήσει σε τεχνητά χαμηλό False Positive Rate με αποτέλεσμα την υπερεκτίμηση της απόδοσης του μοντέλου.

- Precision-Recall Curve

Η καμπύλη Precision-Recall εστιάζει αποκλειστικά στην *κλάση ενδιαφέροντος* (θετική κλάση), οπτικοποιώντας το *πόσο ακριβείς είναι οι θετικές προβλέψεις*, σε σχέση με το *πόσες από τις θετικές περιπτώσεις ανιχνεύονται*, καθώς μεταβάλλεται το κατώφλι ταξινόμησης.

- **Άξονας Y:** Εμφανίζει την Recall.
- **Άξονας X:** Εμφανίζει την Precision.

Ιδιότητες:

- Καθώς χαμηλώνει το κατώφλι πρόβλεψης, το μοντέλο προβλέπει περισσότερα δείγματα ως θετικά (αυξάνεται η Recall και ενδεχομένως μειώνεται η Precision)
- Το ιδανικό μοντέλο βρίσκεται στην πάνω δεξιά γωνία (Precision = 1, Recall = 1).
- Το *εμβαδόν κάτω από την καμπύλη (PR AUC)* εμφανίζει τη συνολική απόδοση.

Η καμπύλη Precision - Recall εστιάζει αποκλειστικά στην απόδοση του ταξινομητή ως προς την θετική κλάση, αγνοώντας τα True Negatives. Ως εκ τούτου, αποτελεί πιο αξιόπιστο δείκτη σε περιπτώσεις εντοπισμού σπάνιων αλλά κρίσιμων παραδειγμάτων, όπως τα σήματα των αστέρων νετρονίων στο πλαίσιο της παρούσας έρευνας.

Εργαλεία υλοποίησης

Για την υλοποίηση των τεσσάρων κλασικών αλγορίθμων ταξινόμησης – Logistic Regression, Random Forest, Support Vector Machine (SVM) και XGBoost – αξιοποιήθηκαν οι ευρέως χρησιμοποιούμενες βιβλιοθήκες της γλώσσας προγραμματισμού Python. Συγκεκριμένα, η

scikit-learn παρείχε τις έτοιμες κλάσεις για την εκπαίδευση, την αξιολόγηση και τη ρύθμιση των παραμέτρων των μοντέλων των Logistic Regression, Random Forest Classifier και SVM. Ενώ, για το XGBoost χρησιμοποιήθηκε η *βιβλιοθήκη* xgboost, η οποία επιτρέπει λεπτομερή έλεγχο των παραμέτρων και υψηλή απόδοση. Από την άλλη πλευρά, για την υλοποίηση των δύο μοντέλων fuzzy λογικής αναπτύχθηκαν (με χρήση της Python) δύο custom ταξινομητές, βασισμένοι στην αρχιτεκτονική του API της scikit-learn. Το μοντέλο Fuzzy k-NN που επέκτεινε το κλασικό k-NN με fuzzy weighting μηχανισμό στις πιθανότητες ταξινόμησης, λαμβάνοντας υπόψη την απόσταση κάθε γείτονα ως βαθμό συμμετοχής. Καθώς και το Fuzzy Decision Tree, το οποίο δημιουργεί "soft fuzzification" μέσω συναρτήσεων υπερβολικής εφαπτομένης και παράγει πολλαπλά fuzzy χαρακτηριστικά ανά κατηγόρημα, στηριζόμενο στη δομή ενός συμβατικού decision tree. Επιπλέον, αναπτύχθηκαν δύο απλοί αλγόριθμοι εξαγωγής fuzzy κανόνων σε φυσική γλώσσα, ώστε να διερευνηθεί η δυνατότητα ερμηνευσιμότητας των αποφάσεων του μοντέλου.

Logistic Regression

- ### Περιγραφή αλγόριθμου

Η Logistic Regression αποτελεί ένα από τα πιο θεμελιώδη και διαδεδομένα μοντέλα στη στατιστική και τη μηχανική μάθηση. Παρά την ονομασία της, δεν πρόκειται για μια μέθοδο παλινδρόμησης όπως η γραμμική παλινδρόμηση, αλλά για ένα πιθανολογικό μη γραμμικό μοντέλο ταξινόμησης, κατάλληλο για δυαδικές καθώς και για πολυκατηγορικές προβλέψεις. Η χρησιμότητά της έγκειται στην απλότητα, στην ερμηνευσιμότητα και στην αποδοτικότητα της, ιδιότητες που την καθιστούν κατάλληλη για ένα ευρύ φάσμα εφαρμογών.

Η κεντρική μαθηματική ιδέα αφορά στη μοντελοποίηση της πιθανότητας το παρατηρούμενο δείγμα να ανήκει σε μία από δύο κατηγορίες, δηλαδή σε μια δυαδική μεταβλητή στόχο y ∈ {0,1}. Αντί να προβλέπει απευθείας την κλάση, το μοντέλο υπολογίζει την πιθανότητα P(y = 1 | **x**), όπου **x** είναι το διάνυσμα των χαρακτηριστικών (features). Το μοντέλο βασίζεται στη χρήση της λογιστικής συνάρτησης (sigmoid) για τον περιορισμό των προβλέψεων στο διάστημα (0, 1). Συγκεκριμένα ο γραμμικός συνδυασμός των γνωρισμάτων μετασχηματίζεται μέσω της συνάρτησης:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

όπου $z = \mathbf{w}^T x + \mathbf{b}$, με τα **w** και **b** να είναι τα παραμετρικά βάρη και το διάνυσμα μετατόπισης αντίστοιχα. Συνεπώς, η τελική μορφή του μοντέλου που παράγει την εκτιμώμενη πιθανότατα η δυαδική μεταβλητή στόχος να παίρνει την τιμή 1 είναι:

$$P(y = 1 \mid \mathbf{x}) = \sigma(z)$$

Συνακόλουθα, η αντίστοιχη πιθανότητα η y να λαμβάνει την τιμή 0 προκύπτει:

$$P(y = 0 \mid \mathbf{x}) = 1 - \sigma(z)$$

Η εποπτευόμενη (supervised) εκπαίδευση του αλγόριθμου πραγματοποιείται μέσω της μέγιστης πιθανοφάνειας (Maximum Likelihood Estimation - MLE). Θεωρώντας ένα σύνολο N ανεξάρτητων παρατηρήσεων ($\mathbf{x}_i$, y_i), η συνάρτηση πιθανοφάνειας προκύπτει στην ακόλουθη μορφή:

$$L(\boldsymbol{w}, b) = \prod_{i=1}^{N} [\sigma(z_i)]^{y_i} [1 - \sigma(z_i)]^{1-y_i}$$

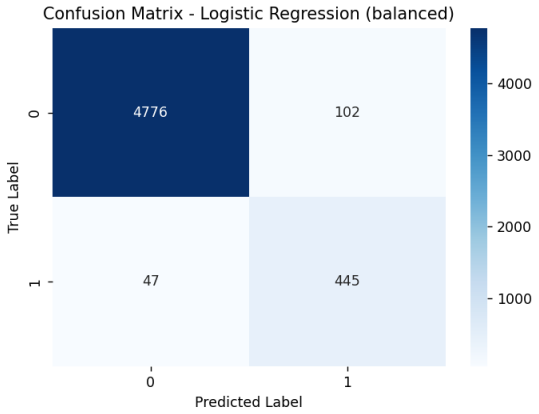
Για πρακτικούς λόγους που έχουν να κάνουν με τη μαθηματική δομή του προβλήματος και την πιθανοκρατική φύση του μοντέλου ο υπολογισμός προσανατολίζεται στην εκτίμηση των όρων μεγιστοποίησης της λογαριθμικής (λογιστικής) πιθανοφάνειας ή των αντίστοιχων της ελαχιστοποίησης της αρνητικής εκδοχής της, γνωστής συνάρτησης κόστους (binary Cross Entropy Loss). Οι παράγωγοι των συγκεκριμένων συναρτήσεων ως προς τα βάρη **w** και την παράμετρο b υποδεικνύουν την “κατεύθυνση” μεταβολής των βαρών ώστε να μειωθεί το σφάλμα. Οι νέες αυτές παράμετροι βελτιστοποίησης υπολογίζονται μέσω της εφαρμογής μεθόδων βαθμίδωσης (gradient descent). Ενώ η περίπτωση υπερπροσαρμογής (overfitting), προστίθεται κανονικοποίηση (regularization), συχνά τύπου L2 (ridge).

Η ξεκάθαρη μαθηματική δομή και η σχετικά απλή κατανόηση των πιθανοκρατικών όρων του συγκεκριμένου μοντέλου του προσδίδουν ορισμένα σημαντικά πλεονεκτήματα όπως: Η απλότητα κι η ταχύτητα, το χαμηλό κόστος των πόρων συστήματος (μνήμη και επεξεργαστική

ισχύ), η εξαιρετική απόδοση σε σύνολα γραμμικά διαχωρίσιμων δεδομένων, η πολύ ισχυρή στατιστική θεμελίωση που οδηγεί σε σαφείς ερμηνείες των αποτελεσμάτων, καθώς και η εύκολη κανονικοποίηση. Από την άλλη μεριά, ο αλγόριθμος εμφανίζεται κάπως αδύναμος στη μοντελοποίηση μη γραμμικών σχέσεων μεταξύ των χαρακτηριστικών και του λογάριθμου των πιθανοτήτων, ευαίσθητος στα πολυσυσχετισμένα (multicollinearity) χαρακτηριστικά και στη διαχείριση των outliers, καθώς και λιγότερο αποδοτικός σε πολυπλοκότερα προβλήματα.

- Αποτελέσματα εκπαίδευσης

Η εκπαίδευση της Logistic Regression με ενσωματωμένο μηχανισμό εξισορρόπησης των κλάσεων (balanced) επέστρεψε ιδιαίτερα ικανοποιητικά αποτελέσματα, αναδεικνύοντας αρκετά καλή διαχωριστική ικανότητα ως προς τον εντοπισμό των πραγματικών σημάτων των αστέρων νετρονίων έναντι των σημάτων που οφείλονται σε θόρυβο (RFI).



Εικόνα 22. Confusion Matrix Logistic Regression (balanced)

Η μήτρα σύγχυσης του αλγόριθμου (Εικόνα 22) υποδεικνύει:

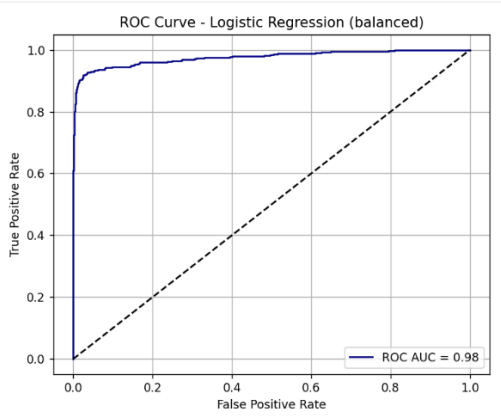
- 4776 σωστές προβλέψεις της αρνητικής κλάσης (True Negatives)
- 445 σωστές προβλέψεις της θετικής κλάσης (True Positives)
- 102 False Positives (σήματα θορύβου ταξινομήθηκαν εσφαλμένα ως pulsars)
- 47 False Negatives (πραγματικά pulsars που δεν αναγνωρίστηκαν)

```
=== Αποτελέσματα για Logistic Regression (balanced) ===
              precision    recall  f1-score   support

           0     0.9903    0.9791    0.9846      4878
           1     0.8135    0.9045    0.8566       492

    accuracy                         0.9723      5370
   macro avg     0.9019    0.9418    0.9206      5370
weighted avg     0.9741    0.9723    0.9729      5370

Accuracy             : 0.9723
Balanced Accuracy    : 0.9418
ROC AUC              : 0.9758
Precision-Recall AUC : 0.9252
```

Εικόνα 23. Results Logistic Regression (balanced)

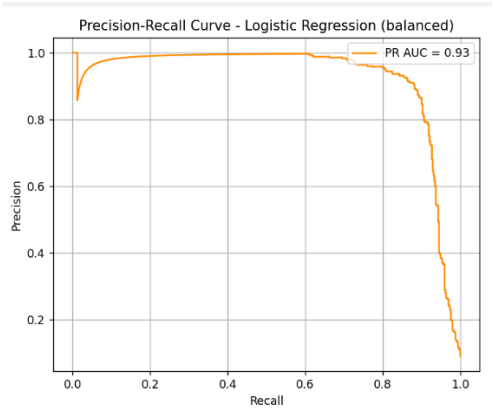
Η ταξινομητική επάρκεια του μοντέλου επαληθεύεται και από τις τιμές των μετρικών που επιστρέφει η εκπαίδευση (Εικόνα 23). Η *accuracy* ανέρχεται στο **97.23%**, γεγονός που δείχνει ότι το μοντέλο ταξινόμησε σωστά τη συντριπτική πλειοψηφία των παραδειγμάτων. Ωστόσο, λόγω της ανισορροπίας στο dataset (πολύ περισσότερα αρνητικά δείγματα από θετικά), η *balanced accuracy* αποτελεί πιο αξιόπιστη ένδειξη, φτάνοντας το **94.18%**. Η υψηλή αυτή τιμή

επιβεβαιώνει ότι το μοντέλο αποδίδει καλά και στις δύο κλάσεις, εντοπίζοντας επαρκώς και τα στιγμιότυπα της μειοψηφικής (πραγματικοί pulsars). Η *precision* για την κλάση 1 (θετικά παραδείγματα) ανήλθε σε **81.35%**, υποδεικνύοντας ότι περίπου 8 στις 10 προβλέψεις ως pulsars ήταν όντως σωστές. Παράλληλα, η *recall* για την ίδια κλάση φτάνει το **90.45%**, γεγονός που υποδηλώνει ότι το μοντέλο κατάφερε να ανιχνεύσει τη συντριπτική πλειοψηφία των πραγματικών θετικών περιπτώσεων (445 από τις 492). Ο *f1-score* (**85.66%**) για την κλάση 1 ενσωματώνει την ισορροπία μεταξύ *precision* και *recall* και επιβεβαιώνει την καλή απόδοση ως προς τον εντοπισμό των θετικών περιπτώσεων. Η ελαφρώς μεγαλύτερη τιμή δε των False Positives στη μήτρα σύγχυσης φανερώνει ότι το μοντέλο είναι ελάχιστα πιο "επιθετικό" στον χαρακτηρισμό παραδειγμάτων ως θετικά, ενδεχομένως ως αποτέλεσμα της εξισορρόπησης κλάσεων, κάτι που είναι επιθυμητό σε περιβάλλοντα, όπου το κόστος της μη αναγνώρισης ενός πραγματικού γεγονότος είναι υψηλότερο από εκείνο ενός ψευδώς θετικού.



Εικόνα 24. ROC Curve Logistic Regression

Η καμπύλη ROC (Εικόνα 24) αναπαριστά γραφικά την εξαιρετική διαχωριστική ικανότητα του μοντέλου, καθώς η γραμμή προσεγγίζει το άνω αριστερό μέρος του διαγράμματος, το οποίο αντιπροσωπεύει την ιδανική συμπεριφορά (υψηλό TPR, χαμηλό FPR). Η περιοχή κάτω από την καμπύλη (AUC) ανέρχεται σε **0.98**, επιβεβαιώνοντας τη δυνατότητα του μοντέλου να διαχωρίζει αποτελεσματικά τις δύο κλάσεις ανεξαρτήτως επιλογής κατωφλίου.



Εικόνα 25. Precision-Recall Curve Logistic Regression

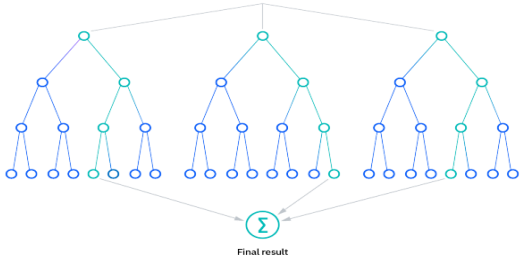
Δεδομένης της έντονης ανισορροπίας στο dataset (πολύ λιγότερα θετικά παραδείγματα), η καμπύλη *Precision-Recall* (Εικόνα 25) αποτελεί ακόμη πιο αντιπροσωπευτικό εργαλείο αξιολόγησης. Η PR αναπαράσταση της Logistic Regression δείχνει ότι το μοντέλο διατηρεί υψηλές τιμές *precision* (δηλαδή λίγες ψευδώς θετικές προβλέψεις) ακόμη και για σχετικά

υψηλά επίπεδα *recall*. Η AUC της καμπύλης PR (**0.93**) φανερώνει πολύ καλή επίδοση στην ανίχνευση των σπάνιων θετικών παραδειγμάτων, επιβεβαιώνοντας ότι το μοντέλο έχει μικρή πιθανότητα να παραβλέψει πραγματικά σήματα αστέρων νετρονίων. Η μικρή πτώση της precision για recall πολύ κοντά στο 1 είναι αναμενόμενη, καθώς το μοντέλο στην προσπάθειά του να εντοπίσει το 100% των θετικών περιπτώσεων, καταλήγει να αυξάνει κάπως τον αριθμό των ψευδώς θετικών προβλέψεων.

Random Forest Classifier

- Περιγραφή αλγόριθμου

Ο Random Forest Classifier αποτελεί έναν από τους πιο διαδεδομένους και αποδοτικούς αλγορίθμους επίβλεψης (supervised learning) στο πεδίο της μηχανικής μάθησης. Βασίζεται στην έννοια του ensemble learning, κατά την οποία συνδυάζονται πολλαπλά μοντέλα για την παραγωγή ενός ισχυρότερου τελικού μοντέλου (Εικόνα 26). Στην περίπτωση του Random Forest, τα επιμέρους μοντέλα είναι δέντρα απόφασης (decision trees) που εκπαιδεύονται με τυχαίες υποδειγματοληψίες των δεδομένων, αλλά και των χαρακτηριστικών. Ο αλγόριθμος εκμεταλλεύεται την ποικιλομορφία των δέντρων για να μειώσει την υπερπροσαρμογή (overfitting) και να ενισχύσει τη γενίκευση.



Εικόνα 26. Διάγραμμα δομής Random Forest

Συγκεκριμένα, το μοντέλο εκπαιδεύεται δημιουργώντας ένα σύνολο Τ της μορφής $\{h_1(x), h_2(x), \ldots, h_T(x)\}$, όπου κάθε στοιχείο $h_t(x)$ είναι ένα διακριτό δέντρο απόφασης που λαμβάνει ως είσοδο το διάνυσμα χαρακτηριστικών $x \in \mathbb{R}$ και παράγει (ως έξοδο) μια προβλεπόμενη ετικέτα $y \in Y$. Η διαδικασία κατασκευής κάθε τέτοιου δέντρου περιλαμβάνει δύο βασικά στάδια τυχαιότητας:

1. **Bootstrap sampling:** Για κάθε δέντρο, δημιουργείται ένα υποσύνολο εκπαίδευσης D_t με δειγματοληψία και με επανατοποθέτηση (sampling with replacement) από το αρχικό σύνολο δεδομένων *D*.
2. **Τυχαία υποεπιλογή χαρακτηριστικών:** Κατά την κατασκευή κάθε κόμβου του δέντρου, αντί να εξετάζονται όλα τα χαρακτηριστικά για διαχωρισμό, εξετάζεται μόνο ένα τυχαίο υποσύνολο μεγέθους *m* (με $m < d$, όπου *d* το μέγεθος όλων των χαρακτηριστικών). Η διαδικασία αυτή συμβάλλει στη διαφοροποίηση των δέντρων και μειώνει τη συσχέτισή τους.

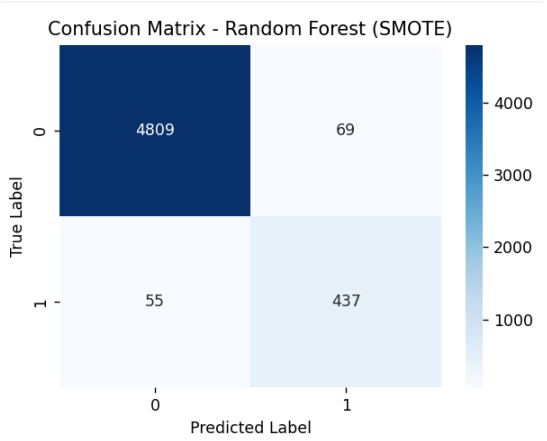
Το κάθε δέντρο κατασκευάζεται έτσι ώστε να ελαχιστοποιείται το κατάλληλο κριτήριο καθαρότητας. Τα κριτήρια που συνήθως χρησιμοποιούνται είναι η *εντροπία* και ο *δείκτης Gini*. Η τελική πρόβλεψη του μοντέλου στα προβλήματα ταξινόμησης, όπως το πρόβλημα της παρούσας μελέτης, προκύπτει από την τιμή που εμφανίζει η πλειοψηφία των μεμονωμένων δέντρων (majority voting).

Η βασική ιδέα ανάπτυξης του Random Forest αφορά στη μείωση της διασποράς (variance) των προβλέψεων, χωρίς να προκύψει σημαντική αύξηση της μεροληψίας (bias) του μοντέλου. Έτσι, καθώς κάθε επιμέρους δέντρο έχει υψηλή διασπορά αλλά χαμηλή μεροληψία, ο συνδυασμός των δύο οδηγεί σε σταθερότερο και περισσότερο γενικεύσιμο μοντέλο. Επίσης, ο ίδιος συνδυασμός κάνει τον Random Forest ανθεκτικό στον θόρυβο και αρκετά αποτελεσματικό στη διαχείριση συνόλων δεδομένων με μη γραμμικές συσχετίσεις ανάμεσα στα γνωρίσματά τους. Τέλος, καθίσταται ικανός να χειριστεί τόσο κατηγορικές όσο και συνεχής μεταβλητές και εμφανίζει καλή συμπεριφορά σε datasets με ελλείπουσες τιμές και μη ισορροπημένες κατηγορίες. Τα πλεονεκτήματά του αυτά συμπληρώνονται από την ισχυρή ικανότητα γενίκευσης και την ευκολία παραμετροποίησης, οι οποίες αυξάνουν κατά πολύ την αποτελεσματικότητά του σε πρακτικά προβλήματα. Παρόλο που δεν είναι τόσο διαφανής όσο ένα μεμονωμένο δέντρο, παρέχει επαρκή πληροφορία για τη σημασία των μεταβλητών.

Τέλος, ορισμένα μειονεκτήματα του μοντέλου είναι ότι μπορεί να καταστεί αργό στην πρόβλεψη όταν το πλήθος των δέντρων είναι μεγάλο ή όταν τα δεδομένα είναι υψηλής διάστασης, ενώ δεν λειτουργεί καλά και σε προβλήματα που απαιτούν πολύπλοκες χωρικές εξαρτήσεις, όπως σε ορισμένες εφαρμογές εικόνας ή φυσικής γλώσσας, όπου τα νευρωνικά δίκτυα υπερτερούν.

- Αποτελέσματα εκπαίδευσης

Η εκπαίδευση του μοντέλου Random Forest σε συνδυασμό με τη μέθοδο SMOTE είχε ως αποτέλεσμα την επίτευξη ιδιαίτερα ικανοποιητικών επιδόσεων στο πρόβλημα ταξινόμησης των αστρονομικών σημάτων. Η επιλογή του συγκεκριμένου αλγορίθμου βασίστηκε στην ικανότητά του να χειρίζεται μη γραμμικές σχέσεις και αλληλεπιδράσεις μεταξύ μεταβλητών, ενώ η χρήση του SMOTE συνέβαλε στην αντιμετώπιση της έντονης ανισορροπίας του συνόλου δεδομένων, αυξάνοντας τεχνητά τα δείγματα της μειοψηφικής κλάσης (πραγματικά σήματα αστέρων νετρονίων).



Εικόνα 27. Confusion Matrix Random Forest (SMOTE)

Η μήτρα σύγχυσης του αλγόριθμου (Εικόνα 27) υποδεικνύει:

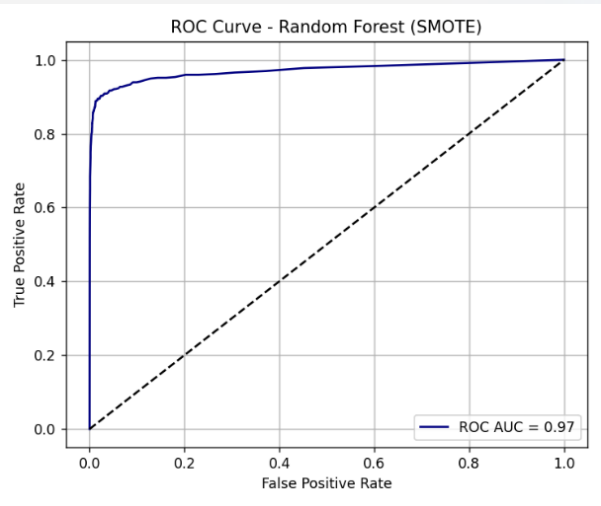
- 4809 σωστές προβλέψεις της αρνητικής κλάσης (True Negatives)
- 437 σωστές προβλέψεις της θετικής κλάσης (True Positives)
- 69 False Positives (σήματα θορύβου ταξινομήθηκαν εσφαλμένα ως pulsars)
- 55 False Negatives (πραγματικά pulsars που δεν αναγνωρίστηκαν)

```
=== Αποτελέσματα για Random Forest (SMOTE) ===
              precision    recall  f1-score   support

           0     0.9887    0.9859    0.9873      4878
           1     0.8636    0.8882    0.8758       492

    accuracy                         0.9769      5370
   macro avg     0.9262    0.9370    0.9315      5370
weighted avg     0.9772    0.9769    0.9771      5370

Accuracy             : 0.9769
Balanced Accuracy    : 0.9370
ROC AUC              : 0.9707
Precision-Recall AUC : 0.9200
```

Εικόνα 28. Results Random Forest (SMOTE)

Η οθόνη των αποτελεσμάτων (Εικόνα 28) εμφανίζει συνολική ακρίβεια (*accuracy*) της τάξης του **97.69%**, επιτυγχάνοντας εξαιρετικά ποσοστά σωστής ταξινόμησης στο σύνολο των παρατηρήσεων. Ωστόσο, δεδομένου ότι η ακρίβεια από μόνη της μπορεί να διαμορφώσει παραπλανητική εικόνα σε περιπτώσεις ανισόρροπων κλάσεων, μεγαλύτερη βαρύτητα έχει η μετρική της *balanced accuracy*, η οποία ανήλθε στο **93.70%** και η οποία υπολογίζει τον μέσο όρο της ευαισθησίας (*recall*) και για τις δύο κλάσεις, παρέχοντας μια πιο αντιπροσωπευτική εικόνα της πραγματικής απόδοσης του μοντέλου.

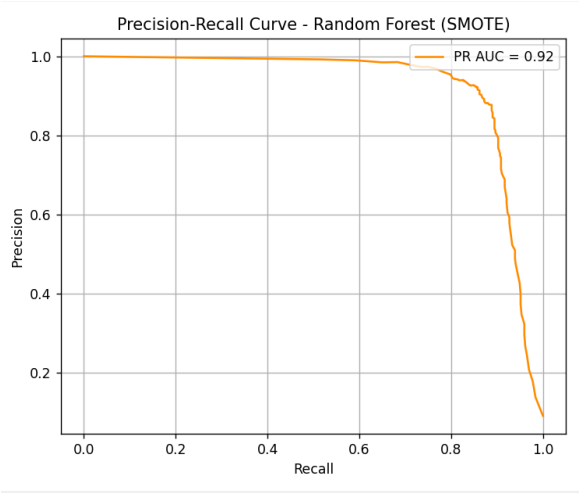
Σε ό,τι αφορά τις επιμέρους μετρικές της μειοψηφικής θετικής κλάσης (σήματα αστέρων νετρονίων), η *recall* ανήλθε στο **88.82%**, γεγονός που φανερώνει ότι το μοντέλο εντόπισε επιτυχώς τη συντριπτική πλειονότητα των πραγματικών σημάτων. Η *precision*, δηλαδή η ικανότητα του μοντέλου να αποφεύγει τις ψευδώς θετικές προβλέψεις, καταγράφηκε στο **86.36%**, ενώ ο *f1-score*, που συνδυάζει τις δύο αυτές πτυχές, ήταν **87.58%**. Τα αποτελέσματα αυτά δείχνουν πως το μοντέλο όχι μόνο εντοπίζει τα περισσότερα σήματα, αλλά το κάνει και με ικανοποιητική ακρίβεια, αποφεύγοντας την υπερταξινόμηση των αρνητικών περιπτώσεων ως θετικές.



Εικόνα 29. ROC Curve Random Forest

Η επάρκεια της διαχωριστικής ικανότητας του μοντέλου επικυρώνεται και από την οπτική αναπαράσταση της καμπύλης ROC (Εικόνα 29). Η καμπύλη πλησιάζει το άνω αριστερό όριο του διαγράμματος, γεγονός που υποδεικνύει ότι το μοντέλο επιτυγχάνει υψηλό ποσοστό εντοπισμού των θετικών περιπτώσεων (TP), με ταυτόχρονα χαμηλό αντίστοιχο ποσοστό των ψευδώς θετικών (FP). Η περιοχή κάτω από την καμπύλη (*AUC*) έχει τιμή **0.9707**, δηλαδή πολύ

κοντά στο μέγιστο θεωρητικό όριο 1.0, αποδεικνύοντας πως το μοντέλο διαχωρίζει αρκετά αποτελεσματικά τις δύο κλάσεις ανεξαρτήτως του ορίου ταξινόμησης.



Εικόνα 30. Precision-Recall Curve Random Forest

Από την άλλη μεριά, η *Precision-Recall Curve* (Εικόνα 30) εμφανίζει επίσης ικανοποιητικά αποτελέσματα, με το *PR AUC* στο υψηλό **0.9200**. Η καμπύλη διατηρεί υψηλές τιμές precision ακόμα και για μεγάλη recall, γεγονός που δείχνει πως το μοντέλο κατορθώνει να εντοπίζει μεγάλο ποσοστό των θετικών περιπτώσεων χωρίς να επιβαρύνεται υπερβολικά από τις ψευδώς θετικές, πράγμα ιδιαιτέρως σημαντικό στο πλαίσιο της παρούσας έρευνας, όπου η κλάση των θετικών (σήματα αστέρων νετρονίων) είναι πολύτιμη και σπάνια.

Συνοψίζοντας, η χρήση του Random Forest σε συνδυασμό με την τεχνική του SMOTE αποδεικνύεται ιδιαίτερα αποτελεσματική στην παρούσα εφαρμογή ταξινόμησης. Το μοντέλο επιδεικνύει αξιοσημείωτη ικανότητα να ισορροπεί ανάμεσα στην αναγνώριση των σπάνιων σημάτων και στη διατήρηση υψηλής ακρίβειας και σταθερότητας στη συνολική πρόβλεψη. Το γεγονός αυτό, σε συνδυασμό με την καλή επίδοση στις κρίσιμες μετρικές (recall, precision, balanced accuracy και PR AUC), καθιστά το συγκεκριμένο μοντέλο κατάλληλο για εφαρμογές όπου η σωστή ανίχνευση σπάνιων φαινομένων καθίσταται ζωτικής σημασίας.

Support Vector Machine (SVM)

- Περιγραφή αλγόριθμου

Ο αλγόριθμος Support Vector Machine (SVM) αποτελεί ένα από τα σημαντικότερα εργαλεία της εποπτευόμενης μηχανικής μάθησης. Σχεδιάστηκε για να λύνει προβλήματα ταξινόμησης, δηλαδή να διαχωρίζει τα δεδομένα σε δύο ή περισσότερες κατηγορίες με τον καλύτερο δυνατό τρόπο. Συνδυάζει το αυστηρό μαθηματικό υπόβαθρο με την πρακτική ευελιξία, γι' αυτό και γνωρίζει ευρεία χρήση σε μια μεγάλη περιοχή εφαρμογών – από την ανάλυση εικόνας μέχρι τη βιοπληροφορική και τα συστήματα ασφαλείας.

Ο πυρήνας της βασικής ιδέας του συγκεκριμένου αλγόριθμου αφορά στην επίλυση ενός κατά βάση γεωμετρικού προβλήματος: *Η εύρεση του ορίου που διαχωρίζει όσο το δυνατόν καλύτερα τα δεδομένα μεταξύ δύο ή περισσότερων κατηγοριών*. Ανάλογα με τον αριθμό των διαστάσεων του χώρου των δεδομένων το όριο αυτό μπορεί να είναι μία γραμμή (σε δύο διαστάσεις), ένα επίπεδο (σε τρεις διαστάσεις) ή, γενικότερα, ένα υπερεπίπεδο (σε

περισσότερες διαστάσεις). Ζητούμενο δεν είναι μόνο να χωριστούν σωστά τα δεδομένα, αλλά *να τοποθετηθεί αυτό το όριο όσο το δυνατόν μακρύτερα από τα κοντινότερα σημεία κάθε κατηγορίας* — μια προσέγγιση που μεγιστοποιεί το "περιθώριο ασφαλείας" και μειώνει τον κίνδυνο σφαλμάτων σε νέα άγνωστα δεδομένα. Τα "κρίσιμα" σημεία που καθορίζουν το όριο λέγονται support vectors, και είναι οι παρατηρήσεις που βρίσκονται πιο κοντά στο διαχωριστικό υπερεπίπεδο. Αξίζει να σημειωθεί ότι μόνο αυτά επηρεάζουν άμεσα τη θέση του μοντέλου, γεγονός που καθιστά τον SVM υπολογιστικά αποδοτικό σε προβλήματα με μεγάλο αριθμό χαρακτηριστικών.

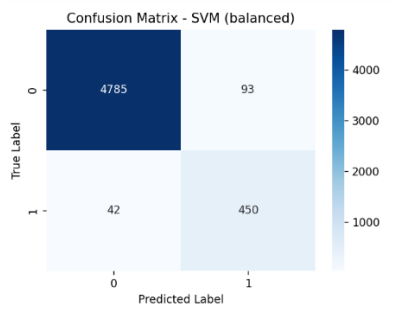
Στην πράξη ωστόσο τα δεδομένα σπάνια είναι τέλεια διαχωρίσιμα. Συχνά υπάρχουν επικαλύψεις ή ακραίες τιμές. Στις περιπτώσεις αυτές εισάγονται ειδικές μεταβλητές ανοχής (slack variables), οι οποίες επιτρέπουν ελεγχόμενες παραβιάσεις του κανόνα ταξινόμησης με κέρδος τη μεγαλύτερη ευελιξία του μοντέλου. Κάπως έτσι ο αλγόριθμος προσαρμόζεται και σε πραγματικά σενάρια όπου ο θόρυβος ή οι επικαλύψεις είναι αναπόφευκτες.

Επίσης, υπάρχουν προβλήματα που οι κλάσεις δεν μπορούν να διαχωριστούν γραμμικά. Εδώ ο SVM επιδεικνύει μια εξαιρετική καινοτομία: τη χρήση των *πυρηνικών συναρτήσεων (kernels)*. Μέσω αυτών, τα δεδομένα προβάλλονται σε έναν νέο, υψηλότερων διαστάσεων χώρο, όπου μπορεί να γίνει γραμμικός διαχωρισμός, ακόμη κι αν αυτό δεν είναι δυνατό στον αρχικό χώρο. Αυτό γίνεται χωρίς να χρειάζεται ρητός υπολογισμός των νέων διαστάσεων. Ουσιαστικά, υπολογίζονται μόνο τα εσωτερικά γινόμενα μεταξύ σημείων, χρησιμοποιώντας ειδικές συναρτήσεις (όπως τον RBF kernel ή τον πολυωνυμικό kernel). Αυτή η τεχνική, γνωστή ως *“kernel trick”*, δίνει τη δυνατότητα επίλυσης πολύπλοκων προβλημάτων με εντυπωσιακή αποδοτικότητα.

Τα κυριότερα πλεονεκτήματα που προκύπτουν από τη μαθηματική δομή και από τα λειτουργικά χαρακτηριστικά του αλγόριθμου είναι: Η πολύ υψηλή ακρίβεια ταξινόμησης, η ανθεκτικότητα στην υπερεκπαίδευση (overfitting), ειδικά σε μικρά σύνολα, καθώς και η δυνατότητα διαχείρισης σύνθετων, μη γραμμικών προβλημάτων. Από την άλλη, κάποιοι από τους περιορισμούς του μοντέλου είναι: η χαμηλή του απόδοση σε πολύ μεγάλα σύνολα δεδομένων, ο απαιτούμενος πειραματισμός για τη ρύθμιση παραμέτρων κι οι σημαντικές αδυναμίες στην ερμηνεία των αποτελεσμάτων του.

- Αποτελέσματα εκπαίδευσης

Το μοντέλο Support Vector Machine (SVM), με ενσωμάτωση της ρύθμισης εξισορρόπησης κλάσεων (balanced class weights), παρουσίασε εξαιρετικά πειστική απόδοση στο πρόβλημα ταξινόμησης των σημάτων αστρονομικής παρατήρησης. Παρά την εγγενή ανισορροπία του dataset, ο αλγόριθμος κατόρθωσε να διατηρήσει υψηλή γενική ακρίβεια και παράλληλα να εντοπίσει με αξιοσημείωτη επιτυχία τις θετικές περιπτώσεις.



Εικόνα 31. Confusion Matrix SVM (balanced)

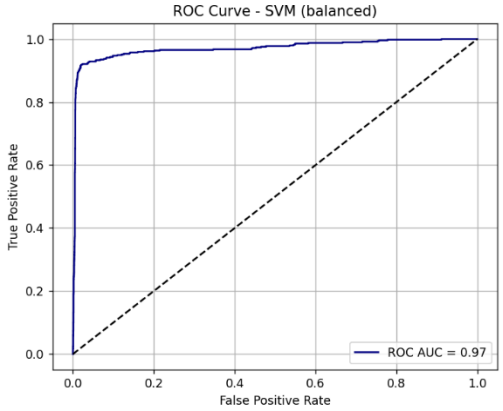
Η μήτρα σύγχυσης του αλγόριθμου (Εικόνα 31) υποδεικνύει:

- 4785 σωστές προβλέψεις της αρνητικής κλάσης (True Negatives)
- 450 σωστές προβλέψεις της θετικής κλάσης (True Positives)
- 93 False Positives (σήματα θορύβου ταξινομήθηκαν εσφαλμένα ως pulsars)
- 42 False Negatives (πραγματικά pulsars που δεν αναγνωρίστηκαν)

```
=== Αποτελέσματα για SVM (balanced) ===
              precision    recall  f1-score   support

           0     0.9913    0.9809    0.9861      4878
           1     0.8287    0.9146    0.8696       492

    accuracy                         0.9749      5370
   macro avg     0.9100    0.9478    0.9278      5370
weighted avg     0.9764    0.9749    0.9754      5370

Accuracy             : 0.9749
Balanced Accuracy    : 0.9478
ROC AUC              : 0.9710
Precision-Recall AUC : 0.8710
```

Εικόνα 32. Results SVM (balanced)

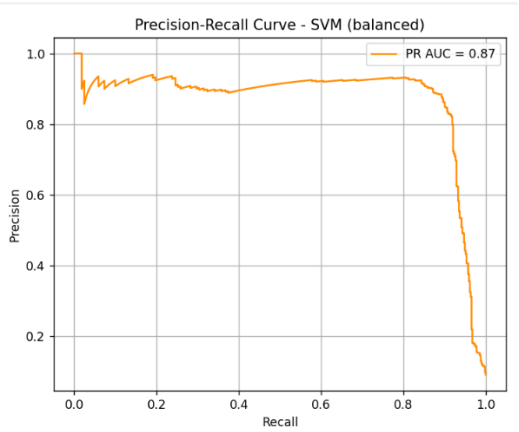
Οι τιμές των κρίσιμων μετρικών που επιστρέφει η οθόνη των αποτελεσμάτων (Εικόνα 32) επιβεβαιώνουν την καλή επίδοση του μοντέλου. Η precision της θετικής κλάσης κυμαίνεται στο **0.8287**, δηλαδή από τα γεγονότα που το μοντέλο προέβλεψε ως θετικά, το 82.87% ήταν πράγματι θετικά. Η τιμή της αντίστοιχης recall (θετικής κλάσης) ανέρχεται στο **0.9146**, που σημαίνει ότι το μοντέλο ανιχνεύει σωστά 9 στα 10 θετικά γεγονότα περιορίζοντας εξαιρετικά την απώλεια σημαντικής επιστημονικής πληροφορίας. Έτσι το F1-score για τη θετική κλάση διαμορφώνεται στο **0.8696**, τιμή που επίσης αντανακλά την επάρκεια του μοντέλου. Για την αρνητική κλάση οι τιμές των αντίστοιχων μετρικών είναι precision **0.9913**, recall **0.9809** και F1-score **0.9861**. Το σχεδόν απόλυτο των συγκεκριμένων επιδόσεων οφείλεται σίγουρα στην κυριαρχία των αρνητικών στιγμιότυπων, λόγω της ανισορροπίας των κλάσεων. Τέλος, η τιμή της συνολικής ακρίβειας (*accuracy*) ανέρχεται στο εντυπωσιακό **0.9749** και η αντίστοιχη της ακόμη αντιπροσωπευτικότερης balanced accuracy (του μέσου όρου της recall και για τις δύο κλάσεις) στο **0.9478**.



Εικόνα 33. ROC Curve SVM

Η καμπύλη ROC (Εικόνα 33), με εμβαδόν κάτω από την καμπύλη (AUC) ίσο με **0.97**, καταδεικνύει ότι το μοντέλο είναι ιδιαίτερα ικανό στο να διαχωρίζει τις δύο τάξεις, ακόμα και όταν αλλάζει το κατώφλι ταξινόμησης. Η κλίση της καμπύλης προς το άνω αριστερό άκρο

του γραφήματος συνιστά ασφαλή ένδειξη πως πρόκειται για μοντέλο που επιτυγχάνει υψηλό ποσοστό αληθώς θετικών προβλέψεων (True Positives), διατηρώντας ταυτόχρονα χαμηλό το αντίστοιχο ποσοστό των ψευδώς θετικών (False Positives).



Εικόνα 34. Precision-Recall Curve SVM

Η καμπύλη Precision-Recall (Εικόνα 34) παρουσιάζει επίσης εξαιρετική εικόνα. Με AUC ίσο με **0.87**, δεικνύει ότι η απόδοση του μοντέλου παραμένει υψηλή και στην περίπτωση που το βάρος πέφτει στη θετική κλάση. Η υψηλή ακρίβεια (precision) και η σταθερή ανάκληση (recall) σε όλο το φάσμα της διακύμανσης των thresholds καταδεικνύουν ότι το μοντέλο καταφέρνει να αναγνωρίζει επιτυχώς τις θετικές περιπτώσεις χωρίς σημαντική επιβάρυνση σε λανθασμένες θετικές προβλέψεις.

Συμπερασματικά, το μοντέλο SVM στην balanced εκδοχή του, επιδεικνύοντας ισχυρή απόδοση και σταθερότητα στην ταξινόμηση και αντιμετωπίζοντας πολύ ικανοποιητικά την ανισορροπία των κλάσεων, συνιστά μια ακόμη αξιόπιστη επιλογή για την αυτοματοποιημένη διαλογή των αστρονομικών δεδομένων.

XGBoost (Extreme Gradient Boosting)

- Περιγραφή αλγόριθμου

Ο XGBoost, συντομογραφία του *Extreme Gradient Boosting*, είναι ένας επίσης εξαιρετικά αποτελεσματικός αλγόριθμος μηχανικής μάθησης, ο οποίος ανήκει στην κατηγορία των ενισχυμένων μοντέλων (*boosting models*). Η επιτυχία του βασίζεται τόσο στη θεωρητική του δομή όσο και στην αποτελεσματική του υλοποίηση. Αναπτύχθηκε με σκοπό να προσφέρει υψηλή ακρίβεια πρόβλεψης και μεγάλη ταχύτητα εκπαίδευσης, πράγμα που τον καθιστά ιδανικό για μεγάλης κλίμακας προβλήματα σε πραγματικά δεδομένα.

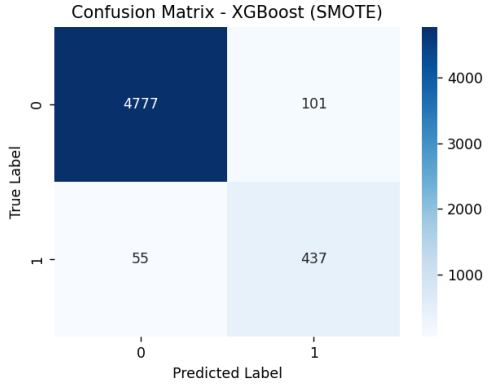
Η βασική ιδέα του XGBoost είναι απλή: δημιουργεί μια ακολουθία από πολλά απλά μοντέλα, συνήθως δέντρα απόφασης (decision trees), και κάθε νέο μοντέλο μαθαίνει από τα λάθη που έκαναν τα προηγούμενα. Αντί να εκπαιδεύεται ένα μοντέλο μόνο του, τα μοντέλα συνεργάζονται, ώστε να βελτιώνουν διαρκώς την πρόβλεψη. Αυτό το είδος εκπαίδευσης λέγεται *ενίσχυση μέσω βαθμίδωσης* (gradient boosting) και επιτυγχάνεται με τη σταδιακή βελτιστοποίηση συναρτήσεων κόστους μέσω παραγώγων. Στην ουσία, κάθε νέο δέντρο «σπρώχνει» το σύνολο του μοντέλου προς τη σωστή κατεύθυνση, μαθαίνοντας από τα σφάλματα του προηγούμενου. Ο XGBoost αναπτύσσει αυτή τη βασική ιδέα με τεχνικές που

κάνουν την εκπαίδευση γρηγορότερη, πιο ακριβή και πιο ανθεκτική σε προβλήματα όπως η υπερπροσαρμογή (overfitting).

Τα παραπάνω χαρακτηριστικά προσδίδουν στο μοντέλο σημαντικά πλεονεκτήματα όπως η ταχύτητα και η υψηλή απόδοση, η αποτελεσματική αντιμετώπιση της πολυπλοκότητας, η διαχείριση ελλιπών ή ακαθάριστων δεδομένων, η επιτυχής αξιολόγηση της σημαντικότητας των χαρακτηριστικών, καθώς και η προσαρμοστικότητα σε πολλά προβλήματα. Από την άλλη μεριά, ορισμένα μειονεκτήματα του συγκεκριμένου αλγόριθμου είναι: η πολυπλοκότητα της παραμετροποίησης, η μεγάλη κατανάλωση πόρων (μνήμης και επεξεργαστικής ισχύος), η δυσκολία στην ερμηνευσιμότητα (λόγω της πολύπλοκης δομής των πολλών δέντρων δεν παρέχει άμεση σχέση εισόδου - εξόδου), καθώς και η μικρή απόδοση σε μη δομημένα δεδομένα (εικόνες, κείμενο, ήχος).

- Αποτελέσματα εκπαίδευσης

Το μοντέλο XGBoost σε συνδυασμό με την τεχνική SMOTE για την εξισορρόπηση των κλάσεων, παρουσίασε επίσης πολύ καλή συνολική απόδοση στο πρόβλημα ταξινόμησης που απασχολεί την παρούσα εργασία. Η συμπεριφορά του ως προς την αναγνώριση τόσο της πλειοψηφικής όσο και της μειοψηφικής κλάσης ήταν ικανοποιητική, κάτι που αντανακλάται στις τιμές όλων των κρίσιμων μετρικών, οι οποίες φανερώνουν ισχυρή ισορροπία μεταξύ της ανίχνευσης των θετικών περιπτώσεων και του περιορισμού των ψευδών συναγερμών.



Εικόνα 35. Confusion Matrix XGBoost (SMOTE)

Η μήτρα σύγχυσης του αλγόριθμου (Εικόνα 35) υποδεικνύει:

- 4777 σωστές προβλέψεις της αρνητικής κλάσης (True Negatives)
- 437 σωστές προβλέψεις της θετικής κλάσης (True Positives)
- 101 False Positives (σήματα θορύβου ταξινομήθηκαν εσφαλμένα ως pulsars)
- 55 False Negatives (πραγματικά pulsars που δεν αναγνωρίστηκαν)

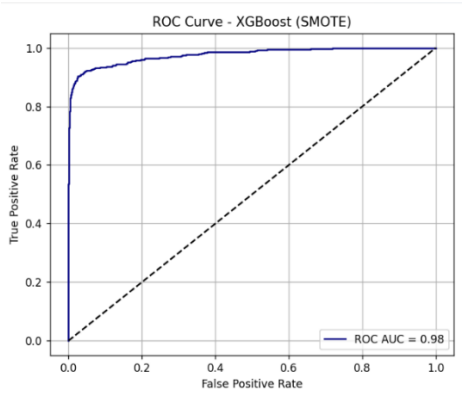
```
=== Αποτελέσματα για XGBoost (SMOTE) ===
              precision    recall  f1-score   support

           0     0.9886    0.9793    0.9839      4878
           1     0.8123    0.8882    0.8485       492

    accuracy                         0.9709      5370
   macro avg     0.9004    0.9338    0.9162      5370
weighted avg     0.9725    0.9709    0.9715      5370

Accuracy             : 0.9709
Balanced Accuracy    : 0.9338
ROC AUC              : 0.9770
Precision-Recall AUC : 0.9245
```

Εικόνα 36. Results XGBoost (SMOTE)

Στην οθόνη των αποτελεσμάτων (Εικόνα 36) η τιμή της *accuracy* παρατηρείται να φτάνει το **97.09%**, ποσοστό που φανερώνει τη γενικά επιτυχημένη πρόβλεψη και των δύο κλάσεων. Η ακόμη πιο αξιόπιστη σε προβλήματα ανισορροπίας κλάσεων *balanced accuracy* ανέρχεται στο **93.38%**, επιβεβαιώνοντας ότι το μοντέλο δεν υπερεστιάζει στη συχνότερη κλάση.

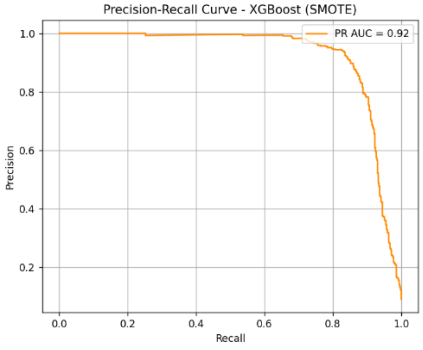
Αξιολογώντας επιμέρους μετρικές ανά κατηγορία, διαπιστώνεται ότι για την αρνητική κλάση (σήματα θορύβου), η *precision* είναι ιδιαίτερα υψηλή (**0.9886**), όπως και η *recall* (**0.9793**), οδηγώντας σε f1-score ίσο με **0.9839**. Αυτό σημαίνει πως το μοντέλο καταφέρνει να αναγνωρίζει σωστά την πλειονότητα των ψευδών σημάτων, με ελάχιστες λανθασμένες ταξινομήσεις. Όσον αφορά στη θετική κλάση (πραγματικά σήματα αστέρων νετρονίων), η *precision* φτάνει το **81.23%**, το οποίο υποδηλώνει ότι περίπου 8 στις 10 θετικές προβλέψεις είναι ορθές, ενώ η *recall* εμφανίζεται στο **88.82%**, δηλαδή το μοντέλο εντοπίζει σχεδόν 9 στους 10 πραγματικούς αστέρες. Το f1-score για την θετική κλάση διαμορφώνεται με αυτόν τον τρόπο στο **0.8485**.

Στους λιγότερο κρίσιμους δείκτες η macro average των μετρικών (δηλαδή ο μέσος όρος των τιμών ανά κλάση χωρίς στάθμιση ως προς τη συχνότητά τους) είναι επίσης ενδεικτική της συνολικής ποιότητας του ταξινομητή: macro precision **0.9004**, recall **0.9338** και f1-score **0.9162**. Τιμές που, όπως κι οι αντίστοιχες των weighted averages, υπογραμμίζουν τη γενικά αξιόπιστη απόδοση του μοντέλου ανεξάρτητα του πλήθους των δειγμάτων ανά κατηγορία.



Εικόνα 37. ROC Curve XGBoost

Και στην περίπτωση του συγκεκριμένου ταξινομητή η καμπύλη ROC είναι σχεδόν άριστη, πλησιάζοντας την επάνω αριστερή γωνία του διαγράμματος, γεγονός που υποδεικνύει την υψηλή διαχωριστική ικανότητα μεταξύ των δύο κλάσεων. Η αντίστοιχη τιμή του *ROC AUC* ανέρχεται στο **0.98**, κάτι που σημαίνει ότι το μοντέλο έχει πολύ μεγάλη πιθανότητα να διακρίνει σωστά μεταξύ ενός θετικού και ενός αρνητικού παραδείγματος (Εικόνα 37).



Εικόνα 38. Precision-Recall Curve XGBoost

Η Precision-Recall καμπύλη ενισχύει περαιτέρω τη θετική εικόνα, επιβεβαιώνοντας την αποτελεσματικότητα του ταξινομητή σε συνθήκες ανισορροπίας κλάσης. Από τη μορφή της προκύπτει ότι το μοντέλο διατηρεί υψηλή ακρίβεια (precision) για ευρύ φάσμα τιμών recall, κάτι που είναι ιδιαίτερα σημαντικό σε εφαρμογές όπου η λανθασμένη πρόβλεψη θετικού σήματος (false positive) πρέπει να ελαχιστοποιείται. Η περιοχή κάτω από την καμπύλη (PR AUC) ανέρχεται στο **0.9245**, αντανακλώντας τη συνολικά ισχυρή επίδοση του μοντέλου και σε αυτό το πιο απαιτητικό πλαίσιο της ανισορροπίας κλάσεων (Εικόνα 38).

Fuzzy k-NN

- ### Περιγραφή αλγόριθμου

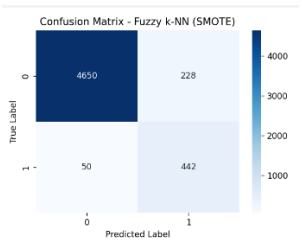
Ο Fuzzy k-Nearest Neighbors (Fuzzy k-NN) είναι μια εξελιγμένη εκδοχή του κλασικού αλγορίθμου k-NN, που χρησιμοποιείται ευρέως στη μηχανική μάθηση για την επίλυση προβλημάτων ταξινόμησης. Ο παραδοσιακός k-NN κατατάσσει κάθε νέο δεδομένο στην κατηγορία που εμφανίζεται πιο συχνά στους “k” κοντινότερους γείτονές του, δηλαδή σε εκείνα τα δεδομένα του συνόλου εκπαίδευσης που μοιάζουν περισσότερο με αυτό. Ωστόσο, αυτός ο αλγόριθμος παίρνει μια “σκληρή” απόφαση: επιλέγει μόνο μία κατηγορία, χωρίς να λαμβάνει υπόψη την πιθανότητα το νέο δείγμα να ανήκει μερικώς και σε άλλες.

Ο Fuzzy k-NN έρχεται να διορθώσει αυτή την αυστηρότητα, εισάγοντας την έννοια της αβεβαιότητας μέσω της ασαφούς λογικής (fuzzy logic). Σύμφωνα με αυτήν την προσέγγιση, ένα δείγμα δεν ανήκει αποκλειστικά σε μία κατηγορία, αλλά μπορεί να ανήκει σε πολλές ταυτόχρονα, με διαφορετικούς βαθμούς συμμετοχής (π.χ. 70% στην Κατηγορία A και 30% στην Κατηγορία B). Αυτό είναι ιδιαίτερα χρήσιμο όταν τα δεδομένα δεν είναι απόλυτα διαχωρίσιμα ή όταν υπάρχει επικαλυπτόμενη πληροφορία μεταξύ των κατηγοριών.

Η ευελιξία, η προσαρμοστικότητά του και η αναλυτική του ερμηνευσιμότητα, λόγω της περισσότερης πληροφορίας που παρέχει η έξοδός του σε σχέση με τον κλασικό κ-NN, τον καθιστούν αποτελεσματικό σε προβλήματα με επικάλυψη χαρακτηριστικών, ανθεκτικό στον θόρυβο και την ασάφεια και ιδανικό σε καταστάσεις που απαιτούν πιθανές εκτιμήσεις αντί για βεβαίες αποφάσεις. Από την άλλη μεριά, οι κύριοι περιορισμοί του είναι οι μεγάλες υπολογιστικές του απαιτήσεις, η εξάρτησή του από σύνθετες παραμετροποιήσεις και η προϋπόθεση καλά οργανωμένων δεδομένων εκπαίδευσης.

- ### Αποτελέσματα εκπαίδευσης

Η εκπαίδευση του αλγορίθμου Fuzzy k-Nearest Neighbors (Fuzzy k-NN) σε συνδυασμό με την τεχνική εξισορρόπησης δεδομένων SMOTE, επέστρεψε αποτελέσματα που αποδεικνύουν την ικανότητά του τόσο στην αντιμετώπιση προβλημάτων ανισορροπίας μεταξύ των κλάσεων όσο και στην πραγματοποίηση αξιόπιστης ταξινόμησης με έναν λιγότερο κλασικό τρόπο.



Εικόνα 39. Confusion Fuzzy k-NN (SMOTE)

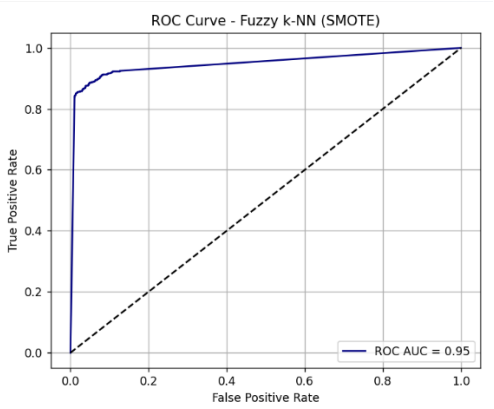
Η μήτρα σύγχυσης του αλγόριθμου (Εικόνα 39) υποδεικνύει:

- 4650 σωστές προβλέψεις της αρνητικής κλάσης (True Negatives)
- 442 σωστές προβλέψεις της θετικής κλάσης (True Positives)
- 228 False Positives (σήματα θορύβου ταξινομήθηκαν εσφαλμένα ως pulsars)
- 50 False Negatives (πραγματικά pulsars που δεν αναγνωρίστηκαν)

```
=== Αποτελέσματα για Fuzzy k-NN (SMOTE) ===
              precision    recall  f1-score   support

           0     0.9894    0.9533    0.9710      4878
           1     0.6597    0.8984    0.7608       492

    accuracy                         0.9482      5370
   macro avg     0.8245    0.9258    0.8659      5370
weighted avg     0.9592    0.9482    0.9517      5370

Accuracy             : 0.9482
Balanced Accuracy    : 0.9258
ROC AUC              : 0.9486
Precision-Recall AUC : 0.8100
```

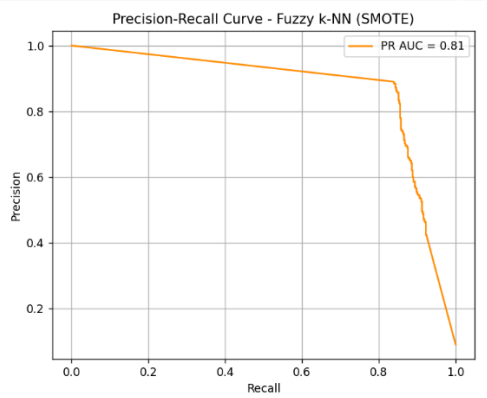
Εικόνα 40. Results Fuzzy k-NN (SMOTE)

Όπως υποδεικνύει ο πίνακας των αποτελεσμάτων, συνολικά το μοντέλο παρουσιάζει ακρίβεια (accuracy) **94.82%**, η οποία μαρτυρά την υψηλή του επίδοση στην ορθή ταξινόμηση του συνόλου των παραδειγμάτων, ανεξαρτήτως κλάσης. Η περισσότερο αξιόπιστη balanced accuracy, η οποία λαμβάνει υπόψη και την απόδοση ως προς την μειοψηφούσα τάξη, φτάνει το **92.58%**, ένδειξη ότι το μοντέλο είναι αρκετά ισορροπημένο ως προς την ικανότητά του να προβλέπει σωστά τόσο τις αρνητικές όσο και τις θετικές περιπτώσεις. Σε επίπεδο precision και recall, παρατηρείται διακύμανση μεταξύ των δύο κλάσεων. Για την αρνητική τάξη (κλάση 0), η precision ανέρχεται σε **98.94%**, ενώ η recall είναι **95.33%**, γεγονός που δείχνει ότι η πλειοψηφία των προβλέψεων για την αρνητική τάξη είναι ορθές. Αντίθετα, για τη θετική τάξη (κλάση 1), το precision μειώνεται στο **65.97%**, γεγονός που φανερώνει μεγαλύτερη δυσκολία του μοντέλου στο να αποφύγει ψευδώς θετικές προβλέψεις (FP). Ωστόσο, το recall για την ίδια τάξη κυμαίνεται στο **89.84%**, υποδεικνύοντας ότι το μοντέλο έχει υψηλή ικανότητα να εντοπίζει τις περισσότερες θετικές περιπτώσεις, πράγμα πολύ σημαντικό σε προβλήματα σαν κι αυτό της παρούσας έρευνας. Στο πεδίο των δευτερευουσών μετρικών, το macro average για precision, recall και f1-score είναι **82.45%**, **92.58%** και **86.59%** αντίστοιχα. Το weighted average, που προσμετρά και την επίδραση του μεγέθους κάθε τάξης, ανέρχεται σε **95.92%** για precision, **94.82%** για recall και **95.17%** για f1-score, υποδηλώνοντας πολύ καλή συνολική απόδοση σε ένα dataset, όπου κυριαρχεί η αρνητική τάξη (Εικόνα 40).



Εικόνα 41. ROC Curve Fuzzy k-NN

Η ROC καμπύλη εμφανίζεται έντονα κυρτή προς τα πάνω αριστερά, κάτι που αντανακλά υψηλή ευαισθησία (True Positive Rate) ακόμα και με σχετικά χαμηλά επίπεδα της ψευδώς θετικής πρόβλεψης (False Positive Rate). Η τιμή AUC (Area Under the Curve) που ανέρχεται στο **0.95** καταδεικνύει την ικανότητα του fuzzy k-NN να ξεχωρίζει με ακρίβεια μεταξύ των δύο τάξεων. Συνεπώς, η γενική εικόνα που παρέχει η καμπύλη ROC είναι ότι το μοντέλο κάνει πολύ λίγα λάθη όταν καλείται να αποφασίσει αν ένα παράδειγμα ανήκει στην κλάση της μειοψηφίας ή της πλειοψηφίας (Εικόνα 41).



Εικόνα 42. Precision-Recall Curve fuzzy k-NN

Η καμπύλη Precision-Recall συνιστά μια αρκετά ακριβέστερη γραφική αναπαράσταση της ταξινομητικής ικανότητας του μοντέλου. Η μορφή της καμπύλης δεικνύει υψηλά επίπεδα precision για σημαντικό εύρος των τιμών της recall, γεγονός που υποδηλώνει ότι το μοντέλο καταφέρνει να εντοπίζει μεγάλο μέρος των θετικών περιπτώσεων χωρίς να παράγει υπερβολικά πολλές ψευδώς θετικές προβλέψεις. Από την άλλη μεριά, η τιμή της PR AUC στο **0.81**, αποτελεί μια αρκετά καλή επίδοση για προβλήματα ανισορροπημένων δεδομένων και επιβεβαιώνει ότι το μοντέλο έχει μια αρκετά ισορροπημένη απόδοση στην ανακάλυψη των θετικών περιπτώσεων με ικανοποιητική ακρίβεια (Εικόνα 42).

```
=== Fuzzy Rules from Fuzzy k-NN Neighbors ===
Rule 1: IF Feature1 is Low AND Feature2 is Low AND Feature3 is High AND Feature4 is High AND Feature5 is Medium AND Feature6 is High AND Feature7 is Low AND Feature8 i
s Low THEN Class = 1
Rule 2: IF Feature1 is High AND Feature2 is Medium AND Feature3 is Low AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Medium AND Feature7 is Medium AND
Feature8 is Medium THEN Class = 0
Rule 3: IF Feature1 is Medium AND Feature2 is Medium AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is High AND Feature6 is High AND Feature7 is Low AND Fe
ature8 is Low THEN Class = 0
Rule 4: IF Feature1 is Medium AND Feature2 is Medium AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Low AND Feature7 is Medium AN
D Feature8 is Medium THEN Class = 0
Rule 5: IF Feature1 is Medium AND Feature2 is Medium AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Low AND Feature7 is High AND
Feature8 is High THEN Class = 0
Rule 6: IF Feature1 is High AND Feature2 is Medium AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Medium AND Feature7 is Medium A
ND Feature8 is Low THEN Class = 0
Rule 7: IF Feature1 is Medium AND Feature2 is High AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Medium AND Feature7 is Medium A
ND Feature8 is Low THEN Class = 0
Rule 8: IF Feature1 is Medium AND Feature2 is Medium AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Low AND Feature7 is Medium AN
D Feature8 is Medium THEN Class = 0
Rule 9: IF Feature1 is Medium AND Feature2 is Medium AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is High AND Feature6 is High AND Feature7 is Low AND Fe
ature8 is Low THEN Class = 1
Rule 10: IF Feature1 is Low AND Feature2 is Low AND Feature3 is Medium AND Feature4 is Medium AND Feature5 is Medium AND Feature6 is Low AND Feature7 is Medium AND Fea
ture8 is Medium THEN Class = 0
```

Εικόνα 43. Fuzzy rule extraction of fuzzy k-NN

Οι ταξινομητές ασαφούς λογικής σε αντίθεση με τα κλασικά μοντέλα ταξινόμησης, τα οποία λειτουργούν σε μεγάλο βαθμό σαν “μαύρα κουτιά” αναφορικά με την ερμηνεία του ταξινομητικού αποτελέσματος, παρέχουν τη δυνατότητα περαιτέρω ερμηνευσιμότητας του μοντέλου μέσω της διαδικασίας του λεγόμενου *fuzzy rule extraction*. Πρόκειται για κανόνες που περιγράφουν πώς οι τιμές των γνωρισμάτων οδηγούν σε μία συγκεκριμένη κατηγορία, χρησιμοποιώντας *ασαφείς γλωσσικούς όρους* (όπως "χαμηλό", "μεσαίο", "υψηλό") αντί για αυστηρά αριθμητικά όρια. Ένα τυπικό παράδειγμα τέτοιου είδους κανόνα, για ένα υποθετικό σύνολο δύο γνωρισμάτων, είναι: *«IF Feature1 is Low AND Feature2 is High THEN Class = 1»*. Κανόνες σαν κι αυτούς επιτρέπουν τη βαθύτερη *εννοιολογική κατανόηση της απόφασης* του

μοντέλου, μεγιστοποιούν τη διαφάνεια της ταξινομητικής διαδικασίας και ενδυναμώνουν την εξήγηση των αποτελεσμάτων. Συνεπώς καθίστανται σημαντικό εργαλείο στο πεδίο του machine learning και του data mining, ιδιαίτερα σε περιπτώσεις που το ενδιαφέρον εστιάζει στην ανακάλυψη λεπτομερέστερων ταξινομητικών μοτίβων, όπως συμβαίνει πολύ συχνά σε προβλήματα ταξινόμησης επιστημονικών παρατηρήσεων σαν και της παρούσας μελέτης.

Στην οθόνη των αποτελεσμάτων (Εικόνα 43) εμφανίζονται 10 τέτοιοι αντιπροσωπευτικοί κανόνες. Μια κάπως συνοπτική ερμηνεία των κανόνων αυτών είναι η ακόλουθη:

- **Rule 1 — Class = 1 (πραγματικός αστέρας νετρονίων):**

Το σήμα εμφανίζει σαφή παλμική δομή με υψηλή εκκεντρότητα (Profile_kurtosis) και ασυμμετρία (Profile _skewness) στο προφίλ, που είναι τυπική για πραγματικά σήματα αστέρων νετρονίων. Ταυτόχρονα, η DM-SNR curve εμφανίζεται στατιστικά ισχυρή, αλλά χωρίς υπερβολικές ανωμαλίες, με χαμηλή εκκεντρότητα (DM_kurtosis) και ασυμμετρία (DM_skewness). Αυτή η ασυμμετρία στο pulse profile, χωρίς υπερβολικά ακραίες τιμές στην DM-SNR καμπύλη, μοιάζει να αποτελεί ισχυρή ένδειξη πραγματικού παλμού.

- **Rule 2 — Class = 0 (Θόρυβος /RFI)**

Ένα σήμα με ομοιόμορφο και αδιάκριτο προφίλ (υψηλό Profile_mean χωρίς απότομες κορυφές) και χαμηλή εκκεντρότητα (skewness), είναι τυπικό για αποσπασματικό RFI ή λευκό θόρυβο. Η απουσία χαρακτηριστικής παλμικής δομής καθιστά το σήμα απίθανο να προέρχεται από αστέρα νετρονίων.

- **Rule 3 — Class = 0 (Θόρυβος/RFI)**

Οι υψηλές τιμές του μέσου όρου (mean) και της τυπικής απόκλισης (stdev) στη DM-SNR μπορεί να οφείλονται σε θορυβώδες σήμα ή RFI, χωρίς όμως ενδείξεις παλμού, αφού οι κατανομές είναι ουδέτερες (χαμηλή εκκεντρότητα και ασυμμετρία). Έτσι, παρά το έντονο ενεργειακά σήμα, η «υπογραφή» παλμικού αστέρα μάλλον απουσιάζει.

- **Rule 4 — Class = 0 (Θόρυβος/RFI)**

Όλα τα χαρακτηριστικά στο μέσο επίπεδο (medium), με χαμηλή stdev της DM-SNR curve συνεπάγεται σταθερότητα του DM-SNR καναλιού και υποδεικνύει απουσία παλμικής διαφοροποίησης, που συχνά χαρακτηρίζει θόρυβο στα ραδιοσήματα.

- **Rule 5 — Class = 0 (Θόρυβος/RFI)**

Μέσοι όροι στα περισσότερα features, χαμηλή DM_stdev και υψηλή εκκεντρότητα (DM_kurtosis) & ασυμμετρία (DM_skewness), διαμορφώνουν έναν παλμό που μοιάζει με θόρυβο και περιέχει εξάρσεις ή σποραδικές κορυφές (spikes), χωρίς όμως σταθερό παλμικό μοτίβο. Είναι πιθανός RFI, ο οποίος συχνά προκαλεί απότομες κορυφές στην DM-SNR curve χωρίς δομημένο προφίλ.

- **Rule 6 — Class = 0 (Θόρυβος)**

Το υψηλό Profile_mean σε συνδυασμό με τη χαμηλή DM_skewness και όλα τα υπόλοιπα features στο μέσο επίπεδο διαμορφώνουν μια *επίπεδη μορφή της DM-SNR curve* και *μη παλμικό προφίλ*, τα οποία υποδηλώνουν χαμηλή πληροφοριακή αξία του σήματος – τυπικό χαρακτηριστικό ασυνεχούς ή τυχαίου θορύβου.

- **Rule 7 — Class = 0 (Θόρυβος)**

Με υψηλή DM_stdev στο integrated profile, χαμηλή DM_skewness στο κανάλι της DM-SNR και όλα τα υπόλοιπα features στο μέσο επίπεδο, το σήμα μπορεί να εμφανίζει παλμικές εξάρσεις στο προφίλ αλλά χωρίς ασυμμετρία ή εκκεντρότητα στην DM-SNR, κάτι που οδηγεί στην απόρριψη του ενδεχομένου pulsar παλμού.

- **Rule 8 — Class = 0 (Θόρυβος)**

Όλα τα γνωρίσματα στο μεσαίο επίπεδο εκτός από την std της DM-SNR διαμορφώνουν ένα συμβατικό προφίλ για λευκό θόρυβο. Η απουσία ισχυρής μεταβολής σε όλες τις διαστάσεις οδηγεί σε ταξινόμηση ως μη παλμός.

- **Rule 9 — Class = 1 (Πραγματικός αστέρας νετρονίων)**

Με μέτρια όλα τα χαρακτηριστικά του integrated profile, υψηλά τα DM_mean και DM_stdev και χαμηλά τα DM_skewness και DM_kurtosis, το μοντέλο μοιάζει να εντοπίζει ισχυρό και καθαρό peak στην DM-SNR curve (υψηλή μέση τιμή και διασπορά) χωρίς παραμορφώσεις. Αυτό συχνά σημαίνει *σταθερό, ανιχνεύσιμο παλμό* – ένδειξη αστέρα νετρονίων.

- **Rule 10 — Class = 0 (Θόρυβος)**

Με χαμηλό mean και std στο integrated profile, χαμηλή std στη DM-SNR curve και όλα τα άλλα γνωρίσματα στο μέτριο επίπεδο, το σήμα παρουσιάζει χαμηλή ένταση και αστάθεια σε όλες τις μορφές, χωρίς διακριτό παλμό, πράγμα που ισοδυναμεί με τυπική περίπτωση μη εντοπίσιμου φυσικού αντικειμένου.

Οι παραπάνω 10 αντιπροσωπευτικοί κανόνες συνιστούν ένα συνοπτικό παράδειγμα των δυνατοτήτων που προσφέρει η ενσωμάτωση *ασαφούς λογικής* (fuzzy logic) σε αλγόριθμους ταξινόμησης. Ακόμη κι από αυτό το μικρό σύνολο κανόνων μπορεί αβίαστα να προκύψει ένα πρώτο συμπέρασμα για τη φύση των γνωρίσματων που ευνοούν την καταχώριση στην κλάση του σήματος αστέρα νετρονίων (class=1):

- Αναγνωρίζουν πραγματικά σήματα αστέρων νετρονίων ως αυτά που παρουσιάζουν *υψηλή εκκεντρότητα***,** *ασυμμετρία ή μεταβλητότητα*, κυρίως στην DM-SNR curve.
- Χαρακτηρίζουν θόρυβο ή RFI σήματα με χαμηλές ή μέσες τιμές των γνωρισμάτων, χωρίς σαφή κατανομή ή με αποσπασματικά peaks.

Τέτοιου είδους πληροφορία, ειδικά όταν παρέχεται από ακόμη περισσότερους κανόνες σαν κι αυτούς, μπορεί να οργανωθεί σε ερμηνευτικά διαγράμματα και να παράσχει ακόμη πιο λεπτές κατευθύνσεις προεπεξεργασίας των δεδομένων (κάτι που υπερβαίνει τα στενά όρια της παρούσας μελέτης) στη βάση των μοτίβων που αναδύονται από τη συμπεριφορά των γνωρισμάτων εντός των συγκεκριμένων κανόνων.

Συνοψίζοντας, το Fuzzy k-NN μοντέλο σε συνδυασμό με την τεχνική SMOTE, καταφέρνει να επιτύχει σημαντικά επίπεδα απόδοσης τόσο στο πεδίο της συνολικής ακρίβειας όσο και ως προς την ανίχνευση της μειοψηφούσας τάξης. Παρόλο που η ακρίβεια των προβλέψεων για τη θετική κλάση δεν είναι εξαιρετική η ικανότητα αναγνώρισής της (recall) είναι υψηλή, κάτι που καθιστά το μοντέλο κατάλληλο για περιπτώσεις (όπως αυτή της παρούσας μελέτης) όπου η μη αναγνώριση μιας θετικής περίπτωσης είναι πιο επιζήμια από μια λανθασμένη πρόβλεψη. Επίσης, καθώς αξιοποιεί την εννοιολογική ασάφεια και τη βαθμωτή συμμετοχή των χαρακτηριστικών για να διακρίνει μεταξύ πραγματικών σημάτων και θορύβου, παρέχει επιπλέον πληροφορία σχετικά με τη λογική της ταξινομητικής διαδικασίας που αδυνατούν

να παράσχουν οι κλασικοί αλγόριθμοι. Η ερμηνευσιμότητα μέσω των fuzzy rules επιτρέπει την κατανόηση του τρόπου λήψης απόφασης από το μοντέλο, κάτι που είναι σημαντικό σε εφαρμογές που απαιτούν υψηλή αξιοπιστία όπως η ανίχνευση αστρονομικών αντικειμένων. Έτσι, αν και αλγόριθμοι όπως ο XGBoost και ο Random Forest πέτυχαν εξαιρετικές επιδόσεις λόγω της ισχυρής δυνατότητας εκμάθησης μη γραμμικών σχέσεων, ο Fuzzy k-NN προσφέρει έναν ισορροπημένο συμβιβασμό μεταξύ ακρίβειας και ερμηνευσιμότητας, καθιστώντας το πολύτιμο εργαλείο για την κατηγοριοποίηση των αστρονομικών σημάτων.

Fuzzy Decision Trees

- Περιγραφή αλγόριθμου

Οι Fuzzy Decision Trees (FDTs) ή ασαφή δένδρα απόφασης, αποτελούν μια εξελιγμένη μορφή των παραδοσιακών δένδρων απόφασης. Ο κύριος στόχος αφορά στη λήψη απόφασης ή την πρόβλεψη καταστάσεων όταν τα δεδομένα είναι δύσκολο να ταξινομηθούν με απόλυτη σαφήνεια. Αντί να λειτουργούν με “σκληρούς” κανόνες του τύπου "αν το χαρακτηριστικό X είναι μεγαλύτερο από y, τότε...", χρησιμοποιούν μια πιο ανθρώπινη, "μαλακή", λογική που προσεγγίζει καλύτερα όψεις της καθημερινής σκέψης.

Στη φυσική μας γλώσσα, συχνά χρησιμοποιούμε όρους όπως “υψηλή θερμοκρασία”, “μέτρια πίεση”, “χαμηλός κίνδυνος” και άλλους παρόμοιους. Αυτοί οι όροι δεν έχουν ακριβή αριθμητικά όρια, αλλά εκφράζουν καταστάσεις υπό μια σχετική σκοπιά. Εκεί ακριβώς βασίζεται η ασαφής λογική (fuzzy logic), η οποία επιτρέπει σε ένα σύστημα να επεξεργάζεται τέτοιες περιγραφές και να λαμβάνει αποφάσεις βασισμένες σε "βαθμούς συμμετοχής" και όχι σε απόλυτα “ναι ή όχι”. Στο πλαίσιο των Fuzzy Decision Trees κάτι τέτοιο σημαίνει ότι κάθε παρατήρηση δεν κατηγοριοποιείται μονοσήμαντα σε κάποια από τις δύο τιμές του χαρακτηριστικού, αλλά μπορεί να είναι *μερικώς* και τα δύο. Για παράδειγμα, η τιμή 13 της πίεσης ενός ασθενούς μπορεί να θεωρηθεί “λίγο υψηλή” αλλά και “σχεδόν φυσιολογική” ταυτόχρονα. Αυτή η ευελιξία φέρνει τη λειτουργία του μοντέλου πιο κοντά στις μορφές των ανθρώπινων τρόπων λήψης αποφάσεων.

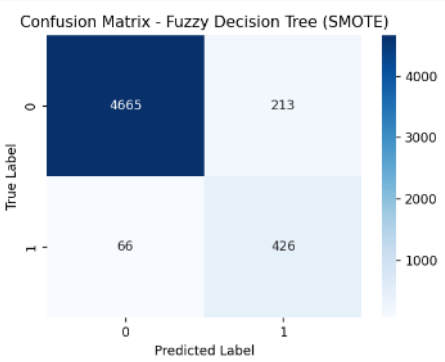
Το Fuzzy Decision Tree μοιάζει με ένα συνηθισμένο δέντρο απόφασης, όπου κάθε "κόμβος" αντιπροσωπεύει ένα χαρακτηριστικό και κάθε διαδρομή στο εσωτερικό του δέντρου καταλήγει σε μία απόφαση ή πρόβλεψη. Η διαφορά είναι ότι εδώ, αντί για αυστηρούς διαχωρισμούς (“είναι ή δεν είναι”), κάθε παρατήρηση μπορεί να “μοιράζεται” σε πολλαπλά σημεία του δέντρου με διαφορετικό βαθμό βαρύτητας. Αυτό επιτρέπει στο δέντρο να “ζυγίζει” καλύτερα την πληροφορία και να παράγει την ταξινόμηση των αποτελεσμάτων βάσει βαθμών συμμετοχής.

Ορισμένα σημαντικά πλεονεκτήματα των FDTs και του τρόπου που αυτά ενσωματώνουν την fuzzy λογική είναι: 1. Η ευέλικτη διαχείριση της αβεβαιότητας: Όπως κάθε αλγόριθμος που ενσωματώνει fuzzy logic, μετατρέπει την ασάφεια που τυχόν ενυπάρχει στα δεδομένα σε πλεονέκτημα για μια ταξινόμηση στη βάση βαθμών συμμετοχής. 2. Το μοντέλο πλησιάζει πολύ περισσότερο, σε σχέση με τους κλασικούς ταξινομητές, τον τρόπο που ανθρώπινος νους παράγει ταξινομήσεις. 3. Η αρκετά καλή ερμηνευσιμότητα των αποτελεσμάτων, καθώς και η προσαρμοστικότητα σε διαφορετικούς τύπους δεδομένων. Όσον αφορά στα μειονεκτήματά του τα κυριότερα είναι: 1. Η αυξημένη υπολογιστική πολυπλοκότητα συγκριτικά με τα απλά decision trees, η οποία απαιτεί αυξημένους πόρους συστήματος. 2. Η πολυπλοκότητα των συναρτήσεων συμμετοχής, η οποία απαιτεί εμπειρική ρύθμιση και μπορεί να οδηγήσει σε

λανθασμένη κατασκευή του αλγόριθμου. 3. Όπως και για τους περισσότερους ταξινομητές ασαφούς λογικής, δεν υπάρχει ένα ενιαίο πλαίσιο τυποποίησης.

- Αποτελέσματα εκπαίδευσης

Το μοντέλο Fuzzy Decision Tree, εκπαιδευμένο με εφαρμογή της τεχνικής SMOTE για την αντιμετώπιση της έντονης ανισορροπίας του dataset της μελέτης, επέστρεψε αποτελέσματα που καταδεικνύουν τη σχετικά ικανοποιητική απόδοση, με ορισμένους περιορισμούς στην ανίχνευση των θετικών περιπτώσεων (πραγματικά σήματα αστέρων νετρονίων).



Εικόνα 44. Confusion Fuzzy Decision Tree (SMOTE)

Η μήτρα σύγχυσης του αλγόριθμου (Εικόνα 44) υποδεικνύει:

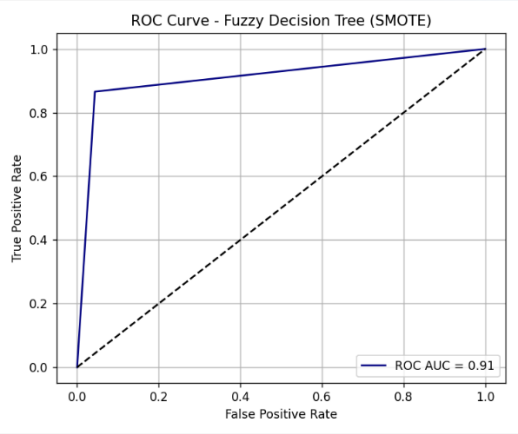
- 4665 σωστές προβλέψεις της αρνητικής κλάσης (True Negatives)
- 426 σωστές προβλέψεις της θετικής κλάσης (True Positives)
- 213 False Positives (σήματα θορύβου ταξινομήθηκαν εσφαλμένα ως pulsars)
- 66 False Negatives (πραγματικά pulsars που δεν αναγνωρίστηκαν)

```
=== Αποτελέσματα για Fuzzy Decision Tree (SMOTE) ===
              precision    recall  f1-score   support

           0     0.9860    0.9563    0.9710      4878
           1     0.6667    0.8659    0.7533       492

    accuracy                         0.9480      5370
   macro avg     0.8264    0.9111    0.8621      5370
weighted avg     0.9568    0.9480    0.9510      5370

Accuracy             : 0.9480
Balanced Accuracy    : 0.9111
ROC AUC              : 0.9111
Precision-Recall AUC : 0.5895
```

Εικόνα 45. Results Fuzzy Decision Tree (SMOTE)

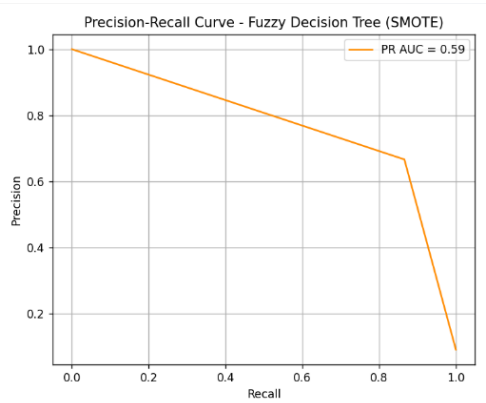
Οι τιμές των μετρικών απόδοσης που επιστρέφονται στην οθόνη των αποτελεσμάτων (Εικόνα 45) αντανακλούν τη σχετική ταξινομητική επάρκεια του μοντέλου, αλλά και κάποιους σοβαρούς περιορισμούς του. Το ποσοστό ακρίβειας (accuracy) του μοντέλου ανέρχεται στο **94.80%**, υποδηλώνοντας ότι ένα σημαντικό ποσοστό των παρατηρήσεων ταξινομήθηκε σωστά. Από την άλλη μεριά, η ακριβέστερη σε προβλήματα ανισορροπίας balanced accuracy φτάνει το **91.11%**, παρέχοντας μια πιο αξιόπιστη εικόνα της συνολικής επίδοσης του μοντέλου τόσο στην κλάση των θετικών όσο και στην κλάση των αρνητικών παραδειγμάτων. Ειδικότερα, για την αρνητική κλάση (σήματα που προέρχονται από θόρυβο), η precision ανέρχεται στο **98.60%**, γεγονός που υποδηλώνει ότι το μοντέλο κάνει ελάχιστα λάθη όταν

προβλέπει ότι ένα σήμα δεν προέρχεται από αστέρα νετρονίων. Η recall της ίδιας κλάσης είναι επίσης υψηλή (**95.63%**), καταδεικνύοντας την ικανότητα του μοντέλου να ανιχνεύει αποτελεσματικά την πλειονότητα των αρνητικών παραδειγμάτων. Όμως, για την κρίσιμη θετική κλάση (πραγματικά σήματα αστέρων νετρονίων) το μοντέλο επιτυγχάνει precision **66.67%**, πράγμα που σημαίνει ότι 1 στις 3 προβλέψεις υπέρ της κλάσης 1 είναι λανθασμένη. Ωστόσο, η recall στην ίδια κλάση αγγίζει το **86.59%**, υποδηλώνοντας πως η πλειονότητα των θετικών παραδειγμάτων εντοπίζεται με επιτυχία. Το f1-score για την κλάση 1 ανέρχεται σε **75.33%**, αντικατοπτρίζοντας μια σχετικά ισορροπημένη (αν και όχι κορυφαία) απόδοση σε precision και recall.



Εικόνα 46. ROC Curve Fuzzy Decision Tree

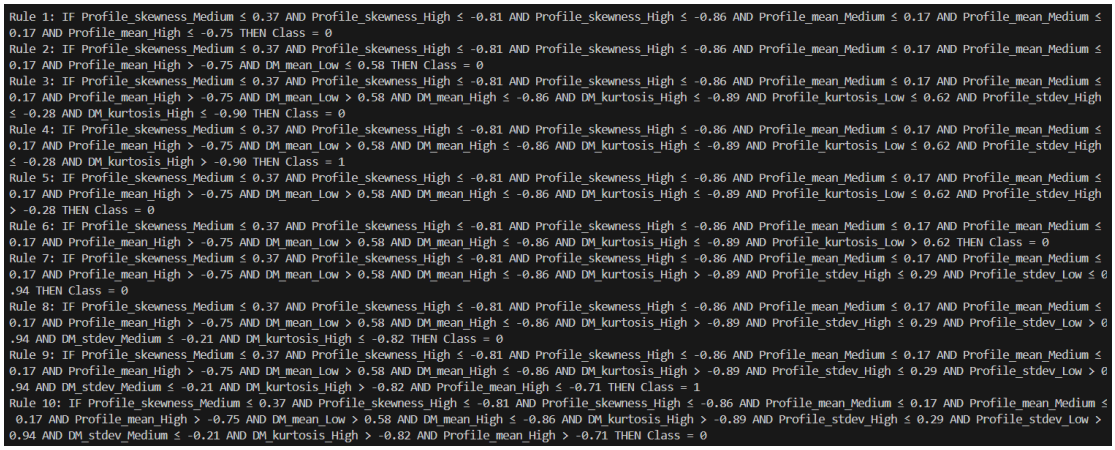
Η καμπύλη ROC (Εικόνα 46) έχει καλή καμπυλότητα προς το άνω αριστερό της άκρο (τρίγωνο), γεγονός που υποδηλώνει ισχυρή διακριτική ικανότητα του μοντέλου μεταξύ των θετικών (σήματα αστέρων νετρονίων) και αρνητικών περιπτώσεων (θόρυβος /RFI). Η καλή διαχωριστική ικανότητα του μοντέλου επιβεβαιώνεται και από την τιμή της AUC, η οποία αγγίζει το **0.91**. Γεγονός που σημαίνει ότι το μοντέλο έχει 91% πιθανότητα να διαχωρίσει σωστά ένα τυχαίο ζεύγος θετικού και αρνητικού δείγματος.



Εικόνα 47. Precision-Recall Curve Fuzzy Decision Tree

Η καμπύλη Precision-Recall αποτυπώνει την ποιότητα αναγνώρισης της μειοψηφικής τάξης μέσω της αναπαράστασης της αλληλεπίδρασης μεταξύ precision (θετικές προβλέψεις που είναι σωστές) και recall (ποσοστό των πραγματικών θετικών που ανιχνεύονται). Η μορφή της συγκεκριμένης καμπύλης (Εικόνα 47) δείχνει ότι η precision καθίσταται μη ικανοποιητική όταν το μοντέλο προσπαθεί να διατηρήσει υψηλό recall. Η αδυναμία αυτή του μοντέλου αποτυπώνεται και στην τιμή της PR AUC, η οποία δεν ξεπερνά το **0.59**, υποδηλώνοντας πως

παρότι το μοντέλο διαχωρίζει σωστά σε γενικές γραμμές τις δύο τάξεις, η ικανότητά του να εντοπίζει με ακρίβεια πραγματικά σήματα pulsars χωρίς την ταυτόχρονη εμφάνιση πολλών ψευδώς θετικών είναι περιορισμένη.



Εικόνα 48. Fuzzy rule extraction of Fuzzy Decision Tree

Παρά την αδυναμία του μοντέλου στην αναγνώριση της μειοψηφικής κλάσης, συγκριτικά με τους άλλους αλγόριθμους που εκπαιδεύτηκαν στο πλαίσιο της παρούσας εργασίας, θα μπορούσε να αξιοποιηθεί συμπληρωματικά ή υποστηρικτικά, ειδικά εκεί όπου απαιτείται λεπτομερής *ερμηνευσιμότητα* των αποτελεσμάτων. Όπως και ο fuzzy k-NN, μέσω της *fuzzy rule extraction* λειτουργίας, μπορεί να ερμηνεύσει λεπτομερέστερα τη διαδικασία λήψης της ταξινομητικής απόφασης, συγκριτικά με τους κλασικούς ταξινομητές.

Ακόμη και οι 10 αντιπροσωπευτικοί κανόνες που επιστρέφει η εκπαίδευση του μοντέλου στο πλαίσιο της παρούσας έρευνας (Εικόνα 48), παρέχουν αρκετά σημαντική πληροφορία αναφορικά με τη διαδικασία της ταξινόμησης, η οποία συνοψίζεται στα ακόλουθα:

- Οι περισσότεροι από τους κανόνες βασίζονται στις μεταβλητές Profile_skewness και Profile_mean, υποδεικνύοντας ότι τα χαρακτηριστικά του ολοκληρωμένου προφίλ σήματος (integrated profile) παίζουν καθοριστικό ρόλο στη διάκριση.
- Οι υπόλοιπες μεταβλητές αφορούν τα χαρακτηριστικά της καμπύλης DM-SNR, που σχετίζεται με την ένταση του σήματος σε συνάρτηση με τη διασπορά.
- Οι τιμές “Low”, “Medium”, “High” αντανακλούν τα fuzzy επίπεδα τιμών (σε ασαφή σύνολα) και τα όρια σύγκρισης προέρχονται από το fuzzification επί των αριθμητικών χαρακτηριστικών.

Κανόνες 1–3, 5–8, 10 (Class = 0 → Θόρυβος)

Αυτοί οι κανόνες διατυπώνουν σενάρια όπου οι παρατηρήσεις έχουν:

- Χαμηλή ή μέτρια τιμή ασυμμετρίας (Profile_skewness) καθώς και μέτρια μέση τιμή (Profile_mean) στο ολοκληρωμένο προφίλ.
- Αδύναμο σήμα στο κανάλι της DM-SNR: Περιπτώσεις όπως *DM_mean_High ≤ -0.86* και *DM_kurtosis_High ≤ -0.89*, οι οποίες υποδηλώνουν φτωχή δομή και συγκέντρωση σήματος στα κρίσιμα μέτρα διασποράς.
- Περιορισμένη τυπική απόκλιση τόσο στα γνωρίσματα του προφίλ όσο και σε εκείνα της DM-SNR (Profile_mean, DM_stdev**)** και έντονη συμμετρία στο προφίλ, τα οποία συνιστούν τυπικά χαρακτηριστικά σε θορυβώδεις ή "τυχαίες" παρατηρήσεις.

Συμπερασματικά: Όλες οι παραπάνω περιπτώσεις αφορούν παρατηρήσεις που εμφανίζουν ομαλό – συμμετρικό προφίλ με ταυτόχρονη ασθενή συγκέντρωση σήματος, παρέχοντας μια συνολική εικόνα που το πιθανότερο είναι να αντιστοιχεί σε θόρυβο.

Κανόνες 4 και 9 (Class = 1 → Πραγματικό σήμα αστέρα νετρονίων)

Οι κανόνες αυτοί απομονώνουν παρατηρήσεις όπου:

- Η ασυμμετρία (skewness) και η κεντρική τάση (mean) του ολοκληρωμένου προφίλ είναι εντός συγκεκριμένων fuzzy ορίων, που δείχνουν ασύμμετρη ή ιδιαίτερη μορφή παλμού.
- Τα χαρακτηριστικά της DM-SNR αποκαλύπτουν σημαντική κορύφωση ή μη-τυπική κατανομή, πιθανότατα σχετιζόμενη με αυθεντικό παλμό από αστέρα νετρονίων.
- Σε συνδυασμό με άλλους περιορισμούς (όπως για παράδειγμα στις fuzzy τιμές των Profile_stdev_High & DM_stdev_Medium), αυτά τα μοτίβα μπορεί να υποδηλώνουν δομημένα, περιοδικά σήματα, όπως αυτά που προέρχονται από pulsar.

Συμπερασματικά: Οι συγκεκριμένη κανόνες απομονώνουν παρατηρήσεις με ιδιαίτερη δομή, έντονη κλίση και ενδεχομένως ανώμαλη κατανομή, οι οποίες είναι αρκετά πιο συμβατές με πραγματικά σήματα αστέρων νετρονίων.

Συνολικά οι παραπάνω κανόνες παρέχουν αρκετά διαφανή κατανόηση των αποφάσεων του μοντέλου, συμβάλλοντας στην ενδυνάμωση της ερμηνευσιμότητας των αποτελεσμάτων. Επίσης, ενισχύουν την ευκαμψία στη διαχείριση του προβλήματος: χάρη στην ασαφή λογική, οι κανόνες επιτρέπουν ανοχή σε αβεβαιότητα και θόρυβο, στοιχείο κρίσιμο σε προβλήματα παρατηρησιακών δεδομένων. Ενώ τέλος, λόγω της προσαρμοστικότητάς τους είναι εύκολο να ενσωματωθούν ή να τροποποιηθούν με βάση την εμπειρική γνώση.

Συμπερασματικό σχόλιο & προοπτικές μελλοντικής έρευνας

Η συγκριτική αξιολόγηση των έξι ταξινομητών που εκπαιδεύτηκαν στο πλαίσιο της παρούσας μελέτης αναδεικνύει ενδιαφέροντα συμπεράσματα τόσο ως προς την αποδοτικότητα των μοντέλων όσο και ως προς την ποιότητα της φύσης του dataset. Οι παραδοσιακοί αλγόριθμοι μηχανικής μάθησης – Logistic Regression, Random Forest, SVM και XGBoost – επέδειξαν ιδιαίτερα υψηλές επιδόσεις, με ακρίβειες που κυμαίνονται από **97.09%** έως **97.69%**, ενώ ταυτόχρονα διατήρησαν υψηλές τιμές recall και f1-score για την ενδιαφέρουσα κλάση (class 1 – pulsars).

Με βάση τα συγκεντρωτικά αποτελέσματα (Εικόνα 49), πιο αποτελεσματικός αλγόριθμος μοιάζει να αναδεικνύεται ο *Random Forest* (στη SMOTE εκδοχή). Ο αλγόριθμος αυτός πέτυχε την υψηλότερη συνολική ακρίβεια (**97.69%**), καθώς και πολύ ικανοποιητικές τιμές recall (**0.8882**) και f1-score (**0.8758**) για την κλάση των pulsars, γεγονός που υποδεικνύει ισχυρή ικανότητα εντοπισμού των θετικών περιστατικών χωρίς να γίνονται υπερβολικά πολλά false positives. Επιπλέον, το ROC AUC (**0.9707**) και το PR AUC (**0.9200**) επιβεβαιώνουν τη σταθερή και αξιόπιστη απόδοσή του σε διαφορετικά κατώφλια πιθανότητας. Αξίζει να σημειωθεί ότι η ισχυρή αποδοτικότητα των δέντρων απόφασης στο πεδίο του συγκεκριμένου προβλήματος συγκλίνει και με τα αποτελέσματα λεπτομερέστερων και αυστηρότερων ερευνών από αυτή που διεξήχθη στο πλαίσιο της παρούσας εργασίας.

Μοντέλο	**Accuracy**	**Balanced Accuracy**	**Precision (class 1)**	**Recall (class 1)**	**F1-score (class 1)**	**ROC AUC**	**PR AUC**
Logistic Regression	0.9723	0.9418	0.8135	0.9045	0.8566	0.9758	0.9252
Random Forest	0.9769	0.9370	0.8636	0.8882	0.8758	0.9707	0.9200
XGBoost	0.9709	0.9338	0.8123	0.8882	0.8485	0.9770	0.9245
SVM	0.9749	0.9478	0.8287	0.9146	0.8696	0.9710	0.8710
Fuzzy k-NN	0.9482	0.9258	0.6597	0.8984	0.7608	0.9486	0.8100
Fuzzy DT	0.9480	0.9111	0.6667	0.8659	0.7533	0.9111	0.5895

Εικόνα 49. Συγκεντρωτικός πίνακας αποτελεσμάτων εκπαίδευσης

Από την άλλη μεριά, ο *Logistic Regression*, αν και γραμμικό μοντέλο, παρουσίασε υψηλή recall (0.9045) και κορυφαία τιμή Precision-Recall AUC (0.9252), γεγονός που υποδηλώνει σταθερή απόδοση σε διάφορα κατώφλια πιθανότητας. Ο *SVM* ξεχωρίζει για την εξαιρετική ισορροπία μεταξύ precision και recall για την κλάση 1 (0.8287 και 0.9146 αντίστοιχα), καθώς και για την υψηλότερη τιμή balanced accuracy (0.9478), η οποία συνιστά κρίσιμο μέτρο σε προβλήματα με ανισόρροπες κλάσεις όπως το παρόν. Το *XGBoost*, γνωστό για την ικανότητά του να αξιοποιεί πολύπλοκες συσχετίσεις στα δεδομένα, διατηρεί επίσης υψηλές επιδόσεις (ROC AUC: 0.9770, PR AUC: 0.9245), γεγονός που επιβεβαιώνει την αποτελεσματικότητά του, έστω και με ελαφρώς μειωμένη precision έναντι του Random Forest. Συνολικά, όλοι οι κλασικοί αλγόριθμοι που εξετάστηκαν επέδειξαν αξιοσημείωτη σταθερότητα και συνέπεια στις μετρικές απόδοσης, επιβεβαιώνοντας την καταλληλόλητά τους για το πρόβλημα.

Όσον αφορά δε στους fuzzy αλγόριθμους, *Fuzzy k-NN* και *Fuzzy Decision Tree*, αν και υπολείπονται σε αρκετές μετρικές απόδοσης, παρέχουν αξιοσημείωτη προστιθέμενη αξία αναφορικά με την ερμηνευσιμότητα των αποτελεσμάτων. Ο Fuzzy k-NN εμφάνισε ακρίβεια 94.82% και balanced accuracy 0.9258, ενώ πέτυχε υψηλή recall για την κλάση 1 (0.8984),

αποδεικνύοντας την ικανότητά του να εντοπίζει τα περισσότερα θετικά περιστατικά, αν και με σαφώς χαμηλότερη precision (0.6597). Η συμπεριφορά αυτή είναι χρήσιμη σε σενάρια όπου προτιμάται η ευαισθησία έναντι της ακρίβειας, όπως για παράδειγμα στις περιπτώσεις που απαιτείται η ελαχιστοποίηση των false negatives. Από την άλλη, το Fuzzy Decision Tree παρουσίασε τις χαμηλότερες επιδόσεις συνολικά (PR AUC: 0.5895), ωστόσο επιτυγχάνει αξιοπρεπές f1-score (0.7533) για την κλάση 1 και balanced accuracy 0.9111, γεγονός που το καθιστά επίσης μια υπολογίσιμη επιλογή όταν η εξαγωγή κανόνων είναι προτεραιότητα.

Το κύριο πλεονέκτημα των fuzzy μοντέλων είναι η *διαφάνεια στη λήψη αποφάσεων* μέσω της εξαγωγής ασαφών κανόνων που είναι εύκολα ερμηνεύσιμοι από τον άνθρωπο – κάτι που είναι δύσκολο να επιτευχθεί με πιο "αδιαφανείς" αλγορίθμους όπως ο XGBoost ή ακόμη και ο Random Forest. Έτσι, η ενσωμάτωση των fuzzy μοντέλων στο pipeline της έρευνας μπορεί να λειτουργήσει συμπληρωματικά: τα ακριβέστερα μοντέλα παρέχουν προβλέψεις υψηλής απόδοσης, ενώ τα fuzzy μοντέλα προσφέρουν νοηματικές εξηγήσεις των χαρακτηριστικών που οδηγούν σε θετική ή αρνητική διάγνωση των παραμέτρων της πραγματικής κατάστασης. Αυτή η συνέργεια καθίσταται κρίσιμη σε επιστημονικά πεδία όπως η αστρονομία, όπου η διαφάνεια και η αιτιολόγηση των αποτελεσμάτων είναι εξίσου σημαντική με την αριθμητική ακρίβεια.

Τα χαρακτηριστικά που περιλαμβάνει το dataset, όπως οι στατιστικές περιγραφές (μέση τιμή, τυπική απόκλιση, μεταβλητότητα, κυρτότητα) του ολοκληρωμένου προφίλ (integrated profile) και οι αντίστοιχες της DM-SNR curve, αποδείχθηκαν εξαιρετικά αποτελεσματικά στην ταξινόμηση των σημάτων των pulsars. Η επιτυχία των διαφορετικών μοντέλων, ακόμα και με θεμελιωδώς διαφορετικές αρχιτεκτονικές, αναδεικνύει την *υψηλή πληροφορική αξία των επιλεγμένων γνωρισμάτων*, τα οποία αποτυπώνουν ουσιαστικές διαφοροποιήσεις μεταξύ αστρονομικού θορύβου και πραγματικών σημάτων. Συνεπώς, η ερευνητική πρόταση του να αποκτήσουν τα γνωρίσματα αυτά υπόσταση καθολικών μεταβλητών, ενισχύεται ουσιαστικά από τα αποτελέσματα της παρούσας μελέτης, η οποία επιβεβαιώνει την αξιοπιστία και την γενικευσιμότητά τους στη διαδικασία διαλογής σημάτων pulsars.

Όπως είναι φανερό από όσο προηγήθηκαν, η παρούσα μελέτη επικεντρώθηκε στην αξιολόγηση της απόδοσης, ορισμένων κλασικών και fuzzy αλγορίθμων ταξινόμησης για τον εντοπισμό πραγματικών σημάτων αστέρων νετρονίων (pulsars) έναντι ψευδών θετικών (false positive), χρησιμοποιώντας ένα καλά επιμελημένο και επιστημονικά τεκμηριωμένο σύνολο δεδομένων. Παρά τα σημαντικά αποτελέσματα που προέκυψαν από τα επιλεγμένα μοντέλα, υπάρχουν αρκετές πτυχές οι οποίες θα μπορούσαν να αποτελέσουν αντικείμενο μελλοντικής έρευνας και εμβάθυνσης, τόσο ως προς τη βελτίωση της ακρίβειας όσο και ως προς την ενίσχυση της ερμηνευσιμότητας και της γενίκευσης των μοντέλων.

Μια πρώτη κατεύθυνση αφορά στη *βελτιστοποίηση των fuzzy μοντέλων*. Στο πλαίσιο της παρούσας μελέτης, εφαρμόστηκαν σχετικά απλοί μηχανισμοί fuzzification, όπως χρήση της υπερβολικής εφαπτομένης ή ευθύγραμμη κανονικοποίηση βάσει αποστάσεων. Στο μέλλον, θα μπορούσαν να διερευνηθούν πολυπλοκότερες και προσαρμοστικότρες μεθοδολογίες καθορισμού ασαφών όρων και συναρτήσεων συμμετοχής, είτε με βάση στατιστικά κριτήρια είτε μέσω επιβλεπόμενης εκμάθησης. Παράλληλα, η *αυτόματη εξαγωγή και βελτιστοποίηση fuzzy κανόνων* με τεχνικές όπως το genetic rule learning, το rule pruning ή και η ελεγχόμενη μείωση πολυπλοκότητας θα μπορούσε να οδηγήσει σε μοντέλα πιο συνεκτικά και κατανοητά από τον άνθρωπο, χωρίς υποβάθμιση της απόδοσης. Μια δεύτερη σημαντική προοπτική αφορά στην *εξερεύνηση εναλλακτικών fuzzy αρχιτεκτονικών ή υβριδικών συστημάτων*, όπως νευροασαφείς ταξινομητές (neuro-fuzzy systems), Takagi–Sugeno fuzzy inference models ή

fuzzy ensemble learning, αλλά και δοκιμές με κλασικές νευρωνικές (deep learning) μεθόδους. Η ενσωμάτωση της fuzzy λογικής σε πολυεπίπεδα μοντέλα θα μπορούσε να γεφυρώσει την ερμηνευσιμότητα με την υπολογιστική ισχύ των εξαιρετικά αδιαφανών νευρωνικών τεχνικών μηχανικής μάθησης. Επιπλέον, ενδιαφέρον παρουσιάζει και η καθαρά *συγκριτική μελέτη των fuzzy ταξινομητών και των deep learning προσεγγίσεων*, ειδικά στα πολύ μεγάλης κλίμακας datasets που αναμένεται να αποφέρουν στο άμεσο μέλλον οι σύγχρονες παρατηρησιακές δυνατότητες. Και σε αυτό το σενάριο, η χαμηλή ερμηνευσιμότητα των νευρωνικών δικτύων θα μπορούσε ενδεχομένως να αντισταθμιστεί με την ενσωμάτωση fuzzy στοιχείων ή τεχνικών εξήγησης (explainability).

Τέλος, αξίζει να διερευνηθεί η μεταφορά και γενίκευση των μοντέλων σε άλλες μορφές αστρονομικών δεδομένων, πέραν των γνωρισμάτων της DM-SNR curve και του integrated profile. Με τον τρόπο αυτόν θα μπορούσε να ελεγχθεί περαιτέρω δυνατότητα να καταστούν τα χαρακτηριστικά αυτά καθολικές οντότητες για το σύνολο της μελλοντικής έρευνας στο πεδίο του προβλήματος του εντοπισμού των pulsars ή αν θα πρέπει να ενισχυθούν με νέα μετασχηματισμένα ή συνθετικά features, όπως π.χ. temporal-spectral pattern descriptors, statistical entropy measures ή μετασχηματισμοί βασισμένοι στη θεωρία των wavelets. Από την άλλη, η αξιολόγηση των μοντέλων σε περιβάλλοντα με *μη ελεγχόμενο αστρονομικό θόρυβο* θα μπορούσε επίσης να αναδείξει την προσαρμοστικότητά τους σε πιο αβέβαιες συνθήκες, γεγονός κρίσιμο για την εφαρμογή τους σε αυτόνομες διατάξεις ανίχνευσης και μεγάλης κλίμακας συστήματα ραδιο-τηλεσκοπίων.

Συνολικά, η παρούσα μελέτη θα μπορούσε να ιδωθεί ως ένα μικρό σημείο εκκίνησης για περαιτέρω διερεύνηση μοντέλων ταξινόμησης με ερμηνευσιμότητα και προσαρμοστικότητα, υποδεικνύοντας ότι η αξιοποίηση της ασαφούς λογικής μπορεί να εμπλουτίσει σημαντικά τη διαδικασία διαλογής αστρονομικών σημάτων σε μελλοντικά έργα μεγάλης κλίμακας.

Βιβλιογραφία

[1] Géron, A., 2022. *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* (3rd ed.). O'Reilly Media.

[2] Haensel P., Potekhin A. Y., Yakovlev D. G., 2007. *Neutron Stars 1: Equation of State and Structure*. Springer, Astrophysics and Space Science Library.

[3] Janert P. K., 2011. *Data Analysis with Open-Source Tools*. O'Reilly Press, Cambridge Massachusetts.

[4] Lyon R. J., Stappers B. W., Cooper S., Brooke J. M., Knowles J. D., 2016. "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach." *Monthly Notices of the Royal Astronomical Society*.

[5] Pang-Ning T., Steinbach M., Kumar V., 2018. *Εισαγωγή στην Εξόρυξη Δεδομένων*. Εκδόσεις Τζιόλα, Αθήνα.

[6] Philippakis M., 2019. *Θεωρία Πιθανοτήτων και Στοιχεία Στατιστικής Ανάλυσης*. Εκδόσεις Τσότρας, Αθήνα.

[7] Raschka S., Mirjalili V., 2022. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2* (4th ed.). Packt Publishing, Birmingham, UK.

[8] VanderPlas J., 2016. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Sebastopol, CA.

[9] Wikipedia, 2024. *Classification (machine learning)*. Available at: https://en.wikipedia.org/wiki/Classification_(machine_learning) (Accessed: July 2025).

[10] Wikipedia, 2024. *Fuzzy logic*. Available at: https://en.wikipedia.org/wiki/Fuzzy_logic (Accessed: July 2025).

[11] Wikipedia, 2024. *Machine learning*. Available at: https://en.wikipedia.org/wiki/Machine_learning (Accessed: July 2025).

[12] Wikipedia, 2024. *Pulsar*. Available at: https://en.wikipedia.org/wiki/Pulsar (Accessed: July 2025).