Εθνικό Μετσόβιο Πολυτεχνείο
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
«Υπολογιστική Μηχανική»

Εφαρμογή της Μηχανικής Μάθησης στην Αντιμετώπιση Προβλημάτων της Διεργασίας Παραγωγής Αλουμινίου

Γιάννης Αντωνόπουλος

Επιβλέπων: Μ. Καβουσανάκης, Επικ. Καθηγητής

Οκτώβριος 2020

Ευχαριστίες

Με την εκπόνηση της παρούσας εργασίας ολοκληρώνεται η πορεία μου στο Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών «Υπολογιστική Μηχανική». Το τέλος αυτό θεωρώ ότι με βρίσκει πολύ πιο καταρτισμένο σε θεωρητικό και πρακτικό επίπεδο, τόσο πάνω στο αντικείμενο του ΔΠΜΣ όσο και σε άλλα, παράπλευρα αντικείμενα με τα οποία μου δόθηκαν η αφορμή και η ευκαιρία να ασχοληθώ λόγω του ΔΠΜΣ. Φυσικά, η διαδικασία της μάθησης δεν είναι ατομική υπόθεση, γι΄ αυτό και θέλω να ευχαριστήσω όλους όσους συνεισέφεραν στην επίτευξη αυτού του αποτελέσματος: τους καθηγητές και τους συμφοιτητές μου. Ιδιαίτερη αναφορά αξίζει στον επιβλέποντα αυτής της εργασίας, τον Επικ. Καθ. Μιχάλη Καβουσανάκη, ο οποίος πίστεψε σε μένα, με τίμησε με τη συνεργασία του και κατέβαλε τη μέγιστη δυνατή προσπάθεια ώστε οι κόποι μου να ευοδωθούν.

Ξεχωριστή αναφορά αρμόζει επίσης σε όλα τα μέλη του Εργαστηρίου Φυσικοχημείας και Εφαρμοσμένης Ηλεκτροχημείας, αρχίζοντας από τον Αναπ. Καθ. Αντώνη Καραντώνη και το (διπλωματούχο πλέον της Σχ. Χημικών Μηχανικών) Γιώργο Στέφα. Αφενός μεν με στήριξαν με κάθε δυνατό τρόπο κατά τη φοίτηση μου στο ΔΠΜΣ, αφετέρου δε - και αυτό είναι πολύ σημαντικότερο - θεωρώ ότι έχουν πολύ μεγάλη συνεισφορά στο ποιος είμαι σήμερα, άρα και σε όσα είμαι σε θέση να πετύχω. Και δεν πρόκειται να το ξεχάσω ποτέ.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου και τους κοντινούς μου ανθρώπους για τη συνεχή τους στήριξη και την υπομονή τους όλα αυτά τα χρόνια. Ο χρόνος μαζί τους είναι ανεκτίμητος και εγώ αποφάσισα να ξοδέψω ένα μέρος του χρόνου αυτού για να πετύχω κάτι. Κατά μία έννοια νομίζω ότι πιο πολύ θα έπρεπε να τους ζητήσω συγγνώμη παρά να τους ευχαριστήσω. Από τη μία εύχομαι να άξιζε, από την άλλη τους υπόσχομαι ότι θα επανορθώσω άμεσα.

Στο άτομο με τη μεγαλύτερη συνεισφορά στην επιτυχή ολοκλήρωση αυτής της εργασίας, των μεταπτυχιακών μου σπουδών αλλά και όχι μόνο δεν έχω να πω κάτι περισσότερο. Τα λόγια δεν είναι αρκετά.

Περίληψη

Στην παρούσα μελέτη επιδιώκεται η ανάπτυξη ενός συστήματος πρόβλεψης προβλημάτων που παρουσιάζονται κατά τη διεργασία παραγωγής αλουμινίου στο εργοστάσιο του Αλουμινίου της Ελλάδος, το οποίο ανήκει στον Τομέα Μεταλλουργίας του Ομίλου Μυτιληναίος. Αυτή η διεργασία είναι μια διεργασία ηλεκτρόλυσης, στην οποία αλουμίνιο παράγεται από αλουμίνα σε κατάλληλα ηλεκτρολυτικά κελιά, τα οποία καλούνται λεκάνες. Τα προβλήματα που ζητείται να προβλεφθούν είναι εξογκώματα στις ανόδους των ηλεκτρολυτικών αυτών κελιών, τα οποία καλούνται *μανιτάρια* και επηρεάζουν αρνητικά τη διεργασία της ηλεκτρόλυσης.

Έως τώρα, ένα σύστημα το οποίο έχει αναπτυχθεί από τους μηχανικούς της Δραστηριότητας Ηλεκτρόλυσης του εργοστασίου προβλέπει την ύπαρξη μανιταριών στις λεκάνες, εξετάζοντας το αν οι τιμές συγκεκριμένων μετρούμενων μεγεθών που χαρακτηρίζουν τις λεκάνες υπερβαίνουν προκαθορισμένα όρια. Τα όρια αυτά έχουν οριστεί από τους μηχανικούς με βάση μη αυτόματη στατιστική ανάλυση επί ενός περιορισμένου όγκου δεδομένων. Η διαφορά της προσέγγισης που ακολουθείται στην παρούσα μελέτη έγκειται στην αποτελεσματικότερη αξιοποίηση της πληροφορίας που περιέχεται στα δεδομένα μέσω τεχνικών Μηχανικής Μάθησης. Μια τέτοια προσέγγιση επιτρέπει την καλύτερη αυτοματοποιημένη εξαγωγή γνώσης από δεδομένα, η οποία με τη σειρά της θα έχει ως αποτέλεσμα τη βελτίωση της αντιμετώπισης του εξεταζόμενου προβλήματος στην πράξη.

Στη μελέτη επιλέγονται να εξεταστούν ηλεκτρικά μεγέθη τα οποία καταγράφονται αυτόματα από αισθητήρες, σε συνδυασμό με την πληροφορία περί παρουσίας μανιταριών, όπως αυτή προκύπτει από δειγματοληπτικούς ελέγχους. Πραγματοποιείται μια προεπεξεργασία των δεδομένων ώστε αυτά να έρθουν σε κατάλληλη μορφή, ενώ επιλέγεται τόσο η στρατηγική συνδυασμού των δεδομένων μεταξύ τους, αφού συνίστανται σε χρονοσειρές με διαφορετική συχνότητα δειγματοληψίας, όσο και η στρατηγική αποτύπωσης της πληροφορίας περί ιστορικής εξέλιξής τους. Επίσης αναπτύσσεται μια τακτική τροποποίησης των αποτελεσμάτων των δειγματοληπτικών ελέγχων για παρουσία μανιταριών πριν αυτά αναπτυχθούν, έτσι ώστε ο τρόπος με τον οποίο λαμβάνονται υπόψη να αντιστοιχεί περισσότερο στην πραγματική κατάσταση.

Τα δεδομένα αντιμετωπίζονται με τη χρήση δύο διαφορετικών αλγορίθμων Μηχανικής Μάθησης, των Δέντρων Αποφάσεων (Decision Trees) και Τυχαίων Δασών (Random

Forests). Για κάθε εκδοχή των δεδομένων και κάθε έναν από τους αλγόριθμούς εξετάζεται ένα ευρύ σύνολο από συνδυασμούς παραμέτρων. Το αποτέλεσμα κάθε προσπάθειας αξιολογείται με ένα κατάλληλο δείκτη αξιολόγησης, συγκεκριμένα του εμβαδού κάτω από την Χαρακτηριστική Καμπύλη Λειτουργίας Δέκτη (Receiver Operating Characteristic, ROC), το οποίο είναι γνωστό απλά ως AUC (εκ του Area Under the Curve). Όλοι οι υπολογισμοί της μελέτης, καθώς επίσης η προεπεξεργασία των δεδομένων αλλά και η ανάλυση των αποτελεσμάτων, πραγματοποιούνται με κώδικα που έχει συνταχθεί σε γλώσσα προγραμματισμού Python 3 και αξιοποιεί διάφορες βιβλιοθήκες, η πιο χρήσιμη εκ των οποίων είναι το scikit-learn.

Από τα παραπάνω προκύπτει ο βέλτιστος συνδυασμός παραμέτρων για κάθε εκδοχή των δεδομένων, καθώς και μια γενικότερη εικόνα της εξάρτησης της επίδοσης των μοντέλων από τις παραμέτρους. Προκύπτει επίσης το καλύτερο μοντέλο, το οποίο είναι βασισμένο στα Τυχαία Δάση, που γενικά υπερέχουν των Δέντρων Αποφάσεων. Η επίδοση του μοντέλου αυτού όσον αφορά το AUC είναι 0.72, έναντι του 0.5 που αντιστοιχεί στην τυχαία πρόβλεψη. Αν και η επίδοση χαρακτηρίζεται ως μέτρια, αποτελεί σημαντική βελτίωση έναντι της υφιστάμενης κατάστασης και αποτελεί ένα σημαντικό πρώτο βήμα στη σωστή κατεύθυνση, καθώς το μοντέλο μπορεί να χρησιμοποιηθεί για την πιο αξιόπιστη πρόβλεψη της παρουσίας μανιταριών, με αποτέλεσμα και τα διαθέσιμα δεδομένα στην πορεία να βελτιωθούν. Επιπλέον, η μελέτη είχε ως παράπλευρο αποτέλεσμα να βρεθούν πολλά στοιχεία που αφορούν στην μέτρηση και την καταγραφή των δεδομένων τα οποία μπορούν να βελτιωθούν, έτσι ώστε καλύτερα αποτελέσματα να προκύψουν χωρίς τροποποιήσεις στην μεθοδολογία που έχει υιοθετηθεί.

Abstract

The aim of the present study is to develop a system that is able to predict problems that occur during the aluminum production process at the Aluminium of Greece plant that belongs to the Metallurgy sector of Mytilineos Group. This is an electrolytic process, in which aluminum is produced from aluminum oxide in appropriate electrolytic cells. The problem sought to be predicted is the evolution of spikes, called *mushrooms*, on the anodes of the cells, which have a detrimental effect on the electrolytic process.

Until now, a system developed by the engineers of the plant's Electrolysis Division predicts the presence of mushrooms in the cells, by testing whether the values of certain measured variables that describe the condition of the cells exceed certain limits. These limits have been set by the engineers based on manual statistical analysis on a limited volume of data. The methodology followed in the present study differs in that the information contained in the data is exploited more efficiently by using Machine Learning techniques. Such an approach allows for a better and finally automated extraction of knowledge from data, which in turn will result in the amelioration of the way the problem under study is addressed.

In the study, certain electrical measures which are automatically recorded by sensors, along with information on the presence of mushrooms, acquired by sample checks, are chosen for examination. The data is preprocessed so that it is in the proper format, while a data combination strategy is devised to account for the fact that the data come in the form of time series of different frequencies. Furthermore, a way to extract the information on the evolution of each measure is also devised. Last but not least, the results of the sample checks for mushrooms are modified prior to further use, so that the way in which they are taken into account better corresponds to the true condition of the system.

The data are processed by two different Machine Learning algorithms, Decision Trees and Random Forests. For each version of the data and each one of the two algorithms, a wide range of parameters is investigated. The result of each attempt is evaluated using an appropriate evaluation metric, specifically the Area Under the Curve (AUC), which is, as the name implies, tha area under the Receiver Operating Characteristic (ROC) curve. Every computation in the study, as well as data preprocessing and result analysis are conducted with code written in the

Python programming language, making use of various libraries, the most important of which is scikit-learn.

After the described procedure, the best combination of parameters for each version of the data, as well as a general perspective of the effect of the parameters on model performance is acquired. The best model, based on Random Forests, is also found. In general, Random Forest-based models outperform the Decision Tree-based ones. The performance of the best model in terms of AUC is 0.72 and can be compared to 0.5, corresponding to random predictions. Even though this performance is considered mediocre, it presents a drastic improvement compared with the present condition. Furthermore, it is an important first step in the right direction, as this model can be employed to predict the presence of mushrooms more accurately compared to the present practice, leading to an improvement in the quality of available data. Finally, the study also helped identify many aspects of data acquisition and storage that may be improved, so that the better results are achieved without changes in the adopted methodology.

Περιεχόμενα

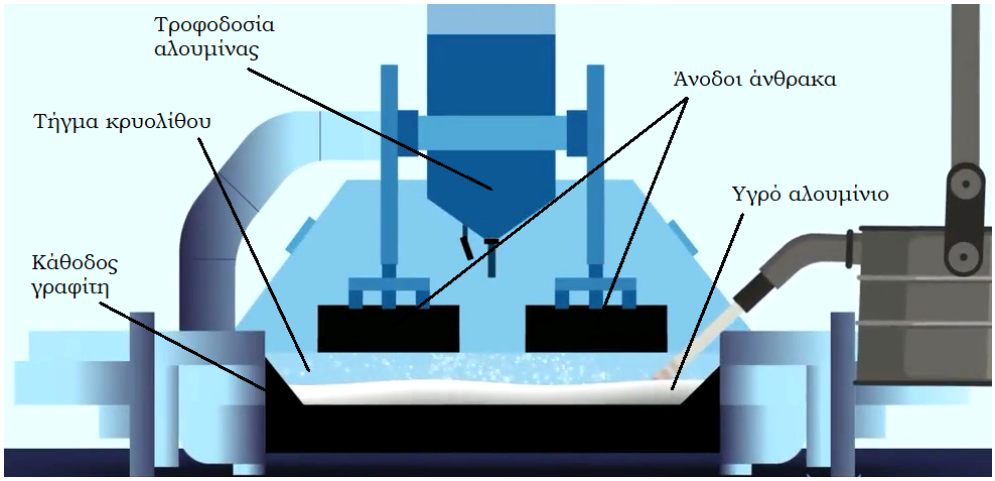
Κεφάλαιο 1

Εισαγωγή

1.1 Η Διεργασία Παραγωγής Αλουμινίου

Η διεργασία παραγωγής αλουμινίου αποτελεί μια ηλεκτρο-μεταλλουργική διεργασία μεγάλης κλίμακας, με απόδοση που αγγίζει τις μερικές δεκάδες δισεκατομμύρια τόνους πρωτόχυτου αλουμινίου ετησίως. Η επικρατούσα μεθοδολογία εφευρέθηκε από τους Hall και Héroult περί το 1885 και καλείται διεργασία Hall-Héroult προς τιμήν τους. Αποτελεί μια διεργασία ηλεκτρόλυσης, στην οποία ως ηλεκτρολύτης χρησιμοποιείται τήγμα κρυολίθου (Na_3AlF_6), ή για την ακρίβεια μίγμα φθοριούχου νατρίου (NaF) και φθοριούχου αλουμινίου (AlF_3), στη μοριακή αναλογία NaF/AlF_3 που απαντάται και στον κρυόλιθο, δηλαδή περίπου 2.3. Η αλουμίνα, Al_2O_3, που αποτελεί την πηγή αλουμινίου, διαλύεται στον ηλεκτρολύτη και διατηρείται σε περιεκτικότητα $2-4\%$ w/w. Επιπλέον, φθοριούχο ασβέστιο (CaF_2) σε περιεκτικότητα $\sim 5\%$ w/w χρησιμοποιείται ως πρόσθετο για την αύξηση της αγωγιμότητας του τήγματος.[1]

Οι τυπικές συνθήκες λειτουργίας του λουτρού είναι σε θερμοκρασία $\sim 950°C$ και πυκνότητα ρεύματος ~ 0.5 A/cm^2. Ως ηλεκτρόδια χρησιμοποιούνται άνοδοι άνθρακα που καταναλώνονται μετατρεπόμενες σε διοξείδιο του άνθρακα (CO_2) στην κορυφή του κελιού, το οποίο καλείται *λεκάνη*, και μία ενιαία κάθοδος γραφίτη στον πυθμένα του. Επειδή η πυκνότητα του παραγόμενου υγρού αλουμινίου είναι υψηλότερη από εκείνη του τήγματος, το προϊόν της ηλεκτρόλυσης συσσωρεύεται στον πυθμένα, ο οποίος μάλιστα απαιτείται να είναι καλυμμένος από υγρό αλουμίνιο για να μην λαμβάνει χώρα κυμάτωση και ανάμιξη στη διεπιφάνεια αλουμινίου/τήγματος. Για το λόγο αυτό, ο γραφίτης της καθόδου συχνά επικαλύπτεται με διβοριούχο τιτάνιο (TiB_2), το οποίο ενισχύει τη διαβροχή του από το υγρό αλουμίνιο.[1] Η λεκάνη ηλεκτρόλυσης και τα κύρια στοιχεία της απεικονίζονται σχηματικά στο Σχ. 1.1.



Σχήμα 1.1: Σχηματική απεικόνιση της λεκάνης ηλεκτρόλυσης και των κύριων στοιχείων της.

Οι αντιδράσεις που πραγματοποιούνται κατά τη διεργασία Hall-Héroult είναι οι εξής:

- Αντίδραση στην κάθοδο:

$$\mathrm{Al^{3+}} + 3\ \mathrm{e^-} \rightarrow \mathrm{Al} \tag{1.1}$$

- Αντίδραση στις ανόδους:

$$\mathrm{O^{2-}} + \mathrm{C} \rightarrow \mathrm{CO} + 2\ \mathrm{e^-} \tag{1.2}$$

ή

$$2\ \mathrm{O^{2-}} + \mathrm{C} \rightarrow \mathrm{CO_2} + 4\ \mathrm{e^-} \tag{1.3}$$

- Συνολική αντίδραση, δεδομένου ότι στις ανόδους παράγεται κυρίως $\mathrm{CO_2}$:

$$2\ \mathrm{Al_2O_3} + 3\ \mathrm{C} \rightarrow 4\ \mathrm{Al} + 3\ \mathrm{CO_2} \tag{1.4}$$

Η διατήρηση ομοιόμορφης κατανομής του ηλεκτρικού πεδίου στο ηλεκτρολυτικό κελί, καθώς και ομοιόμορφης πυκνότητας ρεύματος στα ηλεκτρόδια είναι απαραίτητη για την ομαλή λειτουργία της διεργασίας. Καταρχάς, η πυκνότητα ρεύματος εκφράζει το ρυθμό της ηλεκτροχημικής αντίδρασης, δηλαδή της παραγωγής αλουμινίου, ο οποίος προφανώς και επιδιώκεται να είναι σταθερός. Επιπλέον, οποιαδήποτε ανομοιομορφία στις παραπάνω κατανομές συνεπάγεται ανομοιογένεια άλλων μεταβλητών της διεργασίας, όπως π.χ. της τοπικής θερμοκρασίας ή της συγκέντρωσης των αντιδρώντων, η οποία με τη σειρά της οδηγεί σε απώλεια ελέγχου της διεργασίας και υποβάθμιση της απόδοσής της. Ένα από τα φαινόμενα που διαταράσσουν την ισορροπία στο κελί, το οποίο αποτελεί και το αντικείμενο της παρούσας μελέτης, είναι το ανοδικό πρόβλημα ανάπτυξης *μανιταριών*.

1.2 Το Ανοδικό Πρόβλημα Ανάπτυξης Μανιταριών

Ανοδικά προβλήματα ονομάζονται τα φαινόμενα εκείνα που σχετίζονται με την λειτουργία των ανόδων και έχουν αρνητικό αντίκτυπο στη διεργασία. Ένα από αυτά είναι και η ανάπτυξη εξογκωμάτων στην κατά τα άλλα επίπεδη επιφάνεια των ανόδων, λόγω διαφοράς στο ρυθμό κατανάλωσης στα διάφορα σημεία της επιφάνειας αυτής. Το φαινόμενο αυτό μπορεί να οφείλεται σε πολλές αιτίες, όπως π.χ. σε ανεπιθύμητες επικαθίσεις στην επιφάνεια, που έχουν ως αποτέλεσμα η άνοδος τοπικά να μην αντιδρά, ενώ περιμετρικά συνεχίζει να αντιδρά και να καταναλώνεται με αμείωτο ρυθμό. Τα εξογκώματα που προκύπτουν ονομάζονται *μανιτάρια* και επηρεάζουν αρνητικά τη διεργασία με διάφορους τρόπους, όπως π.χ. προσφέροντας στο ηλεκτρικό ρεύμα μια «συντομότερη» διαδρομή - ουσιαστικά μια διαδρομή χαμηλότερης αντίστασης - από την οποία ευνοείται η διέλευσή του. Η επιλεκτική διέλευση του ρεύματος από τις ανόδους στις οποίες έχουν αναπτυχθεί μανιτάρια έχει ως αποτέλεσμα την διατάραξη της ομοιομορφίας της κατανομής του πεδίου στη λεκάνη, με ο,τι αυτή συνεπάγεται για τη διεργασία της ηλεκτρόλυσης. Ένα μανιτάρι απεικονίζεται να εξέχει από τον πυθμένα μιας ανόδου στη φωτογραφία του Σχ. 1.2. Η άνοδος αυτή, μαζί με τις υπόλοιπες ανόδους της λεκάνης, έχουν απομακρυνθεί στο πλαίσιο εργασιών που πραγματοποιούνται στη λεκάνη. Το μανιτάρι διακρίνεται εύκολα γιατί έχει πυρώσει,

Σχήμα 1.2: Μανιτάρι που προεξέχει από τον πυθμένα ανόδου, η οποία έχει απομακρυνθεί από το εσωτερικό της λεκάνης στο πλαίσιο εργασιών στη λεκάνη.

λόγω της επιλεκτικής διέλευσης του ρεύματος από αυτό.

Το ανοδικό πρόβλημα ανάπτυξης μανιταριών είναι το πρόβλημα που επιδιώκεται να επιλυθεί για λογαριασμό της Δραστηριότητας Ηλεκτρόλυσης του εργοστασίου του Αλουμινίου της Ελλάδας στον Άγιο Νικόλαο Βοιωτίας. Για την κατανόηση των παρακάτω σημειώνεται πως οι λεκάνες της Δραστηριότητας Ηλεκτρόλυσης εντάσσονται - χωροταξικά - σε τρεις σειρές, με κωδικά ονόματα Α, Β και Γ. Κάθε σειρά περιλαμβάνει 260 λεκάνες, οι οποίες με τη σειρά τους αριθμούνται με τα νούμερα μεταξύ 101-165, 201-265, 301-365 και 401-465 (π.χ. Γ465).

Από τα παραπάνω γίνεται φανερό πως η απομάκρυνση των μανιταριών από τις ανόδους είναι ιδιαίτερα σημαντική για την ομαλή λειτουργία των λεκανών. Για αυτό το λόγο, συγκεκριμένοι εργαζόμενοι σε κάθε βάρδια είναι επιφορτισμένοι με το καθήκον του εντοπισμού και της αφαίρεσης μανιταριών. Στο έργο τους συνεισφέρει ένα σύστημα που ενημερώνει για πιθανή ανάπτυξη μανιταριού σε μια λεκάνη, όταν οι τιμές συγκεκριμένων μεγεθών ξεπεράσουν όρια προκαθορισμένα από το χρήστη, που δεν είναι άλλος από τους μηχανικούς παραγωγής της διεργασίας, με βάση την εμπειρία τους. Οι εργαζόμενοι μετρούν τις αντιστάσεις των ανόδων της ύποπτης λεκάνης με φορητό μιλλιβολτόμετρο και αν κάποια από αυτές είναι σημαντικά χαμηλότερη από τις υπόλοιπες, η αντίστοιχη άνοδος απομακρύνεται και ελέγχεται. Αν ο έλεγχος οδηγήσει σε εντοπισμό μανιταριού, τότε αυτό αφαιρείται μηχανικά και η άνοδος επανατοποθετείται. Τέλος, οι εργαζόμενοι καταχωρούν το πλήθος των μανιταριών που εντόπισαν σε ένα πίνακα που περιλαμβάνει το αντίστοιχο πεδίο για κάθε λεκάνη για τη βάρδιά τους.

Τα μεγέθη που αξιοποιούνται από το σύστημα που ενημερώνει για πιθανή ανάπτυξη μανιταριού είναι δύο, ο *θόρυβος* (ή *απόκλιση*), που αποτελεί το μέτρο της *αστάθειας*, και η *τιμή του τεστ*, το οποίο περιγράφεται στη συνέχεια. Σύμφωνα με τα τεχνικά έγγραφα που είναι διαθέσιμα από τη Δραστηριότητα Ηλεκτρόλυσης, ως *αστάθεια* ορίζεται *η ανομοιόμορφη κατανομή του ρεύματος στις ανόδους, η οποία έχει ως αποτέλεσμα σημαντικές κινήσεις του μετάλλου, που επιφέρουν μεταβολή της αντίστασης.* Το μέτρο της αστάθειας είναι ο *θόρυβος*, ο οποίος συμβολίζεται με W_{m} και υπολογίζεται από τη σχέση:

$$W_{\mathrm{m}} = R_{m_{\max}} - R_{m_{\min}}$$

όπου $R_{m_{\max}}$ και $R_{m_{\min}}$ η μέγιστη και η ελάχιστη στιγμιαία τιμή της αντίστασης της λεκάνης που καταγράφονται μέσα σε μια δεδομένη χρονική περίοδο, αντίστοιχα. Συγκεκριμένα, η αντίσταση της λεκάνης καταγράφεται από το μικροϋπολογιστή κάθε λεκάνης με συχνότητα 1 s και ο θόρυβος υπολογίζεται με συχνότητα 1 min = 60 s. Το μέγεθος αυτό προκύπτει ως το άθροισμα των αντιστάσεων όλων των μερών της λεκάνης από τα οποία διέρχεται το ηλεκτρικό ρεύμα, δηλαδή των ανόδων, του λουτρού, του υγρού μετάλλου και των καθόδων.

Η λεγόμενη *τιμή του τεστ* ή απλά *τεστ*, D_{r}, είναι ένα τεχνητό μέγεθος χωρίς φυσικό νόημα που υπολογίζεται από ηλεκτρικές μετρήσεις που πραγματοποιούνται στη λεκάνη. Το τεστ υπολογίζεται μόνο όταν για τη λεκάνη πληρούνται ορισμένες προϋποθέσεις, μια εκ των οποίων είναι π.χ. η λεκάνη να μη βρίσκεται σε καύση, που αποτελεί ένα άλλο πρόβλημα της διεργασίας. Όταν επιτρέπεται να υπολογιστεί το τεστ, το πλαίσιο στην κορυφή της λεκάνης στο οποίο συγκρατούνται οι άνοδοι πραγματοποιεί ορισμένες κινήσεις πάνω και κάτω και για κάθε θέση μετριέται η αντίσταση της λεκάνης. Με βάση αυτές τις μετρήσεις αντίστασης συναρτήσει της απόστασης (δηλαδή κατά μία έννοια ειδικής αντίστασης) και κάποιες εμπειρικές σταθερές της διεργασίας, προκύπτει αυτό το μέγεθος που επιτρέπει μια εκτίμηση της κατάστασης της λεκάνης, με βάση την εμπειρία των μηχανικών της διεργασίας. Αν και ο ακριβής τρόπος υπολογισμού του τεστ δεν κρίνεται σκόπιμο να αναφερθεί, σημειώνεται πως είναι ένα μέγεθος που βασίζεται σε ηλεκτρικές μετρήσεις και πως η συχνότητα με την οποία καταγράφεται είναι ανά 60 s. Φυσικά, η ανάγκη εκπλήρωσης των προϋποθέσεων για την μέτρησή του και ο χρόνος που απαιτείται για αυτήν έχει ως αποτέλεσμα η τιμή που καταγράφεται στο σύστημα να μεταβάλλεται πιο αραιά από ότι με αυτή τη συχνότητα.

Ενδεικτικές τιμές των ορίων που πρέπει να ξεπεράσουν οι τιμές των μεγεθών για να θεωρήσει το σύστημα πιθανή την ανάπτυξη μανιταριού είναι $W_{\mathrm{m}} < 0.21\ \mu\Omega$ και $D_{\mathrm{r}} > 1.5$. Οι τιμές αυτές τροποποιούνται ανάλογα με τις συνθήκες λειτουργίας της διεργασίας, οπότε είναι γενικά διαφορετικές μεταξύ των λεκανών των σειρών Α και Β και αυτών της σειράς Γ, που λειτουργούν διαφορετικά επί μονίμου βάσεως.

1.3 Ο Σκοπός της Μελέτης

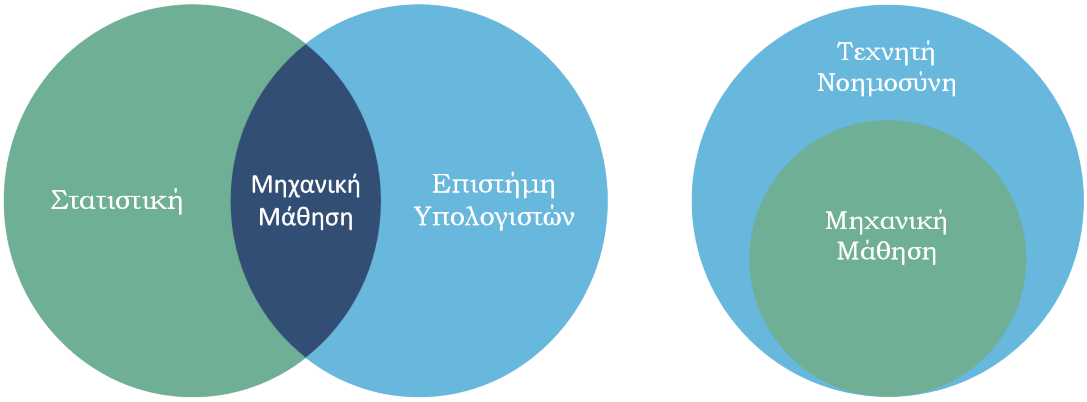
Η παρούσα μελέτη αποσκοπεί στην βελτίωση της υφιστάμενης κατάστασης στη διεργασία της Ηλεκτρόλυσης, μέσω της ανάπτυξης ενός νέου συστήματος ανίχνευσης της ύπαρξης μανιταριών στις λεκάνες. Από το σύστημα αυτό ζητείται να πραγματοποιεί πιο έγκαιρες προβλέψεις που να χαρακτηρίζονται από υψηλότερη αξιοπιστία και να μην απαιτούν την άμεση επέμβαση του ανθρώπινου παράγοντα για τη συντήρηση του. Φυσικά, για την ανάπτυξή του, εκτός από ιστορικά δεδομένα από τη διεργασία, επιβάλλεται να αξιοποιηθεί η εμπειρία των ειδικών, των οποίων όμως η ενεργή συμμετοχή δε θα πρέπει να είναι απαραίτητη μετά την έναρξη της λειτουργίας του συστήματος. Τα παραπάνω ζητούμενα φαίνεται να είναι δυνατό να επιτευχθούν εφαρμόζοντας μια προσέγγιση βασισμένη στη *Μηχανική Μάθηση*.

1.4 Μηχανική Μάθηση

Ο όρος *Μηχανική Μάθηση* (*Machine Learning*) παραπέμπει στην άντληση γνώσης από δεδομένα μέσω μιας αυτοματοποιημένης διαδικασίας που εκτελείται από ένα Η/Υ. Επίσης γνωστή ως *Στατιστική Μάθηση* (*Statistical Learning*), η Μηχανική Μάθηση αποτελεί διεπιστημονικό πεδίο στην τομή Στατιστικής και Επιστήμης Υπολογιστών, ενώ είναι υποσύνολο της Τεχνητής Νοημοσύνης (*Artificial Intelligence*). Η επιρροή της στον τρόπο με τον οποίο διεξάγεται η *καθοδηγούμενη από δεδομένα* (*data-driven*) έρευνα και λήψη αποφάσεων στην εποχή μας είναι τεράστια και αναμένεται να συνεχίσει να παρουσιάζει αύξουσα πορεία με την πάροδο του χρόνου.[2]

Στον αντίποδα της Μηχανικής Μάθησης εντοπίζονται τα πρώιμα «έξυπνα συστήματα» του παρελθόντος, τα οποία βασίζονταν σε επακριβώς ορισμένα σύνολα κανόνων με τη μορφή εντολών if-then για επεξεργασία και προσαρμογή στα δεδομένα εισόδου. Ως παράδειγμα μπορεί να αναφερθεί και το σύστημα που χρησιμοποιείται σήμερα για την αντιμετώπιση του προβλήματος με το οποίο ασχολείται και η παρούσα μελέτη. Συγκεκριμένα, το σύστημα αυτό ενημερώνει το χρήστη για πιθανή ανάπτυξη μανιταριού σε μία λεκάνη όταν οι τιμές δύο συγκεκριμένων μεγεθών ξεπεράσουν όρια προκαθορισμένα από τον ίδιο, με βάση τις συνθήκες λειτουργίας της λεκάνης αυτής. Εκτός από τη σχετικά χαμηλή αξιοπιστία που χαρακτηρίζει τις προβλέψεις του, το υφιστάμενο σύστημα παρουσιάζει δύο επιπλέον αδυναμίες που είναι κοινές σε όλα τα συστήματα αυτού του είδους:

- Η αρχική ανάπτυξη και η συντήρησή του επιτάσσουν την ύπαρξη ενός ειδικού με βαθιά γνώση και εμπειρία της διεργασίας.
- Ακόμη και ένας τέτοιος ειδικός δεν είναι σε θέση να αντιληφθεί όλες τις συσχετίσεις μεταξύ των διαφόρων δεδομένων μέσω της χρήσης εργαλείων Στατι-



Σχήμα 1.3: Σχετική θέση Μηχανικής Μάθησης και άλλων επιστημονικών πεδίων.

στικής Ανάλυσης, οπότε οι κανόνες του συστήματος είναι συνήθως ελλιπείς.

- Η λογική με βάση την οποία πραγματοποιεί τις εκτιμήσεις του εξαρτάται ισχυρά από τα χαρακτηριστικά της διεργασίας, με αποτέλεσμα ακόμη και μια μικρή τροποποίησή τους να δημιουργεί την ανάγκη επανασχεδιασμού του.

Αυτές τις αδυναμίες επιχειρεί να αντιδιαστείλει μια προσέγγιση Μηχανικής Μάθησης, όπου ένας αλγόριθμος μπορεί να γίνει ικανός να πραγματοποιεί αξιόπιστες προβλέψεις αν τροφοδοτηθεί με μια επαρκώς εκτενή συλλογή αξιόπιστων δεδομένων, ενώ η προσαρμογή του σε μια νέα πραγματικότητα αποτελεί μια σε μεγάλο βαθμό αυτοματοποιημένη διαδικασία.

1.5 Τα Εργαλεία της Μελέτης

Όπως έχει ήδη αναφερθεί, η Μηχανική Μάθηση είναι το διεπιστημονικό πεδίο που βρίσκεται στην τομή Στατιστικής και Επιστήμης Υπολογιστών. Επιπλέον, πρόκειται για την αυτοματοποιημένη άντληση γνώσης από δεδομένα από έναν Η/Υ. Συνεπώς, απαιτεί τη χρήση μιας γλώσσας προγραμματισμού για την σύνταξη κώδικα βασισμένου στον αλγόριθμο που περιγράφει την διαδικασία άντλησης γνώσης από τα δεδομένα. Στην παρούσα μελέτη χρησιμοποιείται ως γλώσσα η Python 3, σε συνδυασμό με βιβλιοθήκες της οι οποίες διευκολύνουν τον προγραμματισμό μερών του αλγορίθμου που σχετίζονται π.χ. με την ανάγνωση των δεδομένων ή την εξαγωγή των μοντέλων. Στη συνέχεια πραγματοποιείται μια συνοπτική αναφορά σε αυτά τα εργαλεία.

1.5.1 Python

Η *Python* είναι μια *γλώσσα προγραμματισμού υψηλού επιπέδου* και *γενικής χρήσης* (*high level, general purpose programming language*). Ως υψηλού επιπέδου χαρακτηρίζεται μια γλώσσα της οποίας η μορφή προσομοιάζει αυτή των ανθρώπινων γλώσσων παρά της γλώσσας μηχανής, που αντιλαμβάνεται ο υπολογιστής. Θεωρείται δε γλώσσα γενικής χρήσης επειδή μπορεί να αξιοποιηθεί για την επίτευξη πολλών διαφορετικών σκοπών, μεταξύ των οποίων ενδεικτικά αναφέρονται η προσομοίωση ενός φυσικού συστήματος ή η ανάπτυξη μιας *γραφικής διεπαφής χρήστη* (*Graphical User Interface, GUI*). Επιπλέον, είναι *αντικειμενοστρεφής* (*object-oriented*) γλώσσα, με ότι αυτό συνεπάγεται για τη δομή και τη λειτουργία της[1].[3, 4]

[1]Περαιτέρω ανάλυση του θέματος της αντικειμενοστρέφειας δεν κρίνεται σκόπιμη στο πλαίσιο μιας εισαγωγής στα εργαλεία της παρούσας μελέτης.

Περνώντας στα χαρακτηριστικά της *Python* που την καθιστούν κατάλληλη για την παρούσα μελέτη, σημειώνεται ότι γενικά είναι *διερμηνευόμενη* (*interpreted*), παρέχει όμως τη δυνατότητα *μεταγλώττισης σε πραγματικό χρόνο* (*just-in-time compilation*). Το πρώτο χαρακτηριστικό, που σημαίνει ότι ο κώδικας που είναι γραμμένος σε *Python* μεταγλωττίζεται σε γλώσσα μηχανής κατά την εκτέλεση και όχι πριν από αυτή - σε αντίθεση με τον κώδικα μιας *μεταγλωττισμένης* (*compiled*) γλώσσας - της προσδίδει δυναμικό χαρακτήρα και την καθιστά *διαδραστική* (*interactive*). Η διάδραση αυτή επιτυγχάνεται μέσω *διαδραστικών τερματικών* (*interactive terminals*) που απαντώνται σε εργαλεία όπως τα *Jupyter Notebooks*[2] και έχει τεράστια σημασία για την αποτελεσματική Ανάλυση Δεδομένων και εφαρμογή μεθόδων Μηχανικής Μάθησης. Αυτό οφείλεται στο ότι οι διαδικασίες αυτές είναι κατεξοχήν *επαναληπτικές* (*iterative processes*), δεδομένου ότι είναι καθοδηγούμενες από τα δεδομένα. Το δεύτερο χαρακτηριστικό, δηλαδή η δυνατότητα μεταγλώττισης σε πραγματικό χρόνο, επιτρέπει την γρηγορότερη εκτέλεση τμημάτων του κώδικα με κατάλληλη σύνταξη.[3, 4]

Συνοψίζοντας, η *Python* προσφέρει εύκολη στην κατανόηση σύνταξη, υψηλή διαδραστικότητα και αξιοσημείωτη ισχύ. Λόγω της χρησιμότητάς της σε εφαρμογές από διάφορα επιστημονικά πεδία, πολλοί χρήστες και προγραμματιστές έχουν ασχοληθεί με την ανάπτυξη βιβλιοθηκών που απλοποιούν την επίλυση πληθώρας προβλημάτων, ενώ είναι ανά πάσα στιγμή διαθέσιμοι να προσφέρουν βοήθεια με οποιαδήποτε δυσκολία ανακύπτει κατά τη χρήση αυτών. Επιπλέον, τόσο η *Python* όσο και οι βιβλιοθήκες της συνήθως διατίθενται δωρέαν, συχνά δε είναι και *ανοιχτού κώδικα* (*open source*). Έτσι η *Python* κατέληξε να είναι η πιο δημοφιλής γλώσσα γενικής χρήσης, αλλά και μία από τις δημοφιλέστερες γλώσσες για εφαρμογές Επιστήμης Δεδομένων και Μηχανικής Μάθησης, αντικείμενα για τα οποία διατίθενται εξαιρετικές βιβλιοθήκες. Όλα τα παραπάνω οδήγησαν στην επιλογή της συγκεκριμένης γλώσσας προγραμματισμού για την διεξαγωγή της παρούσας μελέτης.

1.5.2 scikit-learn

Σύμφωνα με τους δημιουργούς του, το *scikit-learn* είναι μια βιβλιοθήκη που περιέχει απλά και αποτελεσματικά εργαλεία *Προβλεπτικής Ανάλυσης* (*Predictive Analytics*). Η *Προβλεπτική Ανάλυση* δεν είναι παρά ένας ακόμη όρος που χρησιμοποιείται αντί της Μηχανικής Μάθησης.[2] Η βιβλιοθήκη είναι δωρεάν και ανοιχτού κώδικα, οπότε ο καθένας μπορεί να έχει πρόσβαση στον κώδικα για να ενημερωθεί για τον τρόπο λειτουργίας του ή για να τροποποιήσει τμήματά του όπως αυτός επιθυμεί για την προσωπική του, εμπορική ή μη, χρήση. Περιλαμβάνει σημαντικό πλήθος αλγορίθμων

[2] Τα *Jupyter Notebooks* αποτελούν βασικό εργαλείο της παρούσας μελέτης και αναλύονται εκτενώς παρακάτω.

Μηχανικής Μάθησης, ακόμη και αν αυτοί έχουν δημοσιευτεί πολύ πρόσφατα, καθώς και αναλυτική τεκμηρίωση και πρόσθετες πηγές για κάθε αλγόριθμο. Προφανώς αυτό είναι εφικτό χάρη στην ενεργή ομάδα που ασχολείται με τη συντήρηση και την ανάπτυξη της, όπως επίσης και στην εξίσου ενεργή κοινότητα χρηστών.[5, 6]

Όλα τα παραπάνω έχουν συνεισφέρει στη μεγάλη δημοφιλία του *scikit-learn* και στην εδραίωση του σαν το κυριότερο εργαλείο Μηχανικής Μάθησης στην *Python*. Αυτά τα χαρακτηριστικά, σε συνδυασμό με την εξαιρετική συνεργασία του με άλλες επιστημονικές βιβλιοθήκες, όπως π.χ. τη *NumPy* ή το *matplotlib*, πάνω στα οποία και έχει αναπτυχθεί, οδήγησαν στην επιλογή του για την παρούσα μελέτη.

1.5.3 NumPy

Η *NumPy* είναι μία από τις θεμελιώδεις, αν όχι η πιο θεμελιώδης, βιβλιοθήκη για την εκτέλεση *Επιστημονικών Υπολογισμών* (*Scientific Computing*) στην *Python*. Πάνω σε αυτήν έχουν στηριχθεί τόσο το *scikit-learn*, που έχει ήδη παρουσιαστεί, όσο και οι βιβλιοθήκες που θα παρουσιαστούν στη συνέχεια, *pandas*, *Matplotlib*, *seaborn* και *Jupyter*. Από μόνη της παρέχει ένα ισχυρό αντικείμενο *N-διάστατης σειράς* (*N-dimensional array*), που μπορεί να εκφράσει διανύσματα ή πίνακες σε όρους Γραμμικής Άλγεβρας, σε συνδυασμό με υψηλού επιπέδου μαθηματικές συναρτήσεις που, μεταξύ άλλων, επιτρέπουν χρήση Γραμμικής Άλγεβρας, *Γένεση Ψευδοτυχαίων Αριθμών* (*Random Number Generation, RNG*), Στατιστική Ανάλυση κ.ά. Τέλος, συνδυάζει υψηλού επιπέδου σύνταξη και υψηλή υπολογιστική ισχύ και ταχύτητα, επειδή βασίζεται σε κώδικα γραμμένο σε C, που είναι πολύ πιο κοντά στη γλώσσα μηχανής σε σχέση με τη βασική *Python*.[7]

1.5.4 pandas

Το *pandas* είναι μια γρήγορη, ισχυρή, ευέλικτη και εύκολη στη χρήση βιβλιοθήκη ανάλυσης και επεξεργασίας δεδομένων, προγραμματισμένη σε *Python*, αλλά με τμήματα γραμμένα σε *Cython* και *C* για επιτάχυνση. Παρέχει ένα ισχυρό αντικείμενο, το *pandas DataFrame*, το οποίο σε υψηλό βαθμό απλούστευσης θα μπορούσε να θεωρηθεί ότι είναι ένας πίνακας με παρόμοιες ιδιότητες π.χ. με ένα φύλλο εργασίας του Microsoft Excel. Το *pandas* παρέχει επίσης εργαλεία για ανάγνωση και εγγραφή δεδομένων, π.χ. από και σε αρχεία *csv*, συνδυασμό δεδομένων, διαχείριση κενών πεδίων, αναδιάταξη πινάκων, κατασκευή πινάκων Pivot, επιλογή και ομαδοποίηση δεδομένων και διαχείριση χρονοσειρών (π.χ. μετατόπιση ή αλλαγή συχνότητας). Όλα τα παραπάνω εργαλεία αξιοποιούνται εκτενώς στην παρούσα μελέτη, καθιστώντας το

pandas το πιο χρήσιμο εργαλείο για τη διεξαγωγή της, μαζί με το *scikit-learn*.[8]

1.5.5 Matplotlib

Η *Matplotlib* είναι η κυριότερη βιβλιοθήκη για επιστημονικά διαγράμματα σε *Python*. Επιτρέπει την παραγωγή διαγραμμάτων επιπέδου επιστημονικών δημοσιεύσεων για οπτικοποίηση κάθε είδους πληροφορίας, όπως π.χ. ιστογράμματα, ραβδογράμματα, διαγράμματα διασποράς κ.ο.κ. Τα σχήματα της *Matplotlib* είναι είτε στατικά, είτε διαδραστικά, γεγονός που τα καθιστά ιδανικά για Επιστήμη Δεδομένων και Μηχανική Μάθηση, οι οποίες είναι από τη φύση τους επαναληπτικές διαδικασίες.[9, 10]

Η βιβλιοθήκη αυτή παρέχει επίσης ειδικές λειτουργίες μέσω πακέτων επέκτασης κατασκευασμένων από τρίτους, όπως είναι π.χ. το *seaborn*, το οποίο αξιοποιείται και σε αυτή τη μελέτη. Ως βιβλιοθήκη που στηρίζεται στη *Matplotlib*, το *seaborn* δεν περιγράφεται σε ξεχωριστή ενότητα, συνοπτικά όμως αναφέρεται ότι είναι μια επέκταση για παραγωγή στατιστικών διαγραμμάτων με στόχο την εξερεύνηση και την κατανόηση σύνθετων δεδομένων μέσω κατάλληλων οπτικοποιήσεών τους.[11]

1.5.6 Jupyter

Τα τετράδια *Jupyter* (*Jupyter Notebooks*) είναι διαδραστικά γραφικά περιβάλλοντα προγραμματισμού, τα οποία επιτρέπουν τη χρήση διαδραστικού κώδικα, διαγραμμάτων, μαθηματικών εξισώσεων και απλού κείμενου ταυτόχρονα σε ένα παράθυρο οποιουδήποτε *web browser*. Πρόκειται για ένα εξαιρετικό εργαλείο το οποίο χρησιμοποιείται πολύ συχνά για ανάλυση και επεξεργασία δεδομένων, στατιστική μοντελοποίηση, μηχανική μάθηση και σε άλλα σχετικά πεδία.[12] Όλα όσα παρουσιάζονται στην παρούσα μελέτη έχουν προγραμματιστεί και εκτελεστεί στο περιβάλλον που παρέχεται από ένα τετράδιο *Jupyter*.

1.5.7 imbalanced-learn

Παρότι η χρήση του δεν παρουσιάζεται στο παρόν κείμενο, μια βιβλιοθήκη που αξιοποιήθηκε σε μεγάλο βαθμό κατά την εκτέλεση της μελέτης είναι το *imbalanced-learn*. Πρόκειται για μια βιβλιοθήκη που είναι συμβατή με το *scikit-learn* και μπορεί να χρησιμοποιηθεί σαν επέκτασή του. Παρέχει μια σειρά αλγορίθμων για τη *αναδειγματοληψία* (*resampling*) συνόλων δεδομένων στα οποία παρατηρείται έντονα το φαινόμενο των *Ανισοκατανεμημένων Τάξεων*. Το φαινόμενο αυτό συχνά δημιουργεί την ανάγκη

ειδικού χειρισμού των δεδομένων και συναντάται στα δεδομένα αυτής της μελέτης, οπότε στην πορεία το *imbalanced-learn* αξιοποιήθηκε στην προσπάθεια αντιμετώπισής του. Αν και τελικά προκρίθηκε μια προσέγγιση χωρίς αναδειγματοληψία, οφείλω να αναγνωρίσω τη συνεισφορά του συγκεκριμένου εργαλείου στη μελέτη και να αποδώσω τα δέοντα στους δημιουργούς του, καθώς και να ενημερώσω τον αναγνώστη για την ύπαρξή του.[13, 14]

Κεφάλαιο 2

Γενικές Έννοιες

Σε αυτό το κεφάλαιο του κειμένου αναπτύσσεται, κατά το δυνατόν πιο συνοπτικά, το θεωρητικό υπόβαθρο που απαιτείται για την κατανόηση της υπόλοιπης εργασίας. Συγκεκριμένα, εισάγονται ορισμένες γενικές έννοιες που προέρχονται από το πεδίο της Μηχανικής Μάθησης και αφορούν όχι μόνο την παρούσα μελέτη αλλά όλες τις αντίστοιχες, με αφετηρία αυτές τις *Επιτηρούμενης* και *Μη Επιτηρούμενης Μάθησης*.

2.1 Επιτηρούμενη και Μη Επιτηρούμενη Μάθηση

Στο υποπεδίο της Μηχανικής Μάθησης που είναι γνωστό ως *Επιτηρούμενη Μάθηση* (*Supervised Learning*), ο χρήστης παρέχει στον αλγόριθμο ζεύγη εισόδων και αντίστοιχων ορθών αποτελεσμάτων και ο αλγόριθμος καταλήγει σε μια συνάρτηση που παράγει αποτελέσματα από νέες, άγνωστες στη φάση της εκπαίδευσης, εισόδους. Στο τρέχον παράδειγμα, ο χρήστης θα παρείχε στον αλγόριθμο τιμές συγκεκριμένων μεγεθών ενδιαφέροντος, π.χ. των μεγεθών που αξιοποιεί και το υπάρχον σύστημα, μαζί με την πληροφορία του αν οι τιμές αυτές αντιστοιχούν σε ανάπτυξη μανιταριού και ο αλγόριθμος θα προέβλεπε την ύπαρξη μανιταριού για άλλες, παρόμοιες τιμές. Τα ορθά αποτελέσματα συχνά αναφέρονται ως ετικέτες (*labels*), ενώ το σύνολο των εισόδων που αντιστοιχούν σε μία ετικέτα ως *χαρακτηριστικά ή ιδιότητες* (*features*).[2]

Αυτού του είδους η Μηχανική Μάθηση καλείται *επιτηρούμενη* επειδή ο χρήστης, σαν ένας ιδιότυπος δάσκαλος, *επιτηρεί* τον αλγόριθμο τροφοδοτώντας τον με τη σωστή απάντηση για κάθε πρόβλημα με βάση το οποίο αυτός μαθαίνει. Αν και η συλλογή των κατάλληλων παρατηρήσεων για την εκπαίδευση συχνά αποτελεί μια απαιτητική, μη αυτοματοποιημένη διαδικασία, η λειτουργία τέτοιων αλγορίθμων είναι σχετικά εύκολο να κατανοηθεί και η επίδοση τους είναι δυνατό να ποσοτικοποιηθεί επακριβώς. Αυτό

ισχύει επειδή, μετά την εκπαίδευση, ο χρήστης μπορεί να δώσει στον αλγόριθμο εισόδους για τις οποίες είναι γνωστό το αποτέλεσμα και να μετρήσει τη συχνότητα με την οποία ο αλγόριθμος καταλήγει στο σωστό αποτέλεσμα. Στην περίπτωση όμως της παρούσας μελέτης, ακόμη το να συλλέξει κανείς τα δεδομένα δεν είναι ιδιαίτερα δύσκολο, αφού αυτά καταγράφονται από αισθητήρες ή εργαζόμενους και εισάγονται στη βάση δεδομένων της διεργασίας, από όπου μπορούν να ανακτηθούν.

Για λόγους πληρότητας αναφέρεται το έτερο ευρύ υποπεδίο της Μηχανικής Μάθησης, η *Μη Επιτηρούμενη Μάθηση*, η οποία αντιμετωπίζει περιπτώσεις όπου είναι γνωστές μόνο είσοδοι και όχι αποτελέσματα. Όπως είναι αναμενόμενο, στις περιπτώσεις αυτές το ζητούμενο δεν είναι αποτελέσματα αλλά π.χ. η ταξινόμηση των εισόδων σε ομάδες ή η ανίχνευση ιδιαίτερων περιπτώσεων, όπως άκυρων εγγραφών, στις εισόδους. Οι αντίστοιχοι αλγόριθμοι είναι δυσκολότερο να κατανοηθούν και να αξιολογηθούν σε σύγκριση με αυτούς της Επιτηρούμενης Μάθησης.[2] Ορισμένοι από αυτούς αξιοποιούνται σε συγκεκριμένα βήματα της παρούσας μελέτης, οπότε μια εκτενέστερη αναφορά σε εκείνους τους αλγορίθμους θα πραγματοποιηθεί παρακάτω.

2.2 Ταξινόμηση και Παλινδρόμηση

Από τα παραπάνω γίνεται κατανοητό ότι η Επιτηρούμενη Μάθηση είναι το υποπεδίο της Μηχανικής Μάθησης που ασχολείται με παρατηρήσεις οι οποίες φέρουν ετικέτες. Σε αυτό υπάγονται με τη σειρά τους η *Ταξινόμηση* και η *Παλινδρόμηση*, εκ των οποίων η πρώτη αποβλέπει στην πρόβλεψη μιας ετικέτας μεταξύ διακριτών επιλογών ενώ η δεύτερη στην πρόβλεψη μιας τιμής από το συνεχές σύνολο των πραγματικών αριθμών. [2] Χρησιμοποιώντας το αντικείμενο της παρούσας μελέτης ως παράδειγμα, πρόβλημα Ταξινόμησης αποτελεί η πρόβλεψη της ύπαρξης (ή μη) μανιταριού σε μια λεκάνη, ενώ Παλινδρόμησης η πρόβλεψη του πλήθους των μανιταριών που υπάρχουν στη λεκάνη.

Πιο συγκεκριμένα, το παραπάνω πρόβλημα Ταξινόμησης, που είναι και το πρόβλημα που επιχειρεί να επιλύσει η μελέτη, είναι πρόβλημα *Δυαδικής Ταξινόμησης*, δηλαδή πρόβλεψης μεταξύ δύο μόνο ετικετών. Εναλλακτικά, μπορεί να τεθεί ως ερώτημα που επιδέχεται θετική ή αρνητική απάντηση: «Υπάρχει μανιτάρι σε αυτήν τη λεκάνη;» Αυτά τα προβλήματα επιδέχονται διαφορετική αντιμετώπιση σε σχέση με εκείνα που απαιτούν πρόβλεψη μεταξύ πολλών ετικετών, δηλαδή *Πολυταξικής Ταξινόμησης*.[2]

Στη Δυαδική Ταξινόμηση συχνά γίνεται λόγος για *θετική* και *αρνητική* τάξη. Ο όρος *θετική* δεν υποδηλώνει ότι η συγκεκριμένη τάξη έχει κάποια αξία ή χρησιμότητα, υποδεικνύει απλά το ποιο είναι το αντικείμενο της μελέτης. Συνεπώς, η αντιστοίχιση του

σε μία από τις δύο τάξεις είναι υποκειμενική και εξαρτάται από το εκάστοτε πεδίο.[2] Στην παρούσα μελέτη, η θετική τάξη συνδέεται με την παρουσία μανιταριού, δηλαδή με ένα γεγονός επιζήμιο για την διεργασία, ενώ η αρνητική με απουσία μανιταριού.

2.3 Μη Ισοκατανεμημένα Δεδομένα

Ένα χαρακτηριστικό των συγκεκριμένων τάξεων, το οποίο μάλιστα είναι μάλλον ο κανόνας παρά η εξαίρεση σε προβλήματα Ταξινόμησης, είναι ότι δεν είναι ισοκατανεμημένες. Αυτό σημαίνει πως η μία τάξη, στην προκείμενη περίπτωση η αρνητική, εμφανίζεται με πολύ υψηλότερη συχνότητα σε σχέση με την άλλη. Τα δεδομένα στα οποία παρατηρείται αυτό το φαινόμενο ονομάζονται *Μη Ισοκατανεμημένα Δεδομένα* (*Imbalanced Datasets*) ή *Δεδομένα με Ανισοκατανεμημένες Τάξεις* (*Datasets with Imbalanced Classes*). Απαιτούν ιδιαίτερο χειρισμό, τόσο όσον αφορά την ανάλυση και την προκατεργασία των δεδομένων, όσο και την αξιολόγηση των μοντέλων που προκύπτουν από τους αλγορίθμους.[2]

Στο πλαίσιο αυτού του ιδιαίτερου χειρισμού εντάσσεται και η αξιοποίηση βιβλιοθηκών όπως το *imbalanced-learn*, η οποία αναφέρθηκε στην Ενότητα 1.5, που αφορούσε στα εργαλεία της μελέτης. Η βιβλιοθήκη αυτή παρέχει τη δυνατότητα *αναδειγματοληψίας* (*resampling*) των δεδομένων, η οποία περαιτέρω διακρίνεται σε *υπερδειγματοληψία* (*oversampling*) και *υποδειγματοληψία* (*undersampling*). Η αναδειγματοληψία γενικά αναφέρεται στην τεχνητή μεταβολή των συχνοτήτων στις οποίες απαντώνται οι διάφορες τάξεις στα δεδομένα, με σκοπό την εξισορρόπηση της διαφοράς τους είτε πλήρως, είτε μερικώς. Στην περίπτωση της υπερδειγματοληψίας αυξάνεται η συχνότητα εμφάνισης της πιο σπάνιας τάξης, ενώ στην αντίθετη περίπτωση, αυτή της υποδειγματοληψίας, μειώνεται η συχνότητα εμφάνισης της πιο κοινής τάξης.

Τα παραπάνω είναι δυνατό να πραγματοποιηθούν με σχετικά απλοϊκό τρόπο, π.χ. αναπαράγοντας αυτούσιες ορισμένες από τις παρατηρήσεις που αντιστοιχούν στην πιο σπάνια τάξη στην περίπτωση της υπερδειγματοληψίας, ή αφαιρώντας ορισμένες από αυτές που αντιστοιχούν στην πιο κοινή στην αντίθετη περίπτωση. Μπορεί όμως να γίνουν και με πιο σύνθετους αλγορίθμους, οι οποίοι δημιουργούν συνθετικά δεδομένα με βάση τις παρατηρήσεις της πιο σπάνιας τάξης στην πρώτη περίπτωση ή διατηρούν μόνο αντιπροσωπευτικές παρατηρήσεις στη δεύτερη. Για την ακρίβεια, στην δεύτερη περίπτωση, αυτή της υποδειγματοληψίας, είναι πιθανό επίσης να δημιουργήσουν συνθετικά δεδομένα, τα οποία εκφράζουν την τάση των αρχικών παρατηρήσεων και να απομακρύνουν τις ίδιες τις αρχικές παρατηρήσεις.

Οι προσεγγίσεις που περιγράφονται παραπάνω ελέγχθηκαν, ειδικά σε συνδυασμό με

αλγορίθμους Μηχανικής Μάθησης που είναι ευαίσθητοι στα Ανισοκατανεμημένα Δεδομένα, όπως π.χ. ο αλγόριθμος των *Εγγύτερων Γειτόνων* (*Nearest Neighbors*). Παρόλα αυτά, οι ολοκληρωμένες προσεγγίσεις που εν μέρει ήταν βασισμένες στην αναδειγματοληψία δεν έδωσαν ικανοποιητικά αποτελέσματα, οπότε χρησιμοποιήθηκαν εναλλακτικοί αλγόριθμοι, μη ευαίσθητοι στο φαινόμενο της Ανισοκατανομής, οι οποίοι δε συνδυάστηκαν με κάποια μορφή αναδειγματοληψίας. Η διαχείριση των δεδομένων, όμως, οδήγησε σε άμβλυνση της αρχικής Ανισοκατανομής, όπως περιγράφεται πιο αναλυτικά στην Ενότητα 3.4, που αφορά σε αυτήν ακριβώς τη διαχείριση.

2.4 Σύνολα Εκπαίδευσης και Ελέγχου

Στη Μηχανική Μάθηση δεν είναι δυνατό να χρησιμοποιηθούν τα ίδια δεδομένα για την παραγωγή ενός μοντέλου, δηλαδή την εκπαίδευση, και για την αξιολόγησή του. Αυτό συμβαίνει επειδή το μοντέλο κατά κάποιο τρόπο «θυμάται» τα δεδομένα από τα οποία έχει προκύψει, με αποτέλεσμα να είναι σε θέση να προβλέψει τη σωστή ετικέτα για κάθε είσοδο που προέρχεται από αυτά. Συνεπώς, μια αξιολόγηση που εκτελείται με τα ίδια δεδομένα δεν παρέχει αξιόπιστες ενδείξεις για το κατά πόσο το μοντέλο *γενικεύει* σε ικανοποιητικό βαθμό. Η έννοια της *γενίκευσης* εξηγείται στη συνέχεια.

Για να αξιολογηθεί σωστά η επίδοση ενός μοντέλου, αυτό τροφοδοτείται με νέες εισόδους, δηλαδή εισόδους οι οποίες δε χρησιμοποιήθηκαν κατά την εκπαίδευση του, για τις οποίες όμως είναι γνωστές οι ετικέτες. Αυτό είναι δυνατό αν το σύνολο των δεδομένων διαχωριστεί εξαρχής σε δύο επιμέρους σύνολα. Το πρώτο σύνολο, το οποίο χρησιμοποιείται για την εκπαίδευση, δηλαδή την παραγωγή του μοντέλου, ονομάζεται *σύνολο εκπαίδευσης* (*training set*). Το δεύτερο χρησιμοποιείται για την αξιολόγηση της επίδοσης του μοντέλου και ονομάζεται *σύνολο ελέγχου* (*test set*). Στην πορεία εξηγείται το πως προκύπτει η ανάγκη χρήσης ενός τρίτου συνόλου, του *συνόλου διακρίβωσης* (*validation set*).

2.5 Γενίκευση, Υπερπροσαρμογή και Υποπροσαρμογή

Στην Επιτηρούμενη Μάθηση, το ζητούμενο είναι η ανάπτυξη ενός μοντέλου βασισμένου σε γνωστά ζεύγη εισόδων και εξόδων, το οποίο θα μπορεί να προβλέψει με επιτυχία άγνωστες εξόδους, για τις οποίες οι είσοδοι έχουν παρόμοια χαρακτηριστικά με εκείνες που χρησιμοποιήθηκαν κατά την εκπαίδευση. Ένα μοντέλο που μπορεί να πετύχει τα παραπάνω χαρακτηρίζεται ως μοντέλο που *γενικεύει* καλά από το σύνολο εκπαίδευσης στο σύνολο ελέγχου. Δηλαδή, το ζητούμενο είναι η ανάπτυξη ενός

μοντέλου που γενικεύει κατά το δυνατόν πιο αποτελεσματικά.

Η διαδικασία της εκπαίδευσης συνίσταται στην ανάπτυξη του μοντέλου με τρόπο τέτοιο που να μπορεί να κάνει πολύ εύστοχες προβλέψεις στο σύνολο εκπαίδευσης. Αν το σύνολο εκπαίδευσης και το σύνολο ελέγχου - ή το σύνολο των άγνωστων δεδομένων που θα προκύψουν στο μέλλον - έχουν προκύψει υπό παρόμοιες συνθήκες, το μοντέλο θα είναι επίσης εύστοχο και σε αυτά τα σύνολα. Όμως, αν η διαδικασία εκπαίδευσης γίνει με τρόπο τέτοιο που το μοντέλο προβλέπει τέλεια το σύνολο εκπαίδευσης, τότε δε θα είναι πια τόσο εύστοχο στα άγνωστα δεδομένα, αφού αυτά δεν εχούν προκύψει υπό ακριβώς τις ίδιες συνθήκες.

Η ανάπτυξη ενός μοντέλου καθιστώντας το όλο και πιο σύνθετο, ανταποκρινόμενοι στα χαρακτηριστικά του συνόλου εκπαίδευσης, ώστε τελικά να μπορεί να προβλέπει το σύνολο αυτό τέλεια ονομάζεται *υπερπροσαρμογή* (*overfitting*). Ο λόγος για τον οποίο η υπερπροσαρμογή αποτελεί πρόβλημα είναι επειδή, όσο κι αν το σύνολο εκπαίδευσης και το σύνολο ελέγχου μοιάζουν, προφανώς τα χαρακτηριστικά τους δεν είναι ακριβώς ίδια, με αποτέλεσμα ένα μοντέλο τέλειο για το σύνολο εκπαίδευσης να μην μπορεί πλέον να ανταποκριθεί στο σύνολο ελέγχου, ούτε σε άγνωστες εισόδους. Θα έλεγε λοιπόν κανείς πως ένα απλό μοντέλο είναι προτιμητέο έναντι ενός πολύ σύνθετου· αυτό ισχύει, αρκεί το μοντέλο να μην είναι ούτε υπερβολικά απλό, δηλαδή να μην οδηγηθεί σε *υποπροσαρμογή* (*underfitting*). Σε αυτήν την περίπτωση, το μοντέλο δεν είναι ικανό να αποκωδικοποιήσει τα διάφορα χαρακτηριστικά των δεδομένων ικανοποιητικά, με συνέπεια να μην είναι εύστοχο ούτε στο σύνολο εκπαίδευσης.

Όπως ίσως είναι εμφανές, υπάρχει μια ισορροπία μεταξύ απλότητας και περιπλοκότητας ενός μοντέλου, ώστε αυτό να μην οδηγείται σε υπερπροσαρμογή ή υποπροσαρμογή, αλλά να έχει τη μέγιστη ικανότητα να γενικεύει. Το σημείο στο οποίο συμβαίνει αυτό είναι δυνατό να φανεί διαγραμματικά. Σχετικά διαγράμματα παρουσιάζονται και αναλύονται επαρκώς στο Κεφάλαιο 5 που αφορά στα αποτελέσματα της μελέτης.

2.6 Όγκος Δεδομένων και Περιπλοκότητα

Σε συνέχεια των παραπάνω, είναι σημαντικό να σημειωθεί ότι το αν ένα μοντέλο θεωρείται απλό ή σύνθετο σχετίζεται άμεσα με την ποικιλία των εισόδων που περιλαμβάνονται στο σύνολο εκπαίδευσης· όσο μεγαλύτερη είναι αυτή η ποικιλία, τόσο πιο σύνθετο μοντέλο μπορεί να χρησιμοποιηθεί χωρίς τον κίνδυνο υπερπροσαρμογής. Συνήθως, απλά και μόνο συλλέγοντας περισσότερα δεδομένα προκύπτει μεγαλύτερη ποικιλία, οπότε γενικά τα μεγαλύτερα σύνολα δεδομένων επιτρέπουν την ανάπτυξη πιο σύνθετων μοντέλων. Φυσικά, όπως γίνεται ίσως κατανοητό, η αντιγραφή σημείων

από τα υπάρχοντα δεδομένα ή η συλλογή δεδομένων με πολύ κοντινά χαρακτηριστικά δεν προσφέρουν κάτι, αφού δεν προσθέτουν ποικιλία στα δεδομένα.

Σε αρκετές πρακτικές εφαρμογές της Επιτηρούμενης Μάθησης υπάρχει η δυνατότητα συλλογής περισσότερων δεδομένων. Αυτός είναι και ένας πολύ αποτελεσματικός τρόπος βελτίωσης της επίδοσης των παραγόμενων μοντέλων, χωρίς π.χ. τη μεταβολή καμίας παραμέτρου των αλγορίθμων, ή γενικότερα χωρίς καμία μεταβολή της ακολουθούμενης διαδικασίας εκπαίδευσης. Η εφαρμογή που εξετάζεται στην παρούσα μελέτη εμπίπτει σε αυτήν ακριβώς την κατηγορία. Δυστυχώς, όμως, επειδή τα δεδομένα που αξιοποιούνται διατηρούνται στη βάση δεδομένων της Δραστηριότητας Ηλεκτρόλυσης για κάτι λιγότερο από δύο μήνες, ο όγκος δεδομένων που είναι διαθέσιμος για την εκπόνηση της μελέτης είναι περιορισμένος. Συγκεκριμένα, συνίσταται στα δεδομένα που καταγράφηκαν από την έναρξη της μελέτης και ύστερα, μαζί με τα δεδομένα των αμέσως δύο προηγούμενων μηνών. Για αυτό το λόγο ίσως δεν είναι επαρκής για την επίτευξη ενός ιδιαίτερα ικανοποιητικού αποτελέσματος. Σε περίπτωση που η μελέτη συνεχιστεί και μετά την ολοκλήρωση της μεταπτυχιακής εργασίας, είναι πολύ πιθανό η ακρίβεια των προβλέψεων να βελτιωθεί δραστικά.

Κεφάλαιο 3

Διαχείριση των Δεδομένων

Πριν αναπτυχθεί η μεθοδολογία για την επίλυση του προβλήματος, πραγματοποιείται μια συνοπτική αναδιατύπωσή του. Κατά την παραγωγική διαδικασία της Δραστηριότητας Ηλεκτρόλυσης του εργοστασίου του Αλουμινίου της Ελλάδας, αλουμίνιο παράγεται από αλουμίνα σε ηλεκτρολυτικά κελιά που καλούνται *λεκάνες*. Πληθώρα μεγεθών που χαρακτηρίζουν τη λεκάνη και τη λειτουργία της, όπως π.χ. η θερμοκρασία της λεκάνης, η τιμή ενός μεγέθους που ονομάζεται *θόρυβος* και αποτελεί το μέτρο της *αστάθειας* της λεκάνης ή το αποτέλεσμα ενός περιοδικού τεστ, καταγράφονται σε πραγματικό χρόνο απο συστήματα αισθητήρων και μικροϋπολογιστών ή εργαζομένους της δραστηριότητας, με διαφορετικές συχνότητες δειγματοληψίας.

Ένα από αυτά τα μεγέθη, το οποίο καταγράφεται από εργαζομένους ανά βάρδια, δηλαδή 3 φορές την ημέρα (συγκεκριμένα στις 6 π.μ, 2 μ.μ. και 10 μ.μ) είναι το πλήθος των μανιταριών που εντοπίστηκαν σε κάθε λεκάνη στη βάρδια, μετά από σχετικό έλεγχο. Τα μανιτάρια αποτελούν ένα από τα ανοδικά προβλήματα των λεκανών και απομακρύνονται από τους εργαζομένους ως ανεπιθύμητα. Για το λόγο αυτό, ένα σύστημα που ενημερώνει για την πιθανή ύπαρξη μανιταριών - με βάση τις τιμές θορύβου και τεστ - ενημερώνει τους εργαζομένους να προχωρήσουν σε έλεγχο συγκεκριμένων λεκανών και αυτοί, αφού απομακρύνουν τυχόν μανιτάρια που εντοπίζουν, καταγράφουν το πλήθος τους στη βάση δεδομένων της Δραστηριότητας.

Συνοψίζοντας, για κάθε βάρδια είναι γνωστές οι τιμές διαφόρων μεγεθών που χαρακτηρίζουν κάθε λεκάνη, μαζί με το πλήθος των μανιταριών, ή απλούστερα την ύπαρξη μανιταριών σε αυτήν. Αν η πρόβλεψη της ύπαρξης μανιταριών διατυπωθεί ως πρόβλημα Μηχανικής Μάθησης, τότε εμπίπτει στο πεδίο της Επιτηρούμενης Μάθησης, αφού τα δεδομένα φέρουν ετικέτες. Εφόσον μάλιστα ζητείται πρόβλεψη μεταξύ πεπερασμένων ενδεχομένων τότε είναι πρόβλημα Ταξινόμησης, συγκεκριμένα δε Δυαδικής Ταξινόμησης, αφού τα πιθανά ενδεχόμενα είναι δύο, *ύπαρξη ή μη ύπαρξη* μανιταριού.

3.1 Αρχική Μορφή των Δεδομένων

Τα δεδομένα που ανακτώνται από τη βάση δεδομένων διακρίνονται σε αυτά από τα οποία παράγονται τα *χαρακτηριστικά* και σε αυτά από τα οποία παράγονται οι *ετικέτες*.

Τα δεδομένα από τα οποία παράγονται τα χαρακτηριστικά έχουν μορφή *χρονοσειράς*. *Χρονοσειρά* ονομάζεται μία αλληλουχία τιμών οι οποίες έχουν ληφθεί σε διαδοχικές χρονικές στιγμές. Εδώ, τα δεδομένα περιλαμβάνουν 4 χρονοσειρές που αντιστοιχούν σε 4 διαφορετικά μεγέθη, τα οποία είναι η *αντίσταση*, R_{m}, ο *θόρυβος*, W_{m}, η *τιμή του τεστ*, D_{r} και η *ζητούμενη αντίσταση*, R_{o} για κάθε μία από τις λεκάνες. Το περιεχόμενο των τριών πρώτων μεγεθών έχει συζητηθεί ήδη στην Εν. 1.2, ενώ το τελευταίο μέγεθος, όπως υποδεικνύεται και από το όνομα του, είναι η επιθυμητή αντίσταση της λεκάνης, η οποία κάθε στιγμή μπορεί να μεταβάλλεται με βάση διάφορα κριτήρια, που όμως δεν κρίνεται σκόπιμο να αναλυθούν.

Γενικά, τα μεγέθη αυτά μετρώνται και καταγράφονται αυτόματα και σε πραγματικό χρόνο από το σύστημα αισθητήρων-μικροϋπολογιστή με το οποίο είναι εξοπλισμένη κάθε λεκάνη. Εξαίρεση αποτελεί η *ζητούμενη αντίσταση*, η οποία αποτελεί μεταβλητή ελέγχου της διεργασίας, οπότε δε μετράται αλλά μόνο καταγράφεται. Η συχνότητα καταγραφής όλων των μεγεθών είναι 1 τιμή/min, δηλαδή η περίοδος καταγραφής είναι 1 min. Για αυτό τα συγκεκριμένα δεδομένα αναφέρονται στη συνέχεια ως *δεδομένα λεπτού*. Η αρχική μορφή των δεδομένων λεπτού σε μορφή πίνακα φαίνεται στον Πίν. 3.1.

Πίνακας 3.1: Αρχική μορφή των δεδομένων λεπτού, συγκεκριμένα των δεδομένων που αφορούν στο θόρυβο, W_{m}, όπως αυτά διαβάζονται (με μικρή επεξεργασία) από τη βάση δεδομένων.

Time	A101	A102	A103	A104	A105	⋯	C461	C462	C463	C464	C465
2020-03-02 06:00:00	0.56	0.59	0.23	0.14	0.14	⋯	0.16	1.33	1.45	0.12	2.73
2020-03-02 06:01:00	0.52	0.48	0.25	0.14	0.16	⋯	0.14	1.36	1.45	0.15	2.77
2020-03-02 06:02:00	0.38	0.38	0.22	0.14	0.16	⋯	0.14	1.42	1.42	0.17	3.11
2020-03-02 06:03:00	0.33	0.36	0.19	0.14	0.16	⋯	0.15	1.47	1.44	0.18	3.18
2020-03-02 06:04:00	0.43	0.33	0.19	0.14	0.16	⋯	0.15	1.44	1.50	0.15	3.18
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	
2020-05-08 13:56:00	0.22	0.15	0.15	0.29	1.23	⋯	0.16	0.20	0.27	0.13	2.51
2020-05-08 13:57:00	0.22	0.15	0.17	0.29	1.31	⋯	0.16	0.20	0.25	0.14	2.50
2020-05-08 13:58:00	0.25	0.15	0.17	0.29	1.41	⋯	0.18	0.21	0.22	0.14	2.57
2020-05-08 13:59:00	0.18	0.15	0.17	0.29	1.28	⋯	0.17	0.23	0.20	0.13	2.52
2020-05-08 14:00:00	0.18	0.15	0.17	0.29	1.21	⋯	0.23	0.23	0.20	0.13	2.33

Τα δεδομένα από τα οποία παράγονται οι ετικέτες έχουν τη μορφή μεμονωμένων καταγραφών, είναι όμως χρήσιμο και απλό να πάρουν και αυτά μορφή χρονοσειράς. Συγκεκριμένα, αποτελούν καταγραφές λεκανών όπου έχουν εντοπιστεί μανιτάρια, μαζί με το χρόνο στον οποίο αυτά εντοπίστηκαν. Τα δεδομένα αυτά διαφέρουν από τα δεδομένα λεπτού ως προς τον τρόπο συλλογής και καταγραφής, αφού τα μανιτάρια εντοπίζονται από εργαζόμενους μέσω δειγματοληπτικού ελέγχου και καταγράφονται χειροκίνητα στη βάση δεδομένων με συχνότητα καταγραφής 1 τιμή/πόστο, δηλαδή με περιόδο καταγραφής 8 h. Η αρχική τους μορφή φαίνεται στον Πίν. 3.2. Για να πάρουν τη μορφή κοινής χρονοσειράς, οι συνδυασμοί λεκάνης-πόστου όπου εντοπίστηκαν μανιτάρια σημειώνονται με 1 (ή True), ενώ αυτοί όπου δεν εντοπίστηκαν σημειώνονται με 0 (ή False). Το αποτέλεσμα αυτής της επεξεργασίας φαίνεται στον Πίν. 3.3. Σύμφωνα με τους ορισμούς που έχουν δωθεί παραπάνω, τα 0 και 1 αποτελούν ετικέτες, για αυτό τα παραπάνω δεδομένα αναφέρονται στη συνέχεια ως *δεδομένα ετικετών*.

Ο χαρακτηρισμός των ελέγχων για εντοπισμό μανιταριών ως δειγματοληπτικών ίσως είναι υπερβολικός. Προφανώς οι εργαζόμενοι έχουν μια αίσθηση για το σε ποια λεκάνη είναι πιο πιθανό να εντοπιστούν μανιτάρια, βασισμένη στην εμπειρία τους και στις πληροφορίες που έχουν για την κατάσταση και την πορεία των λεκανών, δηλαδή δεν επιλέγουν τις λεκάνες που ελέγχουν τυχαία. Επειδή όμως δεν υπάρχει χρόνος να ελεγχθούν πολλές λεκάνες και η αίσθηση τους συχνά είναι άστοχη, πολλά μανιτάρια δεν εντοπίζονται εγκαίρως, με συνέπεια να θεωρείται λανθασμένα ότι οι αντίστοιχες λεκάνες ήταν εντάξει στα πόστα που προηγούνται αυτού στο οποίο βρέθηκε το μανιτάρι, με αποτέλεσμα οι ετικέτες των λεκανών να είναι False αντί για True.

Σε περίπτωση που η παρούσα μελέτη συνεχιστεί, η ακρίβεια των ετικετών είναι δυνατό - και επιβάλλεται - να διασφαλιστεί, επιλέγοντας ένα μικρό υποσύνολο των λεκανών, το οποίο θα ελέγχεται για μανιτάρια σε κάθε πόστο. Εναλλακτικά, θα μπορούσε να καταγράφεται το ποιες ήταν οι λεκάνες που ελέγχθηκαν σε κάθε πόστο, ώστε οι ακριβείς ετικέτες να μην χάνονται μέσα στις υπόλοιπες.

Πίνακας 3.2: Αρχική μορφή των δεδομένων ετικετών, όπως αυτά διαβάζονται (με μικρή επεξεργασία) από τη βάση δεδομένων.

	Time	Cell	Ama
0	2020-01-01 14:00:00	A248	True
1	2020-01-01 14:00:00	B423	True
2	2020-01-01 14:00:00	B465	True
3	2020-01-02 14:00:00	A130	True
4	2020-01-02 14:00:00	A160	True

Από τα παραπάνω στοιχεία για τα δεδομένα γίνονται εμφανείς ορισμένες δυσκολίες που σχετίζονται με την αξιοποίηση τους.

- Τα δεδομένα λεπτού και ετικετών διαφέρουν σε συχνότητα καταγραφής.
- Όλα τα δεδομένα έχουν μορφή χρονοσειράς, οπότε κατά το σχηματισμό του συνόλου δεδομένων πρέπει με κάποιο τρόπο να διατηρηθεί η πληροφορία για τη χρονική εξέλιξη των μεγεθών.
- Τα δεδομένα ετικετών είναι ελλιπή, ή καλύτερα ανακριβή.

Οι στρατηγικές που υιοθετούνται για την αντιμετώπιση των δυσκολιών αναλύονται όσο γίνεται πιο διεξοδικά, αφού είναι ένα από τα στοιχεία που διαφοροποιούν την παρούσα μελέτη από άλλες παρόμοιες και την καθιστούν χρήσιμη στον αναγνώστη. Τονίζεται δε πως οι ιδέες που περιγράφονται παρακάτω δεν έχουν αντληθεί άμεσα από τη βιβλιογραφία, μόνο ορισμένες έχουν εκ των υστέρων επικυρωθεί από αυτήν.

Σημειώνεται πως πέρα από τα παραπάνω δεδομένα, δηλαδή τα δεδομένα λεπτού και τα δεδομένα ετικετών, στη βάση δεδομένων υπάρχουν και *δεδομένα ημέρας*, καθώς και *δεδομένα πόστου*. Τα δεδομένα αυτά, όπως υποδεικνύεται από το όνομά τους, χαρακτηρίζονται από συχνότητα καταγραφής 1 τιμή/ημέρα και 1 τιμή/πόστο, αντίστοιχα. Και αυτά ανακτήθηκαν από τη βάση δεδομένων όπως και όσα έχουν ήδη αναφερθεί, υποβλήθηκαν στην ίδια προκατεργασία και ανάλυση με τα υπόλοιπα και μάλιστα αξιοποιήθηκαν σε δοκιμαστικές εφαρμογές αλγορίθμων Μηχανικής Μάθησης. Παρόλα αυτά, λόγω του ότι είτε αποτελούσαν μέσους όρους των δεδομένων λεπτού, είτε ήταν καταγεγραμμένα πολύ αραιά και με μη ακριβή τρόπο, δηλαδή από εργαζομένους της Δραστηριότητας, η ανάλυση έδειξε ότι δεν παρέχουν ιδιαίτερα χρήσιμες πληροφορίες. Τα δε προκαταρκτικά αποτελέσματα υστερούσαν σημαντικά έναντι αυτών που

Πίνακας 3.3: Μορφή των δεδομένων ετικετών μετά την αρχική τους επεξεργασία.

Cell **Time**	A101	A102	A103	A104	A105	⋯	C461	C462	C463	C464	C465
2020-01-01 14:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-01-01 22:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-01-02 06:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-01-02 14:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-01-02 22:00:00	False	False	False	False	False	⋯	False	False	False	False	False
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	
2020-06-24 22:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-06-25 06:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-06-25 14:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-06-25 22:00:00	False	False	False	False	False	⋯	False	False	False	False	False
2020-06-26 06:00:00	False	False	False	False	False	⋯	False	False	False	False	False

προκύπτουν με τα δεδομένα που αναλύονται παρακάτω. Για το λόγο αυτό δεν αξιοποιούνται περαιτέρω στην παρούσα μελέτη, ενώ ο συνδυασμός τους με τα δεδομένα που τελικά αξιοποιούνται για ένα καλύτερο αποτέλεσμα θα εξεταστεί μελλοντικά σε περίπτωση συνέχισης της μελέτης.

3.2 Αναγωγή των Δεδομένων στην Ίδια Συχνότητα

Όπως αναφέρθηκε παραπάνω, τα δεδομένα λεπτού και ετικετών διαφέρουν ως προς τη συχνότητα καταγραφής. Συνεπώς, για να συνδυαστούν και να αξιοποιηθούν θα πρέπει πρώτα να αναχθούν στην ίδια συχνότητα. Η συχνότητα αυτή θα μπορούσε να είναι μία από τις δύο αρχικές τους συχνότητες, δηλαδή 1 τιμή/min ή 1 τιμή/πόστο. Στην πραγματικότητα, θα μπορούσε να είναι και οποιαδήποτε άλλη συχνότητα, είτε μεταξύ των δύο αυτών συχνοτήτων, είτε και εκτός του εύρους που αυτές ορίζουν.

Για να αναχθούν τα δεδομένα λεπτού στο πόστο πρέπει οι τιμές κάθε μεγέθους κατά τη διάρκεια του πόστου να μετατραπούν σε μία μόνο τιμή. Το αρχικό πλήθος των τιμών αυτών προκύπτει ως εξής: $8\ \mathrm{h} \times 60\ \frac{\mathrm{min}}{\mathrm{h}} \times 1\ \frac{\text{τιμή}}{\mathrm{min}} = 480$ τιμές. Αυτό μπορεί να επιτευχθεί αν τη θέση των αρχικών τιμών πάρει κάποιο στατιστικό μέγεθος που να τις περιγράφει, π.χ. ο μέσος όρος τους. Φυσικά, ανακύπτει το ερώτημα του ποιο είναι το καταλληλότερο στατιστικό μέγεθος για να αποτυπώσει το μέρος εκείνο της αρχικής πληροφορίας που είναι πιο χρήσιμο στην παρούσα μελέτη. Αυτό είναι μάλλον δύσκολο να απαντηθεί προκαταβολικά, οπότε κατα την εφαρμογή αυτής της προσέγγισης παράγονται διάφορα στατιστικά μεγέθη και πραγματοποιείται μια σχετική διερεύνηση σε επόμενο στάδιο της μελέτης. Η προφανής αδυναμία της προσέγγισης είναι ότι συνεπάγεται την απώλεια ενός σημαντικού μέρους της αρχικής πληροφορίας.

Η εναλλακτική επιλογή, δηλαδή η αναγωγή των δεδομένων ετικετών στο λεπτό, μπορεί να επιτευχθεί αν η πληροφορία για την παρουσία μανιταριού αποδοθεί σε κάθε μεμονωμένο λεπτό του αντίστοιχου πόστου. Η προσέγγιση αυτή παρουσιάζει μάλλον αποκλειστικά αδυναμίες. Από τη μία, αυξάνει κατακόρυφα τον όγκο των δεδομένων, σε βαθμό που δεν είναι διαχειρίσιμος με συμβατικές μεθόδους Μηχανικής Μάθησης, ενώ από την άλλη ενισχύει την αβεβαιότητα για την αξιοπιστία των ετικετών, ή καλύτερα δημιουργεί ετικέτες που είναι σίγουρα αναξιόπιστες. Η δεύτερη αδυναμία μπορεί να εξηγηθεί μέσα από ένα παράδειγμα. Αν σε μία λεκάνη εντοπιστεί μανιτάρι στις 07:00, για τη λεκάνη θα καταγραφεί παρουσία μανιταριού στο πρωινό πόστο, το οποίο διαρκεί μεταξύ 06:00 και 14:00. Αν η πληροφορία αυτή αναχθεί σε όλα τα λεπτά του πόστου, θα δημιουργηθούν 60 παρατηρήσεις για τις οποίες οι τιμές των χαρακτηριστικών αντιστοιχούν σε παρουσία μανιταριού, ενώ ταυτόχρονα θα δημιουργηθούν 420 παρατηρήσεις για τις οποίες δεν αντιστοιχούν σε παρουσία μανιταριού,

αφού τα μανιτάρια έχουν ήδη αφαιρεθεί. Απο τα παραπάνω προκύπτει πως η τελευταία προσέγγιση δεν προσφέρει απολύτως τίποτα αλλά δημιουργεί πρόσθετες δυσκολίες. Τέλος, μια προσέγγιση μεταξύ των δύο θα υποφέρει από τις αδυναμίες της δεύτερης, οπότε με σχετική σιγουριά επιλέγεται η αναγωγή των δεδομένων λεπτού στο πόστο.

Τα στατιστικά μεγέθη τα οποία επιλέγονται για την αναγωγή είναι 8: ο *μέσος όρος*, η *τυπική απόκλιση*, το *ελάχιστο*, το *μέγιστο*, η *διάμεσος*, το *κατώτερο τεταρτημόριο* (Q1), το *ανώτερο τεταρτημόριο* (Q3) και το *διατεταρτημοριακό εύρος* (IQR) των τιμών κάθε μεγέθους στο πόστο. Τα 4 πρώτα είναι τα πλέον κοινά στατιστικά μεγέθη, η ικανότητα τους όμως να περιγράφουν έναν πληθυσμό υποβαθμίζεται σημαντικά από την παρουσία ακραίων τιμών στο δείγμα. Τέτοιες τιμές είναι δεδομένο ότι θα υπάρχουν στην περίπτωση μεγεθών που καταγράφονται αυτόματα από αισθητήρες και μάλιστα σε βιομηχανικές συνθήκες. Για παράδειγμα, πολλοί βιομηχανικοί αισθητήρες, σε περίπτωση δυσλειτουργίας καταγράφουν αυθαίρετα τη μέγιστη τιμή ή την ελάχιστη τιμή που μπορούν να μετρήσουν. Οι ακραίες τιμές είναι δυνατό αλλά συχνά δύσκολο να αφαιρεθούν από τα δεδομένα, γι΄ αυτό εκτός από τα κοινά στατιστικά μεγέθη, στην παρούσα μελέτη αξιοποιούνται και στατιστικά μεγέθη τα οποία επηρέαζονται λιγότερο από την παρουσία ακραίων τιμών στο δείγμα. Τα πρώτα, λοιπόν, χαρακτηριστικά που αξιοποιούνται σε αυτήν τη μελέτη είναι τα παραπάνω στατιστικά μεγέθη για κάθε καταγεγγραμένο μέγεθος, π.χ. ο μέσος όρος των τιμών της αντίστασης που καταγράφηκαν σε ένα συγκεκριμένο πόστο για μία λεκάνη.

Σε αυτό το σημείο αξίζει να διασαφηνιστεί το περιεχόμενο των στατιστικών μεγεθών που ενδεχομένως δεν είναι γνωστά σε όλους τους αναγνώστες, IQR, Q1 και Q3. Το IQR είναι η διαφορά μεταξύ των τιμών του *τρίτου* και του *πρώτου τεταρτημορίου*

Πίνακας 3.4: Τελική μορφή των δεδομένων λεπτού που αφορούν στο θόρυβο, αφού αυτά αναχθούν στο πόστο χρησιμοποιώντας το μέσο όρο τους (Wm_avg).

Time	A101	A102	A103	A104	A105	⋯	C461	C462	C463	C464	C465
2020-03-02 14:00:00	0.368	0.513	0.723	0.144	0.158	⋯	0.140	0.970	0.517	0.126	2.092
2020-03-02 22:00:00	0.333	0.584	0.585	0.148	0.335	⋯	0.179	0.146	0.124	0.143	1.523
2020-03-03 06:00:00	0.325	0.592	0.520	0.144	0.207	⋯	0.137	0.194	0.174	0.168	1.456
2020-03-03 14:00:00	1.194	0.833	0.680	0.182	0.203	⋯	0.203	0.255	0.321	0.184	0.965
2020-03-03 22:00:00	1.056	1.075	0.558	0.298	0.439	⋯	0.113	0.254	0.575	0.165	0.186
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	
2020-05-07 06:00:00	0.182	0.143	0.213	0.223	0.213	⋯	1.016	0.262	1.289	1.275	1.795
2020-05-07 14:00:00	0.186	0.147	0.176	0.243	0.893	⋯	0.552	0.281	2.038	0.666	1.272
2020-05-07 22:00:00	0.158	0.140	0.158	0.228	0.251	⋯	0.415	0.352	1.624	0.269	1.270
2020-05-08 06:00:00	0.157	0.152	0.160	0.308	0.245	⋯	1.426	0.254	1.740	0.474	1.626
2020-05-08 14:00:00	0.175	0.228	0.199	0.396	0.317	⋯	0.958	0.227	0.268	0.696	1.900

(*first* and *third quartile*), ή αλλιώς του *ανώτερου* και του *κατώτερου τεταρτημορίου* (*upper* and *lower quartile*), συμβολικά IQR = Q3 − Q1. Η τιμή Q3 είναι η πρώτη τιμή που είναι μεγαλύτερη αυτών που ανήκουν στα τρία πρώτα τεταρτημόρια, δηλαδή στο χαμηλότερο 75% από το σύνολο των τιμών. Αντίστοιχα, η τιμή Q1 είναι η πρώτη τιμή που είναι μεγαλύτερη αυτών που ανήκουν στο πρώτο τεταρτημόριο, δηλαδή στο χαμηλότερο 25% των τιμών. Για προφανείς λόγους αναφέρονται εναλλακτικά ως τιμές *75ου* και *25ου ποσοστημορίου* ή *εκατοστημορίου* (*75th* and *25th quantile/percentile*).

Ανακεφαλαιώνοντας, από τις τιμές που καταγράφονται ανά λεπτό για κάθε ένα από τα αρχικά μεγέθη, υπολογίζονται 8 στατιστικά μεγέθη για κάθε πόστο. Συγκεκριμένα, από τα 3 αρχικά μεγέθη, τα οποία είναι ο θόρυβος, W_{m}, η τιμή του τεστ, D_{r} και η απόλυτη τιμή της διαφοράς μεταξύ μετρούμενης και ζητούμενης αντίστασης της λεκάνης, $R_{\mathrm{d}} = |R_{\mathrm{m}} - R_{\mathrm{o}}|$, προκύπτουν 24 χαρακτηριστικά[1]. Αυτά, ανάλογα με το σε ποιο αρχικό μέγεθος και σε ποιο στατιστικό μέγεθος βασίζονται, συμβολίζονται π.χ. ως Wm_avg, που αντιστοιχεί στο μέσο όρο του θορύβου στο πόστο. Η τελική μορφή των δεδομένων που αφορούν σε αυτό το μέγεθος φαίνεται στον Πίν. 3.4. Αντίστοιχα σύμβολα χρησιμοποιούνται και για τα υπόλοιπα παράγωγα μεγέθη - χαρακτηριστικά.

Συμπληρώνοντας τα όσα ειπώθηκαν στο τέλος της προηγούμενης ενότητας για τα δεδομένα πόστου και ημέρας, τα οποία δεν παρουσιάζονται, αναφέρεται ότι τα δεδομένα πόστου προφανώς δεν απαιτούν παρόμοια διαχείριση με αυτά του λεπτού. Αυτό συμβαίνει επειδή εξαρχής βρίσκονται στην ίδια συχνότητα με τα δεδομένα ετικετών. Όσον αφορά τα δεδομένα ημέρας, αυτά δεν συνοψίζονται με χρήση κάποιου στατιστικού μεγέθους, απλούστατα επειδή η συχνότητα καταγραφής τους είναι υψηλότερη από τη ζητούμενη. Αντίθετα, ζητείται να παραχθούν 3 τιμές, μία για κάθε πόστο της εκάστοτε ημέρας, από τη μία διαθέσιμη τιμή. Στο πλαίσιο της διερεύνησης που έγινε στην παρούσα μελέτη, αυτό πραγματοποιήθηκε με δύο τρόπους: με γραμμική παρεμβολη, αποδίδοντας την τιμή της ημέρας σε ένα από τα τρία πόστα και συμπληρώνοντας τα υπόλοιπα, είτε απλά αντιστοιχίζοντας την τιμή σε όλα τα πόστα της ημέρας.

3.3 Αξιοποίηση της Χρονικής Εξέλιξης των Μεγεθών

Επειδή τα δεδομένα που αφορούν στα χαρακτηριστικά έχουν τη μορφή χρονοσειράς, είναι δηλαδή γνωστή η χρονική τους εξέλιξη, η πληροφορία αυτή μπορεί - και απαιτείται - να αξιοποιηθεί για την εκτίμηση που θα πραγματοποιηθεί. Για παράδειγμα, αν στον αλγόριθμο Μηχανικής Μάθησης τροφοδοτηθούν ως χαρακτηριστικά τόσο

[1] Υπενθυμίζεται ότι ο θόρυβος, που αποτελεί το μέτρο της αστάθειας (δηλαδή της ανομοιόμορφης κατανομής του ρεύματος στις ανόδους) δίνεται από τη σχέση $W_{\mathrm{m}} = R_{\max} - R_{\min}$, όπου $R_{\max}$ και $R_{\min}$ η μέγιστη και η ελάχιστη στιγμιαία τιμή της αντίστασης της λεκάνης ανά λεπτό, αντίστοιχα.

η τρέχουσα κατάσταση των υπό εξέταση μεγεθών, όσο και ιστορικά δεδομένα που τα αφορούν, μπορεί το αποτέλεσμα της πρόβλεψης που θα προκύψει να είναι σαφώς βελτιωμένο.

Αντίστοιχα με την αναγωγή των δεδομένων στην ίδια συχνότητα, η αξιοποίηση της χρονικής εξέλιξης των μεγεθών μπορεί να πραγματοποιηθεί με διαφορετικούς τρόπους. Ο πρώτος είναι η αξιοποίηση μεθόδων *Ανάλυσης Χρονοσειρών* (*Time Series Analysis*). Όμως, στην παρούσα μελέτη έχει γίνει η επιλογή να αξιοποιηθούν μόνο οι συμβατικοί αλγόριθμοι Μηχανικής Μάθησης, πλην ίσως των Νευρωνικών Δικτύων, οπότε η προσέγγιση της Ανάλυσης Χρονοσειρών δεν προκρίνεται. Στη θέση της αναζητείται μια εναλλακτική που να είναι συμβατή με τους συγκεκριμένους αλγορίθμους.

Μια εναλλακτική επιλογή είναι η εισαγωγή προηγούμενων τιμών του κάθε χαρακτηρισμού υπό μορφή νέων χαρακτηριστικών. Για παράδειγμα, πέραν των 24 χαρακτηριστικών, θα μπορούσαν να εισαχθούν 24 νέα, τα οποία θα περιείχαν τις τιμές των αρχικών, απλά μετατοπισμένες κατά ένα πόστο στο παρελθόν. Ο αλγόριθμος, γνωρίζοντας την τρέχουσα και την αμέσως προηγούμενη τιμή κάθε χαρακτηριστικού ενδεχομένως να μπορεί να αντιστοιχίσει την μεταβολή των χαρακτηριστικών με την εμφάνιση μανιταριών, δηλαδή τη μεταβολή των ετικετών. Φυσικά, η παραπάνω προσέγγιση μπορεί να επεκταθεί εισάγοντας όλο και περισσότερα πόστα από το παρελθόν ως νέα χαρακτηριστικά, με τα χαρακτηριστικά που προκύπτουν να είναι $(\mathrm{n}+1)\times 24$, οπου n το πλήθος των πόστων τα οποία εξετάζονται.

Από τα παραπάνω γίνεται εμφανές πως η χρονική εξέλιξη των μεγεθών μπορεί να τροφοδοτηθεί στον αλγόριθμο με πολλούς διαφορετικούς τρόπους. Φυσικά, αυτός που τελικά θα επιλεγεί πρέπει να έχει νόημα και να μην αυξάνει δραματικά τον όγκο των δεδομένων, σε βαθμό που αυτός να μην είναι πλέον διαχειρίσιμος. Στην παρούσα μελέτη χρησιμοποιείται δοκιμαστικά ένας τρόπος μεταξύ των πολλών, βασισμένος σε στατιστικά μεγέθη, όπως και τα παράγωγα χαρακτηριστικά που δημιουργήθηκαν. Για κάθε ένα από τα 24 χαρακτηριστικά λαμβάνεται η διάμεσος, η τυπική απόκλιση, το Q1 και το Q3 των τιμών του κατά τα προηγούμενα 5 πόστα. Έτσι επιδιώκεται να αποτυπωθεί η τάση κάθε μεγέθους, υπό την έννοια του αν αυτό λαμβάνει χαμηλές ή υψηλές τιμές, καθώς και αν παρουσιάζει σημαντικές διακυμάνσεις ή είναι σταθερό, για μια μικρή χρονική περίοδο πριν την εκάστοτε χρονική στιγμή.

Η προσέγγιση που ακολουθείται οδηγεί στην παραγωγή $4\times 24 = 96$ νέων χαρακτηριστικών τα οποία, σε συνδυασμό με τα 24 αρχικά, ανεβάζουν το συνολικό πλήθος των χαρακτηριστικών στα 120. Το πλήθος αυτό θα μπορούσε να θεωρηθεί μεγάλο, ειδικά σε συνδυασμό με το μεγάλο πλήθος των διαθέσιμων παρατηρήσεων. Το κατά πόσο όμως είναι διαχειρίσιμο εξαρτάται από τον αλγόριθμο που θα χρησιμοποιηθεί.

Τα χαρακτηριστικά αυτά συμβολίζονται π.χ. ως Wm_avg_std_5, όνομα το οποίο υποδεικνύει πως το χαρακτηριστικό αντιστοιχεί στην τυπική απόκλιση του μέσου όρου του θορύβου για τα 5 αμέσως προηγούμενα πόστα.

3.4 Αντιμετώπιση της Αβεβαιότητας των Ετικετών

Ανακεφαλαιώνοντας, κάθε παρατήρηση περιγράφει την κατάσταση μιας λεκάνης σε ένα συγκεκριμένο πόστο και αποτελείται από ορισμένα χαρακτηριστικά και μία ετικέτα. Τα χαρακτηριστικά της παρατήρησης είναι οι τιμές ορισμένων στατιστικών μεγεθών που υπολογίζονται για τρία διαφορετικά φυσικά μεγέθη, τα οποία καταγράφονται σε πραγματικό χρόνο για τη λεκάνη. Η ετικέτα είναι ο αριθμός 1 αν στο συγκεκριμένο πόστο η λεκάνη ελέγχθηκε και σε αυτήν εντοπίστηκε μανιτάρι, ενώ είναι 0 σε κάθε άλλη περίπτωση. Πιο συγκεκριμένα, η ετικέτα είναι 0 αν η λεκάνη δεν ελέγχθηκε, ή ελέγχθηκε και σε αυτήν δεν εντοπίστηκε μανιτάρι.

Ένα βασικό πρόβλημα που αναφέρθηκε και παραπάνω είναι ότι οι τιμές των χαρακτηριστικών στο πόστο που εντοπίζεται το μανιτάρι δεν αντιστοιχούν αποκλειστικά σε παρουσία μανιταριού, αφού κάποια στιγμή μέσα στο πόστο, το μανιτάρι εντοπίστηκε και αφαιρέθηκε. Άρα, οι αντίστοιχες παρατηρήσεις μπορεί να παραπλανήσουν τον αλγόριθμο. Γι΄ αυτό το λόγο, μια πρώτη τροποποίηση των δεδομένων είναι οι παρατηρήσεις αυτές να αφαιρεθούν και η ετικέτα 1 να αποδοθεί στο αμέσως προηγούμενο πόστο. Αυτό πραγματοποιείται με την παραδοχή ότι το μανιτάρι σίγουρα υπήρχε στο προηγούμενο πόστο και τα χαρακτηριστικά του αντιστοιχούν σε παρουσία μανιταριού.

Συχνά, η ετικέτα 0 αντιστοιχεί σε λεκάνες που δεν ελέγχθηκαν. Αν είχαν ελεγχθεί, είναι πιθανό κάποιες από αυτές να είχαν την ετικέτα 1. Φυσικά, σε επόμενο χρόνο οι λεκάνες αυτές ελέγχονται και τελικά τους αποδίδεται η ετικέτα 1. Είναι λοιπόν εμφανές ότι η ετικέτα 1 υποδεικνύει το πόστο στο οποίο εντοπίστηκε το μανιτάρι και όχι αυτό στο οποίο εμφανιστήκε, ούτε αυτά που μεσολάβησαν μέχρι να εντοπιστεί. Κατά συνέπεια, κάποιες παρατηρήσεις είναι παραπλανητικές, αφού χαρακτηριστικά που συνδέονται με παρουσία μανιταριού αντιστοιχίζονται στην ετικέτα 0.

Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με δύο τρόπους. Ο πρώτος, που είναι και ο πιο συντηρητικός και ασφαλής τρόπος, είναι οι αβέβαιες παρατηρήσεις να αφαιρεθούν από τα δεδομένα. Ωστόσο, δεν είναι γνωστό το πόσες είναι οι αβέβαιες παρατηρήσεις. Σύμφωνα με τους αρμόδιους μηχανικούς, ένα μανιτάρι μπορεί να έχει εμφανιστεί ακόμη και 10 ημέρες πριν τον εντοπισμό του. Έτσι, δημιουργείται ένα νέο σύνολο δεδομένων στο οποίο έχουν αφαιρεθεί οι παρατηρήσεις έως και 15 ημέρες πριν από κάθε παρατήρηση με ετικέτα 1. Στο εξής, το σύνολο αυτό αναφέρεται για ευκολία

ως *καθαρά δεδομένα*.

Πέραν όμως από τις 15 μέρες πριν από το πόστο εντοπισμού του μανιταριού, εν μέρει προβληματικές είναι και αυτές αμέσως μετά την αφαίρεση του. Ο λόγος είναι ότι μετά τις εργασίες που πραγματοποιούνται σε αυτή, η λεκάνη ενδεχομένως χρειάζεται κάποιο χρόνο για να επανέλθει στην κανονική λειτουργία της. Με αυτή τη λογική, δημιουργούνται 9 νέα σύνολα δεδομένων, τα οποία αναφέρονται ως *ασφαλή δεδομένα*, κάθε ένα από τα οποία βασίζεται στα καθαρά δεδομένα, αφαιρώντας επιπλέον τις παρατηρήσεις 1 έως 9 ημερών μετά από κάθε παρατήρηση με ετικέτα 1. Τα σύνολα αυτά συμβολίζονται ως S1-S9 (από τη λέξη *ασφαλή*, *Safe*) και ελέγχονται όλα, με την προσδοκία κάποιο από αυτά να δώσει καλύτερα αποτελέσματα, προφανώς όταν αυτά εξετάζονται με τον ίδιο ακριβώς αλγόριθμο, για λόγους σύγκρισης.

Σε αυτό το σημείο αξίζει να σημειωθεί πως τόσο στην περίπτωση των καθαρών δεδομένων, όσο και σε αυτή των ασφαλών, που είναι και αυτά που εξετάζονται διεξοδικά, δεν αφαιρούνται παρατηρήσεις με ετικέτα 1, παρά μόνο παρατηρήσεις με ετικέτα 0, που αντιστοιχούν δηλαδή σε απουσία μανιταριού. Η τάξη στην οποία ανήκουν οι τελευταίες παρατηρήσεις είναι η πιο κοινή τάξη, με αποτέλεσμα οι χειρισμοί για τη δημιουργία των ασφαλών δεδομένων να αποτελούν μια μέθοδο υποδειγματοληψίας, βασισμένη όχι σε κάποιον αλγόριθμο αλλά σε μια λογική εμπνευσμένη από τις γνώσεις μας για τη διεργασία. Με τον τρόπο αυτό αναιρείται μερικώς το πρόβλημα των Ανισοκατανεμημένων Δεδομένων που περιγράφηκε στην Εν. 2.3. Η παραδοχή που γίνεται, και η οποία πιθανότατα επαληθεύεται, είναι ότι οι παρατηρήσεις που αντιστοιχούν στην πιο κοινή τάξη είναι στην πραγματικότητα τόσες πολλές, που αφαιρώντας όσες αναφέρθηκε ουσιαστικά δεν προκύπτει απώλεια πληροφορίας.

Ο δεύτερος τρόπος μεταβολής των ετικετών, ο οποίος είναι πιο ριψοκίνδυνος, είναι στις αβέβαιες παρατηρήσεις να αποδοθεί η ετικέτα 1, υποθέτοντας ότι το μανιτάρι υπήρχε κάποιες μέρες πριν τον εντοπισμό του. Προκειμένου όμως να μη δημιουργηθούν παρατηρήσεις που θα φέρουν λανθασμένα την ετικέτα 1, αυτή αποδίδεται μόνο σε παρατηρήσεις χρονικά κοντά στο πόστο όπου εντοπίστηκε το μανιτάρι. Δοκιμαστικά, δημιουργούνται 9 επιπλέον σύνολα δεδομένων, τα οποία αναφέρονται ως *μη ασφαλή δεδομένα*, κάθε ένα από τα οποία βασίζεται στην εκδοχή των καθαρών δεδομένων με την καλύτερη επίδοση (προκαταβολικά αναφέρεται ότι αυτή είναι η S2). Σε αυτά, τα οποία συμβολίζονται ως U1-U9 (από τον όρο *μη ασφαλή*, *Unsafe*), η ετικέτα 1 αποδίδεται σε παρατηρήσεις για 1 έως 9 ημερών πριν από κάθε παρατήρηση με ετικέτα 1, δηλαδή πριν τον εντοπισμό του μανιταριού, αντίστοιχα.

3.5 Συνδυασμός Χαρακτηριστικών και Ετικετών

Οι ετικέτες από κάθε ένα εκ των 18 συνόλων ετικετών που δημιουργούνται, S1-S9 και U1-U9, συνδυάζονται με τα χαρακτηριστικά που έχουν προκύψει όπως περιγράφεται στις Εν. 3.2 και 3.3. Όλα τα πλήρη σύνολα δεδομένων που προκύπτουν τροφοδοτούνται στους αλγόριθμους Μηχανικής Μάθησης, με σκοπό να φανεί το αν η επίδοση του μοντέλου που θα βασιστεί σε κάποιο από αυτά θα ξεχωρίσει. Αυτό θα αποτελεί μια ένδειξη ότι η λογική πίσω από την κατασκευή του αντίστοιχου συνόλου είναι πιο κοντά στην πραγματικότητα. Φυσικά, το αποτέλεσμα κρίνεται τόσο με βάση τη λογική, όσο και με βάση την εμπειρία των αρμόδιων μηχανικών της Δραστηριότητας.

Κεφάλαιο 4

Εφαρμογή της Μηχανικής Μάθησης

Σε αυτό το κεφάλαιο αναπτύσσεται, περιεκτικά αλλά με την πρόθεση να μην λείπει καμία χρήσιμη πληροφορία, η μεθοδολογία που ακολουθείται κατά την εφαρμογή της Μηχανικής Μάθησης για την επίλυση του προβλήματος της μελέτης. Πολλά από τα εργαλεία και τις προσεγγίσεις που διερευνήθηκαν κατά τη διάρκεια της μελέτης απουσιάζουν από το κείμενο, καθώς οδήγησαν σε μετριότερο αποτέλεσμα από αυτά που παρουσιάζονται. Παρόλα αυτά, πραγματοποιείται μια σύντομη αναφορά σε αυτά με στόχο την ενημέρωση για το τι έχει δοκιμαστεί χωρίς επιτυχία, καθώς και τους κύριους λόγους για τους οποίους οι εναλλακτικές προσεγγίσεις απορρίφθηκαν υπέρ αυτών που τελικά αξιοποιήθηκαν.

4.1 Περίγραμμα της Μεθοδολογίας

Όλα τα τετράδια Jupyter που χρησιμοποιούνται για την εφαρμογή της μεθοδολογίας Μηχανικής Μάθησης παρουσιάζουν την ίδια μορφή. Συγκεκριμένα, περιλαμβάνουν τα εξής βήματα:

1. Φόρτωση των απαραίτητων βιβλιοθηκών.
2. Φόρτωση ενός εκ των συνόλων δεδομένων S1-S9 και U1-U9.
3. Διαχωρισμός του συνόλου δεδομένων σε έναν πίνακα που περιλαμβάνει τα χαρακτηριστικά και ένα διάνυσμα που περιλαμβάνει τις ετικέτες.
4. Περαιτέρω διαχωρισμός των χαρακτηριστικών και των ετικετών σε αυτά που ανήκουν στα σύνολα εκπαίδευσης, ελέγχου και διακρίβωσης (βλ. Εν. 2.4).

5. Αρχικοποίηση των *Ταξινομητών* (*Classifiers*), δηλαδή των αλγορίθμων Μηχανικής Μάθησης που πρόκειται να χρησιμοποιηθούν για την Ταξινόμηση, καθώς και ενός *Ψευδοταξινομητή* (*Dummy Classifier*) (βλ. παρακάτω, Εν. 4.3). Ο τελευταίος αξιοποιείται ως μέτρο σύγκρισης.

6. Αρχικοποίηση των *Δεικτών Αξιολόγησης* (*Evaluation Metrics*) που χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων (βλ. παρακάτω, Εν. 4.4).

7. Αρχικοποίηση μιας *Αναζήτησης Πλέγματος* (*Grid Search*), που επίσης αποτελεί ένα εργαλείο του *scikit-learn* για την βελτιστοποίηση των παραμέτρων των αλγορίθμων (βλ. παρακάτω, Εν. 4.5).

8. Εκτέλεση της Αναζήτησης Πλέγματος για την εκπαίδευση διαφορετικών μοντέλων, με διαφοροποιημένες τιμές των κρίσιμων παραμέτρων, και σύγκριση τους με βάση τους Δείκτες Αξιολόγησης.

9. Αποτύπωση των αποτελεσμάτων της Αναζήτησης Πλέγματος σε πίνακα ώστε να είναι διαθέσιμα για περαιτέρω ανάλυση.

10. Εφαρμογή του βέλτιστου μοντέλου που προέκυψε από την διαδικασία που προηγήθηκε στο σύνολο διακρίβωσης και υπολογισμός της επίδοσής του, με βάση ορισμένους από τους δείκτες.

Όπως φαίνεται στην παραπάνω λίστα, πολλά από τα εργαλεία που αξιοποιούνται αναλύονται στις επόμενες ενότητες, ώστε να γίνει κατανοητή η λειτουργία τους και ο τρόπος με τον οποίο συνδυάζονται.

4.2 Διαχωρισμός σε Σύνολα Εκπαίδευσης, Ελέγχου και Διακρίβωσης

Στο Κεφ. 2, όπου εισήχθησαν ορισμένες γενικές έννοιες που αφορούν τη Μηχανική Μάθηση, έγινε μια αναφορά στα Σύνολα Εκπαίδευσης, Ελέγχου και Διακρίβωσης. Πιο συγκεκριμένα, αναφέρθηκε ότι ο διαχωρισμός των δεδομένων σε αυτά τα τρία σύνολα είναι απαραίτητος για να εξεταστεί η ικανότητα ενός παραγόμενου μοντέλου να γενικεύει, δηλαδή να προβλέπει με επιτυχία άγνωστες εισόδους, οι οποίες όμως έχουν παρόμοια χαρακτηριστικά με εκείνες βάσει των οποίων έχει εκπαιδευτεί. Στο σημείο αυτό είναι απαραίτητη περαιτέρω εμβάθυνση στο θέμα με την εισαγωγή στη μέθοδο της *αντεπικύρωσης* (*cross-validation*).

Η *αντεπικύρωση*, που στο εξής θα αναφέρεται αποκλειστικά με τον αγγλικό όρο, *cross-validation*, είναι μια στατιστική μέθοδος αξιολόγησης της ικανότητας ενός μοντέλου να γενικεύει, η οποία είναι πιο ενδελεχής και αξιόπιστη από τον απλό διαχωρισμό του σε σύνολα εκπαίδευσης και ελέγχου. Στο cross-validation, τα δεδομένα χωρίζονται επαναληπτικά και παράγονται περισσότερα από ένα μοντέλα, ένα για κάθε διαχωρισμό τους. Η πιο συνήθης εκδοχή του είναι το *k-fold cross-validation*, όπου το k είναι ένας θετικός ακέραιος που επιλέγεται κατά περίπτωση, π.χ. 5 ή 10.

Κατά την εφαρμογή ενός 5-fold cross validation, τα δεδομένα χωρίζονται σε 5 υποσύνολα κατά προσέγγιση ίσου πλήθους παρατηρήσεων, τα οποία ονομάζονται *folds*. Στη συνέχεια, εκπαιδεύονται διαδοχικά μια σειρά από 5 μοντέλα. Το πρώτο μοντέλο εκπαιδεύεται χρησιμοποιώντας το πρώτο υποσύνολο ως σύνολο ελέγχου, και τα υπόλοιπα (2-5) ως σύνολο εκπαίδευσης. Συνεπώς, το μοντέλο παράγεται από τα υποσύνολα 2-5 και αξιολογείται με βάση το 1. Έπειτα, ένα νεό μοντέλο παράγεται, αυτή τη φορά χρησιμοποιώντας το 2 ως σύνολο ελέγχου και τα 1 και 3-5 ως σύνολο εκπαίδευσης κ.ο.κ. Αυτή η διαδικασία επαναλαμβάνεται χρησιμοποιώντας τα υποσύνολα 3, 4 και 5 ως σύνολα ελέγχου. Για κάθε διαχωρισμό των δεδομένων, υπολογίζεται η αξιοπιστία του αντίστοιχου μοντέλου, οπότε τελικά προκύπτουν 5 τιμές του δείκτη που την εκφράζει, μία για κάθε μοντέλο.

Όπως θα εξηγηθεί αναλυτικά παρακάτω, η εφαρμογή του αλγορίθμου Μηχανικής Μάθησης σε αυτή τη μελέτη γίνεται με 5-fold cross-validation, στο πλαίσιο μιας γενικότερης διαδικασίας για την βέλτιστη επιλογή παραμέτρων. Για το λόγο αυτό, όλα τα δεδομένα με τα οποία θα τροφοδοτηθεί ο αλγόριθμος θα χρησιμοποιηθούν κάποια στιγμή ως δεδομένα εκπαίδευσης. Κατά συνέπεια, δεν είναι δυνατό να χρησιμοποιηθούν και για την αξιολόγηση του μοντέλου. Ένας τρόπος να επιλυθεί το πρόβλημα που ανακύπτει είναι τα αρχικά δεδομένα να χωριστούν προκαταβολικά σε δύο υποσύνολα, ένα για τη μοντελοποίηση, από το οποίο θα προκύψουν τα σύνολα εκπαίδευσης και ελέγχου, και άλλο ένα για την αξιολόγηση, το οποίο και είναι το *σύνολο διακρίβωσης* (*validation set*).

Όλοι οι διαχωρισμοί που περιγράφονται παραπάνω εκτελούνται χρησιμοποιώντας τη μέθοδο `train_test_split` της βιβλιοθήκης *scikit-learn*. Η μέθοδος αυτή δέχεται ως είσοδο τα σύνολα που αφορούν στα χαρακτηριστικά και τις ετικέτες, X και y, αντίστοιχα, τα οποία προέκυψαν από το 3ο βήμα της μεθοδολογίας που σκιαγραφήθηκε στην Εν. 4.1. Επιπλέον, διαθέτει τις εξής παραμέτρους που σχετίζονται με την παρούσα μελέτη:

- `train_size` (ή, εναλλακτικά, `test_size`): Το μέγεθος του συνόλου εκπαίδευσης (ή ελέγχου). Χρησιμοποιείται για να ορίσει είτε το πλήθος των παρατηρήσεων, αν εισαχθεί ένας θετικός ακέραιος, είτε το ποσοστό επί των συνολικών παρατηρήσεων, αν εισαχθεί ένας πραγματικός μεταξύ 0 και 1, που θα τοποθε-

τηθούν στο σύνολο εκπαίδευσης, Τα αντίστοιχα ισχύουν για την εναλλακτική επιλογή που αφορά στο σύνολο ελέγχου. Η επιλογή είναι γενικά αυθαίρετη, οπότε διατηρείται η προεπιλογή, που είναι 0.75 και αποτελεί μια συνήθη επιλογή.

- `shuffle`: Καθορίζει το αν τα δεδομένα θα ανακατευτούν πριν το διαχωρισμό τους στα δύο σύνολα. Επειδή η πληροφορία που περιείχαν τα δεδομένα ως χρονοσειρά έχει εισαχθεί σαν αυτοτελή χαρακτηριστικά και η έννοια του χρόνου έχει καταργηθεί, αυτή ορίζεται ως `True`.
- `stratify`: Καθορίζει το αν κατά το διαχωρισμό θα διατηρηθεί η ίδια συχνότητα εμφάνισης της κάθε τάξης στις ετικέτες στα δύο σύνολα που θα προκύψουν. Αυτό είναι επιθυμητό και για να γίνει στην παράμετρο πρέπει να διαβιβαστεί το y, που περιέχει τις ετικέτες.
- `random_state`: Καθορίζει την αρχικοποίηση της γεννήτριας ψευδοτυχαίων αριθμών με την οποία πραγματοποιείται η «τυχαία» τοποθέτηση των δεδομένων στα δύο σύνολα, καθιστώντας το αποτέλεσμα ντετερμινιστικό, ώστε να υπάρχει επαναληψιμότητα στα (υπολογιστικά) πειράματα που εκτελούνται. Τυχαία επιλέγεται εδώ η τιμή 42.

Αυτό που προκύπτει από την εφαρμογή της μεθόδου στα X και y είναι τα σύνολα X_model, y_model, X_valid και y_valid. Τα δύο πρώτα θα χρησιμοποιηθούν κατά τη διαδικασία της αντεπικύρωσης για την παραγωγή των διαφορετικών συνόλων εκπαίδευσης και ελέγχου, ενώ τα δύο τελευταία θα χρησιμοποιηθούν κατά την τελική αξιολόγηση του μοντέλου που θα επιλεγεί ως το βέλτιστο μεταξύ αυτών που θα ελεγχθούν.

4.3 Αλγόριθμοι Ταξινόμησης

Το κύριο στοιχείο μιας μεθοδολογίας Μηχανικής Μάθησης είναι ο αλγόριθμος που αξιοποιείται για την κατασκευή των μοντέλων. Στην περίπτωση που το πρόβλημα που προσεγγίζεται είναι ένα πρόβλημα Ταξινόμησης, ο αλγόριθμος που χρησιμοποιείται καλείται *Αλγόριθμος Ταξινόμησης* ή, απλούστερα, *Ταξινομητής* (*Classifier*). Στη συνέχεια της τρέχουσας ενότητας αναλύονται οι Ταξινομητές που παρουσιάζονται μεταξύ αυτών που δοκιμάστηκαν στην παρούσα μελέτη, πρώτα όμως πραγματοποιείται μια σύντομη αναφορά σε μια ειδική περίπτωση Ταξινομητή.

4.3.1 Ψευδοταξινομητές

Ένας *Ψευδοταξινομητής* (*Dummy Classifier*), όπως υποδεικνύεται και από το όνομά του, δεν αποτελεί έναν πραγματικό Ταξινομητή. Ουσιαστικά, πρόκειται περί ενός μοντέλου που πραγματοποιεί προβλέψεις, χρησιμοποιώντας απλούς κανόνες. Είναι πολύ σημαντικό να τονιστεί ότι στις προβλέψεις του αυτές, ανεξάρτητα από το με ποιον κανόνα λειτουργεί, δε λαμβάνει καθόλου υπόψη τα χαρακτηριστικά. Κατά περίπτωση λαμβάνει υπόψη του αποκλειστικά τις ετικέτες, οπότε στην ουσία δεν *εκπαιδεύεται*, ούτε και *μαθαίνει*. Η χρηστική του αξία έγκειται στο ότι μπορεί να χρησιμοποιηθεί ως *γραμμή βάσης* (*baseline*), δηλαδή ως μέτρο σύγκρισης για τους πραγματικούς Ταξινομητές, οι οποίοι προφανώς απαιτείται να είναι οπωσδήποτε καλύτεροί του.

Η βιβλιοθήκη *scikit-learn* παρέχει ένα Ψευδοταξινομητή, με τη μορφή του αντικειμένου `DummyClassifier`. Οι παράμετροι του που χρησιμοποιούνται στην παρούσα μελέτη είναι οι εξής:

- `strategy`: Η στρατηγική, δηλαδή ο απλός κανόνας που χρησιμοποιείται για την πραγματοποίηση των προβλέψεων. Μεταξύ των διάφορων επιλογών χρησιμοποιούνται αυτές που δίνουν τα καλύτερα αποτελέσματα, συγκεκριμένα οι `uniform` και `stratified`. Όπως ίσως είναι εμφανές, η πρώτη από τις δύο οδηγεί σε ομοιόμορφα τυχαίες προβλέψεις, δηλαδή τυχαία πρόβλεψη τάξης με πιθανότητες 50-50, ενώ η δεύτερη σε επίσης τυχαίες προβλέψεις, για τις οποίες λαμβάνεται υπόψη η συχνότητα των δύο τάξεων. Αυτό σημαίνει πως οι πιθανότητα να προβλεφθεί τυχαία η πιο συχνή τάξη είναι συντριπτικά μεγαλύτερη, εφόσον τα δεδομένα είναι Ανισοκατανεμημένα.
- `random_state`: Όμοίως με την περίπτωση της μεθόδου `train_test_split`, αλλά και κάθε άλλη που θα ακολουθήσει, η παράμετρος αυτή καθορίζει ρητά την αρχικοποίηση της γεννήτριας ψευδοτυχαίων αριθμών. Η παράμετρος αυτή δε θα συζητηθεί περαιτέρω, ενώ η τιμή που της δίνεται είναι πάντα 42.

Από τα παραπάνω γίνεται εμφανές ότι ο Ψευδοταξινομητής δε λαμβάνει καθόλου υπόψη του τα δεδομένα, μόνο τη συχνότητα των ετικετών στο σύνολο εκπαίδευσης. Στη συνέχεια αναλύονται οι πραγματικοί Ταξινομητές που αξιοποιούνται στην παρούσα μελέτη.

4.3.2 Δέντρα Αποφάσεων (Decision Trees)

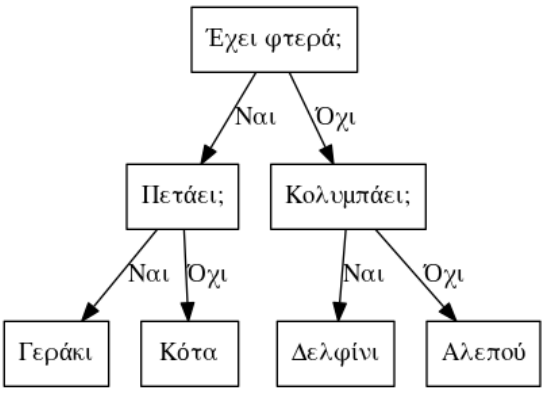
Στην πορεία της μελέτης, το πρόβλημα προσεγγίστηκε με διάφορους αλγορίθμους Μηχανικής Μάθησης. Μεταξύ αυτών διαπιστώθηκε ότι τη βέλτιστη συμπεριφορά παρουσιάζουν αλγόριθμοι βασισμένοι σε *Δέντρα Αποφάσεων* [2, 6, 15, 16, 17]. Επειδή

η φύση των ετικετών δημιούργησε την ανάγκη να εξεταστούν διαφορετικά σύνολα δεδομένων, στο κείμενο που ακολουθεί περιγράφονται για οικονομία το θεωρητικό υπόβαθρο και τα αποτελέσματα που αφορούν μόνο σε αυτούς τους αλγορίθμους.

Τα *Δέντρα Αποφάσεων* (*Decision Trees*) είναι μοντέλα που χρησιμοποιούνται ευρέως τόσο για προβλήματα Ταξινόμησης όσο και Παλινδρόμησης. Πολύ επιγραμματικά, αποτελούνται από μια αλληλουχία συνθηκών if/else οι οποίες οδηγούν στην τελική πρόβλεψη ή *απόφαση*. Γι΄ αυτό και κατά κάποιο τρόπο προσομοιάζουν τα «έξυπνα συστήματα» που παραδοσιακά χρησιμοποιούνται αντί της Μηχανικής Μάθησης, άρα και το σύστημα που προς το παρόν αντιμετωπίζει το πρόβλημα της μελέτης.

Αναλυτικότερα, ένα Δέντρο Αποφάσεων μπορεί να παρομοιαστεί με το ίσως γνώριμο «Παιχνίδι των 20 Ερωτήσεων». Στόχος του παιχνιδιού είναι να εξαχθεί μια άγνωστη πληροφορία, ρωτώντας τις λιγότερες δυνατές ερωτήσεις που να επιδέχονται μονολεκτική απάντηση, δηλαδή «ναι» ή «όχι». Για παράδειγμα, το ζητούμενο θα μπορούσε να είναι να βρεθεί το άγνωστο ζώο μεταξύ των παρακάτω: κότα, αλεπού, γεράκι και δελφίνι. Σε αυτή την περίπτωση, η πρώτη ερώτηση θα μπορούσε να είναι αν το ζώο έχει φτερά ή όχι. Ανεξάρτητα από την απάντηση, η ερώτηση αυτή ελαττώνει τα υποψήφια ζώα σε μόλις δύο: γεράκι και κότα αν η απάντηση είναι «ναι», αλεπού και δελφίνι αν η απάντηση είναι «όχι». Πλέον αρκεί μόνο μία ερώτηση, π.χ. αν το ζώο πετάει για να γίνει η διάκριση μεταξύ των ζώων της πρώτης ομάδας και αν κολυμπάει για να γίνει η διάκριση μεταξύ αυτών της δεύτερης. Η παραπάνω συλλογιστική πορεία προς την ορθή πρόβλεψη απεικονίζεται στο Σχ. 4.1.

Στην απεικόνιση αυτού του απλού δέντρου φαίνεται ότι κάθε κόμβος του, ο οποίος αναπαρίσταται με ένα κουτί, είτε περιέχει μια ερώτηση, είτε αποτελεί *τερματικό κόμβο* (*terminal node*), περιέχει δηλαδή μια απάντηση. Ένας τερματικός κόμβος συχνά



Σχήμα 4.1: Ένα απλό Δέντρο Αποφάσεων που βρίσκει το ζητούμενο μεταξύ 4 ζώων.

καλείται και *φύλλο* (*leaf*). Με βέλη πρακτικά αναπαρίστανται οι απαντήσεις σε ενδιάμεσες ερωτήσεις, οι οποίες οδηγούν σε μια νέα ερώτηση ή στην τελική απάντηση.

Η αντιστοιχία μεταξύ του παραδείγματος και της Μηχανικής Μάθησης είναι εμφανής. Μέσα από την παραπάνω διαδικασία, προέκυψε ένα μοντέλο ικανό να λύσει ένα πρόβλημα Ταξινόμησης και συγκεκριμένα να διακρίνει μεταξύ τεσσάρων τάξεων (κότα, αλεπού, γεράκι, δελφίνι). Αυτό πραγματοποιείται με βάση τρία χαρακτηριστικά («έχει φτερά», «πετάει», «κολυμπάει»). Μοντέλα σαν κι αυτό μπορούν να εξαχθούν αυτόματα από τα δεδομένα με μια διαδικασία Επιτηρούμενης Μάθησης, αντί να κατασκευαστούν χειροκίνητα από έναν άνθρωπο όπως έγινε σε αυτήν εδώ την περίπτωση.

Σημειώνεται πως στην περίπτωση πολλών αλγορίθμων Μηχανικής Μάθησης, η μορφή των δεδομένων με την οποία αυτά εισάγονται στον αλγόριθμο παίζει πολύ σημαντικό ρόλο στην ποιότητα του τελικού αποτελέσματος. Αυτό δεν γίνεται ιδιαίτερα εμφανές στην περίπτωση των αλγορίθμων που βασίζονται σε Δέντρα Αποφάσεων, όπως και τα Τυχαία Δάση που ακολουθούν, καθώς αυτή η κατηγορία αλγορίθμων έχει μάλλον τις χαμηλότερες απαιτήσεις. Για παράδειγμα, δεν απαιτεί *Αναγωγή των χαρακτηριστικών σε Κοινή Κλίμακα* (Scaling) ή *Επιλογή Χαρακτηριστικών* (*Feature Selection*), ενώ στην περίπτωση της μεθόδου του Ταξινομητή του *scikit-learn* που χρησιμοποιείται εδώ, δεν απαιτεί ούτε και Αναδειγματοληψία. Αυτό σημαίνει πως τα δεδομένα μπορούν να τροφοδοτηθούν στον αλγόριθμο στη μορφή που βρίσκονται μετά από την προεπεξεργασία που τους γίνεται ούτως ή άλλως.

4.3.3 Κατασκευή των Δέντρων Αποφάσεων

Στη Μηχανική Μάθηση, οι ερωτήσεις if/else που απαρτίζουν το Δέντρο Αποφάσεων ονομάζονται *δοκιμές* (*tests*). Συχνά, τα χαρακτηριστικά δεν είναι δυαδικά, δεν περιλαμβάνουν δηλαδή δύο περιπτώσεις, αλλά είναι συνεχή, δηλαδή λαμβάνουν τιμές από το σύνολο των πραγματικών αριθμών. Προφάνως, σε αυτή την περίπτωση οι δοκιμές είναι ερωτήσεις της μορφής «Είναι η τιμή του χαρακτηριστικού i μεγαλύτερη από την οριακή τιμή α;». Αυτό ισχύει και για το πρόβλημα που αντιμετωπίζει αυτή η μελέτη.

Για να κατασκευάσει ένα δέντρο, ο αλγόριθμος διερευνά όλες τις υποψήφιες δοκιμές και εντοπίζει αυτή που παρέχει την περισσότερη πληροφορία για τη μεταβλητή-στόχο. Ο ανώτατος κόμβος του δέντρου, ο οποίος ονομάζεται και *ρίζα* (*root*), περιλαμβάνει το σύνολο των δεδομένων εκπαίδευσης. Η πρώτη δοκιμή ελέγχει αν η τιμή κάποιου χαρακτηριστικού i, x_i, είναι μεγαλύτερη από μια σταθερή τιμή x_i^0, ελέγχει δηλαδή την ισχύ της συνθήκης $x_i > \mathrm{x}_i^0$ για κάθε παρατήρηση του συνόλου εκπαίδευσης. Από αυτή προκύπτουν δύο νέοι κόμβοι, δηλαδή δύο συμπληρωματικά υποσύνολα του συνόλου εκπαίδευσης. Στην περίπτωση της Δυαδικής Ταξινόμησης, με την οποία και

ασχολείται η μελέτη, το ιδανικό είναι κάθε ένα από αυτά τα υποσύνολα να περιέχει παρατηρήσεις που ανήκουν σε μία μόνο από τις δύο τάξεις, δηλαδή οι τάξεις να διαχωριστούν πλήρως. Αυτό γενικά δε συμβαίνει, οπότε μεταξύ των δοκιμών επιλέγεται αυτή που πετυχαίνει το μεγαλύτερο διαχωρισμό.

Η κατασκευή του δέντρου είναι λοιπόν μια επαναληπτική διαδικασία, στην οποία σε κάθε βήμα εξετάζεται μεγάλο πλήθος δοκιμών, οι οποίες προκύπτουν με βάση τα διάφορα χαρακτηριστικά και διαφορετικές οριακές τιμές τους, ώστε μεταξύ τους να εντοπιστεί η καλύτερη. Η διαδικασία αυτή μπορεί να συνεχιστεί έως ότου κάθε φύλλο του δέντρου περιλαμβάνει μία μόνο τιμή της μεταβλητής-στόχου, π.χ. μία μόνο τάξη. Ένα τέτοιο φύλλο καλείται *καθαρό* (*pure*). Σε περίπτωση που διακοπεί πρόωρα, ορισμένα φύλλα θα περιέχουν παρατηρήσεις και από τις δύο τάξεις. Η πρόβλεψη της τάξης μιας άγνωστης παρατήρησης πραγματοποιείται εκτελώντας διαδοχικά τις δοκιμές του δέντρου μέχρι το τελευταίο επίπεδο, στο οποίο και αποδίδεται είτε η τάξη του τελικού φύλλου, αν αυτό είναι καθαρό, είτε η τάξη της πλειοψηφίας των παρατηρήσεων που περιλαμβάνονται σε ένα μη καθαρό φύλλο.

Για λόγους πληρότητας αναφέρεται ότι η ίδια διαδικασία με αυτή που περιγράφηκε μπορεί να εφαρμοστεί και για προβλήματα Παλινδρόμησης. Τότε, το αποτέλεσμα της πρόβλεψης είναι είτε η τιμή της μεταβλητής-στόχου για τη μοναδική παρατήρηση ενός καθαρού φύλλου, είτε ο μέσος όρος των τιμών για τις παρατηρήσεις ενός μη καθαρού.

4.3.3.1 Μεγιστοποίηση του Κέρδους Πληροφορίας

Ένα ερώτημα που προκύπτει είναι το πως επιλέγεται η καλύτερη μεταξύ των πιθανών δοκιμών, δηλαδή το πως επιλέγεται το χαρακτηριστικό και η οριακή τιμή του που θα επιτρέψουν τον βέλτιστο διαχωρισμό μεταξύ των δύο τάξεων σε κάθε κόμβο. Προφανώς, πρέπει να οριστεί μια αντικειμενική συνάρτηση η οποία βελτιστοποιείται μέσω του αλγορίθμου κατασκευής του Δέντρου Αποφάσεων. Στη συγκεκριμένη περίπτωση επιδιώκεται η μεγιστοποίηση του συνάρτησης *Κέρδους Πληροφορίας* (*Information Gain, IG*),

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{4.1}$$

που προκύπτει από κάθε διαχωρισμό, όπου

- f: το χαρακτηριστικό με βάση το οποίο πραγματοποιείται ο διαχωρισμός,
- D_p, D_j: το σύνολο των παρατηρήσεων του αρχικού κόμβου και του j-οστού κόμβου από αυτούς που προκύπτουν από τη δοκιμή,

- I: η συνάρτηση εκτίμησης του *βαθμού πρόσμιξης* (*impurity*) σε ένα σύνολο, η οποία ορίζεται παρακάτω, και
- N_p, N_j: το πλήθος των παρατηρήσεων στον αρχικό κόμβο και στον j-οστό κόμβο από αυτούς που προκύπτουν από τη δοκιμή, αντίστοιχα.

Όπως είναι εμφανές από τη σχέση, το Κέρδος Πληροφορίας δεν είναι τίποτα άλλο παρά η διαφορά μεταξύ του βαθμού πρόσμιξης του αρχικού κόμβου και του αθροίσματος των βαθμών πρόσμιξης των κόμβων που προκύπτουν. Αυτό σημαίνει πως όσο μικρότεροι είναι οι τελευταίοι, τόσο μεγαλύτερο το κέρδος πληροφορίας από το διαχωρισμό.

Συνήθως, για λόγους απλότητας και να περιοριστεί σε ένα λογικό πλαίσιο ο χώρος στον οποίο αναζητείται η καλύτερη δοκιμή, οι περισσότερες βιβλιοθήκες που περιέχουν εφαρμογές αλγορίθμων Μηχανικής Μάθησης, συμπεριλαμβανομένου και του scikit-learn, αξιοποιούν δυαδικά Δέντρα Αποφάσεων. Αυτό σημαίνει πως κάθε αρχικός κόμβος διαιρείται σε δύο επιμέρους κόμβους, τον αριστερό και τον δεξί, οι οποίοι συμβολίζονται με D_{left} και D_{right}, αντίστοιχα. Σε αυτή την περίπτωση, η Εξ. 4.1 παίρνει τη μορφή

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \tag{4.2}$$

Τα τρία μέτρα του βαθμού πρόσμιξης (ή κριτήρια διαχωρισμού) που χρησιμοποιούνται στα δυαδικά Δέντρα Αποφάσεων είναι η *πρόσμιξη Gini* (*Gini impurity*), I_G, η *εντροπία* (*entropy*), I_H και το *σφάλμα ταξινόμησης* (*classification error*), I_E.

Για όλες τις μη κενές τάξεις i που απαντώνται σε έναν κόμβο t, δηλαδή αυτές για τις οποίες το ποσοστό στο οποίο απαντώνται στις παρατηρήσεις του κόμβου είναι διάφορο του μηδενός, $p(i|t) \neq 0$, η εντροπία ορίζεται ως

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t) \tag{4.3}$$

όπου c το πλήθος των τάξεων. Εύκολα φαίνεται ότι η εντροπία είναι 0 αν όλες οι παρατηρήσεις σε έναν κόμβο ανήκουν στην ίδια τάξη, ενώ λίγο πιο δύσκολα υπολογίζεται πως η εντροπία μεγιστοποιείται αν υπάρχει ομοιόμορφη κατανομή των παρατηρήσεων στις διάφορες τάξεις. Στην περίπτωση της Δυαδικής Ταξινόμησης, η εντροπία είναι 0 αν $p(i = 1|t) = 1$ ή $p(i = 0|t) = 1$. Αν οι παρατηρήσεις είναι ομοιόμορφα κατανεμημένες και $p(i = 1|t) = 0.5$, $p(i = 0|t) = 0.5$, τότε η εντροπία ισούται με 1.

Η πρόσμιξη Gini δίνεται από τη σχέση

$$I_G(t) = \sum_{i=1}^{c} p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{4.4}$$

Αντιστοιχα με την εντροπία, το μέγεθος αυτό μεγιστοποιείται αν οι τάξεις είναι πλήρως αναμεμιγμένες, π.χ. στην περίπτωση της Δυαδικής Ταξινόμησης

$$I_G(t) = 1 - \sum_{i=1}^{2} 0.5^2 = 0.5$$

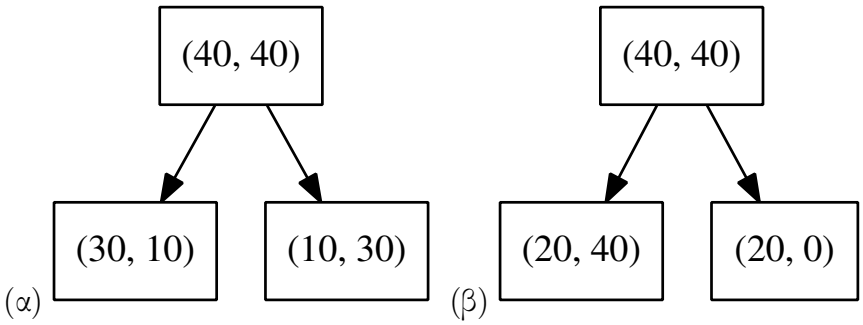
Παρά τη διαφορά στη μέγιστη τιμή, στην πράξη και τα δύο μεγέθη δίνουν παρόμοια αποτελέσματα, έτσι ώστε συχνά δεν έχει αξία να ελεγχθεί το αν επιτυγχάνονται καλύτερα αποτελέσματα χρησιμοποιώντας την τιμή του ενός ή του άλλου ως κριτήριο.

Τέλος, για το σφάλμα ταξινόμησης ισχύει ότι

$$I_E = 1 - \max p(i|t) \tag{4.5}$$

Το κριτήριο αυτό συχνά είναι κατάλληλο για το *κλάδεμα* (*pruning*) ενός δέντρου, το οποίο περιγράφεται λίγο παρακάτω, αλλά όχι για την ανάπτυξη ενός δέντρου, καθώς δεν είναι τόσο ευαίσθητο στις μεταβολής της πιθανότητας εύρεσης κάθε τάξης στους κόμβους.

Όλα τα παραπάνω μπορούν να γίνουν αντιληπτά μέσα από ένα αριθμητικό παράδειγμα. Ας υποτεθεί ότι επιχειρείται η εύρεση του βέλτιστου σεναρίου διαχωρισμού μεταξύ των δύο που εικονίζονται στο Σχ. 4.2. Ο αρχικός κόμβος περιέχει το σύνολο παρατηρήσεων D_p, το οποίο αποτελείται από 40 παρατηρήσεις από δύο τάξεις, έστω 1 και 2, και χωρίζεται σε δύο επιμέρους σύνολα, D_{left} και D_{right}. Το Κέρδος Πληροφορίας, το οποίο επιδιώκεται να μεγιστοποιηθεί, για την περίπτωση που χρησιμοποιείται ως



Σχήμα 4.2: Δύο πιθανά σενάρια διαχωρισμού ενός κόμβου για διαφορετικές δοκιμές.

κριτήριο το σφάλμα ταξινόμησης είναι το ίδιο και για τα δύο σενάρια (α) και (β) και ίσο με $IG_E = 0.25$:

$$
\begin{aligned}
& I_E(D_p) = 1 - 0.5 = 0.5 \\
(\alpha): \; & I_E(D_{left}) = 1 - \frac{30}{40} = 0.25 \\
& I_E(D_{right}) = 1 - \frac{30}{40} = 0.25 \\
& IG_E = 0.5 - \frac{40}{80} \times 0.25 - \frac{40}{80} \times 0.25 = 0.25 \\
(\beta): \; & I_E(D_{left}) = 1 - \frac{40}{60} = \frac{1}{3} \\
& I_E(D_{right}) = 1 - \frac{20}{20} = 0 \\
& IG_E = 0.5 - \frac{60}{80} \times \frac{1}{3} - \frac{20}{80} \times 0 = 0.25
\end{aligned}
$$

Αντίθετα, τα κριτήρια που βασίζονται τόσο στην πρόσμιξη Gini όσο και στην εντροπία υποδεικνύουν ως καλύτερο το διαχωρισμό του σεναρίου (β). Για την πρόσμιξη Gini ισχύουν τα εξής:

$$
\begin{aligned}
& I_G(D_p) = 1 - (0.5^2 + 0.5^2) = 0.5 \\
(\alpha): \; & I_G(D_{left}) = 1 - \left(\left(\frac{30}{40} \right)^2 + \left(\frac{10}{40} \right)^2 \right) = 0.375 \\
& I_G(D_{right}) = 1 - \left(\left(\frac{10}{40} \right)^2 + \left(\frac{30}{40} \right)^2 \right) = 0.375 \\
& IG_G = 0.5 - \frac{40}{80} \times 0.375 - \frac{40}{80} \times 0.375 = 0.125 \\
(\beta): \; & I_G(D_{left}) = 1 - \left(\left(\frac{20}{60} \right)^2 + \left(\frac{40}{60} \right)^2 \right) = \frac{4}{9} \\
& I_G(D_{right}) = 1 - \left(\left(\frac{20}{20} \right)^2 + \left(\frac{0}{20} \right)^2 \right) = 0 \\
& IG_G = 0.5 - \frac{60}{80} \times \frac{4}{9} - 0 = 0.1\bar{6}
\end{aligned}
$$

Αντίστοιχα, για την εντροπία υπολογίζεται ότι:

$$
\begin{aligned}
& I_H(D_p) = -(0.5\log_2(0.5) + 0.5\log_2(0.5)) = 1 \\
(\alpha): \; & I_H(D_{left}) = 1 - \left(\frac{30}{40}\log_2\left(\frac{30}{40}\right) + \frac{10}{40}\log_2\left(\frac{10}{40}\right)\right) = 0.81 \\
& I_H(D_{right}) = 1 - \left(\frac{10}{40}\log_2\left(\frac{10}{40}\right) + \frac{30}{40}\log_2\left(\frac{30}{40}\right)\right) = 0.81 \\
& IG_H = 0.5 - \frac{40}{80} \times 0.81 - \frac{40}{80} \times 0.81 = 0.19 \\
(\beta): \; & I_H(D_{left}) = 1 - \left(\frac{20}{60}\log_2\left(\frac{20}{60}\right) + \frac{40}{60}\log_2\left(\frac{40}{60}\right)\right) = 0.92 \\
& I_H(D_{right}) = 0 \\
& IG_H = 0.5 - \frac{60}{80} \times 0.92 - 0 = 0.31
\end{aligned}
$$

Σύμφωνα με τα όσα αναφέρθηκαν για τα διάφορα κριτήρια, στην παρούσα εργασία παρουσιάζονται αποτελέσματα όπου χρησιμοποιείται η πρόσμιξη Gini, ενώ στο πλαίσιο της μελέτης πραγματοποιήθηκαν δοκιμές και με την εντροπία, χωρίς όμως κάποια ουσιαστική διαφορά στα μοντέλα που προέκυψαν.

4.3.3.2 Κλάδεμα

Στη γενική περίπτωση, η κατασκευή ενός δέντρου εξαντλώντας την προσπάθεια να προκύψουν καθαρά φύλλα οδηγεί σε μοντέλα με υψηλό βαθμό περιπλοκότητας που υπερπροσαρμόζονται στα δεδομένα εκπαίδευσης. Η παρουσία καθαρών φύλλων υποδηλώνει ότι το δέντρο είναι 100% εύστοχο στο σύνολο εκπαίδευσης, όμως το γεγονός αυτό δε το καθιστά απαραίτητα αποτελεσματικό ούτε στο σύνολο ελέγχου, αλλά ούτε και στις άγνωστες εισόδους. Για την αντιμετώπιση της υπερπροσαρμογής μπορούν να αξιοποιηθούν δύο στρατηγικές: το *κλάδεμα* του δέντρου είτε *εκ των προτέρων*, είτε *εκ των υστέρων*. Το *εκ των προτέρων κλάδεμα* (*pre-pruning*) είναι η πρόωρη διακοπή της ανάπτυξης του δέντρου, πριν από τη δημιουργία καθαρών φύλλων. Αυτή μπορεί να πραγματοποιηθεί περιορίζοντας το πλήθος των επιπέδων του δέντρου, το πλήθος των φύλλων του ή ορίζοντας ένα ελάχιστο πλήθος παρατηρήσεων που πρέπει να περιέχονται σε ένα κόμβο για να συνεχίσει αυτός να διαιρείται. Η δεύτερη στρατηγική, δηλαδή το *εκ των υστέρων κλάδεμα* (*post-pruning*), είναι η αφαίρεση κόμβων και κατά συνέπεια επιπέδων που περιέχουν μικρό όγκο πληροφορίας.

Ο αλγόριθμος με τον οποίο κατασκευάζονται Δέντρα Αποφάσεων από το scikit-learn, το οποίο αξιοποιείται στην παρούσα μελέτη, εφαρμόζει μόνο εκ των προτέρων κλάδεμα

των δέντρων, οπότε αυτή είναι και η στρατηγική που αναλύεται περισσότερο στη συνέχεια. Οι παράμετροι του αλγορίθμου αναλύονται στο κεφάλαιο που σχετίζεται με τους υπολογισμούς και τα αποτελέσματά τους.

4.3.4 Ανάλυση των Δέντρων Αποφάσεων

Ένα χαρακτηριστικό των Δέντρων Αποφάσεων που τους δίνει ένα συγκριτικό πλεονέκτημα έναντι άλλων μοντέλων είναι ότι οπτικοποιούνται. Με τον τρόπο αυτό, τα μοντέλα που παράγονται μπορούν να παρουσιαστούν και να επεξηγηθούν εύκολα ακόμη και σε ένα ακροατήριο μη σχετικό με το πεδίο. Αυτό φυσικά ισχύει για δέντρα με περιορισμένο βάθος, καθώς είναι δύσκολο να παρακολουθήσει κανείς ακόμη και δέντρα με μόλις 4 επίπεδα. Δέντρα με μεγαλύτερο βάθος, π.χ. με 10 επίπεδα, τα οποία είναι αρκετά κοινά, είναι ακόμη πιο δύσκολο να κατανοηθούν. Μια λύση ανάγκης είναι κανείς να παρακολουθήσει τη διαδρομή που συνδέει τους κόμβους με τις περισσότερες παρατηρήσεις για κάθε επίπεδο, οι οποίοι είναι και οι πιο χαρακτηριστικοί για τον τρόπο με τον οποίο λειτουργεί το μοντέλο.

Ένας εναλλακτικός τρόπος ανάλυσης ενός Δέντρου Αποφάσεων είναι αξιοποιώντας μεγέθη που συνοψίζουν τον τρόπο λειτουργίας του μοντέλου. Τέτοια μεγέθη π.χ. υπολογίζονται αυτόματα και είναι διαθέσιμα μέσα από το περιβάλλον του scikit-learn. Το πιο κοινό καλείται *σημαντικότητα των χαρακτηριστικών* (*feature importance*) και, όπως υποδεικνύει και το όνομα του, αποτελεί μια βαθμολογία του πόσο σημαντικό είναι κάθε χαρακτηριστικό για την πρόβλεψη που πραγματοποιεί το μοντέλο. Η τιμή της σημαντικότητας για κάθε χαρακτηριστικό κυμαίνεται από 0 έως 1, όπου το 0 σημαίνει πως το χαρακτηριστικό δεν χρησιμοποιείται καθόλου από το μοντέλο και 1 ότι μπορεί από μόνο του να αξιοποιηθεί για απόλυτα εύστοχες προβλέψεις. Επιπλέον, το άθροισμα των τιμών σημαντικότητας για όλα τα χαρακτηριστικά είναι 1.

Αξίζει να σημειωθεί ότι αν ένα χαρακτηριστικό παρουσιάζει χαμηλή τιμή σημαντικότητας, αυτό δε σημαίνει απαραίτητα ότι παρέχει περιορισμένη πληροφορία για τη μεταβλητή-στόχο, σημαίνει απλά ότι αξιοποιείται από το μοντέλο. Αυτό μπορεί π.χ. να οφείλεται στο γεγονός ότι κάποιο άλλο χαρακτηριστικό παρέχει την ίδια ακριβώς πληροφορία με αυτό. Επίσης, σε αντίθεση με τους συντελεστές βαρύτητας ενός γραμμικού μοντέλου, οι τιμές της σημαντικότητας είναι πάντοτε θετικές. Γενικότερα, η τιμή της σημαντικότητας ενός χαρακτηριστικού δε το συνδέει με κάποιο τρόπο με μία από τις πιθανές τάξεις. Στην πραγματικότητα, επειδή η αρχή λειτουργίας των Δέντρων Αποφάσεων τους επιτρέπει να αναγνωρίζουν μη-γραμμικές συσχετίσεις μεταξύ χαρακτηριστικών και στόχων, το ίδιο χαρακτηριστικό μπορεί για ένα εύρος τιμών να καταδεικνύει τη μία τάξη και για τις υπόλοιπες τιμές την άλλη.

4.3.5 Χαρακτηριστικά των Δέντρων Αποφάσεων

Συνοψίζοντας, διαπιστώνεται ότι τα Δέντρα Αποφάσεων παρουσιάζουν αρκετά σημαντικά πλεονεκτήματα. Συγκεκριμένα, λόγω της αρχής λειτουργίας τους μπορούν να αποκωδικοποιήσουν μη-γραμμικές συσχετίσεις μεταξύ χαρακτηριστικών και στόχων. Επιπλέον, δεν απαιτούν οι τιμές των διάφορων χαρακτηριστικών να έχουν αναχθεί σε κοινή κλίμακα, αφού το κάθε χαρακτηριστικό εξετάζεται μόνο του. Φαινόμενα όπως η υπερπροσαρμογή στα δεδομένα μπορούν να αποφευχθούν επιλέγοντας κατάλληλα την τιμή παραμέτρων όπως το μέγιστο βάθος του δέντρου, ακολουθώντας δηλαδή μια στρατηγική εκ των προτέρων κλαδέματος. Τέλος, τα μοντέλα που παράγονται είναι συχνά εύκολο να εξηγηθούν σε μη σχετικό ακροατήριο.

Στον αντίποδα, το βασικό μειονέκτημα των Δέντρων Αποφάσεων είναι ότι ακόμη και εφαρμόζοντας στρατηγικές κλαδέματος τείνουν να υπερπροσαρμόζονται στα δεδομένα. Απαιτείται, επομένως, ιδιαίτερη προσοχή και σύγκριση της απόδοσης του μοντέλου στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου, ώστε η υπερπροσαρμογή να γίνει αντιληπτή και να αναιρεθεί με τις κατάλληλες ενέργειες.

4.3.6 Τυχαία Δάση (Random Forests)

Τα *Σύνολα Μοντέλων* (*Ensembles*) είναι μέθοδοι που συνδυάζουν πολλαπλά μοντέλα Μηχανικής Μάθησης για να δημιουργήσουν ένα πιο ισχυρό μοντέλο από τα αρχικά. Υπάρχουν διάφορα μοντέλα που ανήκουν σε αυτήν την κατηγορία, ένα από αυτά όμως έχει αποδειχτεί ότι είναι ιδιαίτερα αποτελεσματικό σε ένα μεγάλο εύρος προβλημάτων. Το μοντέλο αυτό έχει σαν δομική μονάδα τα Δέντρα Αποφάσεων και δεν είναι άλλο από τα *Τυχαία Δάση* [2, 6, 15, 16, 18, 19].

Όπως έχει αναφερθεί, βασικό ελάττωμα των Δέντρων Αποφάσεων είναι ότι τείνουν να υπερπροσαρμόζονται στα δεδομένα του συνόλου εκπαίδευσης. Τα *Τυχαία Δάση* (*Random Forests*) επιχειρούν να επιλύσουν αυτό το πρόβλημα. Ουσιαστικά, πρόκειται για συλλογές Δέντρων Αποφάσεων, όπου κάθε δέντρο είναι λίγο διαφορετικό από τα υπόλοιπα της συλλογής. Το σκεπτικό με βάση το οποίο λειτουργούν είναι ότι κάθε δέντρο μπορεί να είναι αποτελεσματικό στις προβλέψεις του, τείνει όμως να υπερπροσαρμόζεται λίγο στα δεδομένα. Κατασκευάζοντας όμως πολλά δέντρα τα οποία είναι αποτελεσματικά και υπερπροσαρμόζονται με διαφορετικό τρόπο το ένα από το άλλο, είναι δυνατό να μειωθεί ο βαθμός υπερπροσαρμογής συνδυάζοντας τις προβλέψεις όλων των δέντρων.

Ο όρος *τυχαία* στο όνομα αυτών των μοντέλων οφείλεται στο γεγονός ότι κατά την

κατασκευή κάθε δέντρου εισάγεται μια μικρή ποσότητα τυχαιότητας που διασφαλίζει ότι τα διάφορα δέντρα είναι διαφορετικά. Υπάρχουν δύο κύριοι τρόποι με τους οποίους τα δέντρα ενός Τυχαίου Δάσους διαφοροποιούνται τυχαία. Ο πρώτος έγκειται στην επιλογή διαφορετικών παρατηρήσεων από το σύνολο εκπαίδευσης για την κατασκευή κάθε δέντρου, ενώ ο δεύτερος στην επιλογή διαφορετικών χαρακτηριστικών για κάθε ένα από τα τεστ που περιλαμβάνουν τα δέντρα.

4.3.7 Κατασκευή Τυχαίων Δασών

Για να κατασκευαστεί ένα Τυχαίο Δάσος, αρχικά επιλέγεται το πλήθος των Δέντρων Αποφάσεων από τα οποία αυτό αποτελείται, π.χ. 10 δέντρα. Κάθε δέντρο από αυτά θα κατασκευαστεί ανεξάρτητα από το υπόλοιπα και ο αλγόριθμος θα κάνει διαφορετικές τυχαίες επιλογές για να διασφαλίσει ότι είναι όντως διακριτά. Αρχικά, για κάθε δέντρο επιλέγεται ένα τυχαίο δείγμα παρατηρήσεων μέσα από το σύνολο εκπαίδευσης, το οποίο ονομάζεται *bootstrap sample*. Για την ακρίβεια, το bootstrap sample περιέχει ίδιο πλήθος παρατηρήσεων με το σύνολο εκπαίδευσης, απλά επειδή η τυχαία επιλογή πραγματοποιείται με αντικατάσταση, ορισμένες από τις αρχικές παρατηρήσεις επαναλαμβάνονται και άλλες - περίπου το 1/3 - λείπουν.

Στη συνέχεια, ένα Δέντρο Αποφάσεων κατασκευάζεται με βάση αυτό το δείγμα, με μια μικρή διαφοροποίηση από τον τρόπο που κατασκευάζεται ένα μεμονωμένο δέντρο. Αντί ο αλγόριθμος να αναζητά το βέλτιστο τεστ για το σύνολο των χαρακτηριστικών, επιλέγει επίσης τυχαία ένα υποσύνολο των χαρακτηριστικών και αναζητά το βέλτιστο τεστ χρησιμοποιώντας ένα χαρακτηριστικό από αυτό το υποσύνολο. Το πλήθος των χαρακτηριστικών του υποσυνόλου αποτελεί παράμετρο του αλγορίθμου. Η τυχαία επιλογή χαρακτηριστικών επαναλαμβάνεται σε κάθε επίπεδο του δέντρου, οπότε το υποσύνολο των χαρακτηριστικών από το οποίο προκύπτουν τα τεστ σε κάθε επίπεδο διαφέρει ακόμη και μεταξύ των επιπέδων του ίδιου δέντρου. Οι δύο μηχανισμοί με τους οποίους εισάγεται η τυχαιότητα εξασφαλίζουν ότι όλα τα μεμονωμένα δέντρα είναι διαφορετικά μεταξύ τους.

Το πλήθος των χαρακτηριστικών με βάση τα οποία επιλέγεται το καλύτερο τεστ αποτελεί πολύ σημαντική παράμετρο του αλγορίθμου. Οι ακραίες επιλογές είναι το συνολικό πλήθος των χαρακτηριστικών ή ένα μόνο χαρακτηριστικό. Η πρώτη επιλογή αναιρεί την τυχαιότητα κατά την κατασκευή του δάσους ως προς τα χαρακτηριστικά, αφήνοντας μόνο την τυχαιότητα που οφείλεται στην τυχαία δειγματοληψία. Η δεύτερη εξαναγκάζει το μοντέλο να χρησιμοποιήσει ένα τεστ βασισμένο στο μοναδικό χαρακτηριστικό που επιλέχθηκε τυχαία, χωρίς να μπορεί να ψάξει σε άλλα χαρακτηριστικά για το βέλτιστο τεστ. Συμπερασματικά, αν το μέγιστο πλήθος είναι μεγάλο, τα διάφορα δέντρα μοιάζουν πολύ μεταξύ τους και προσαρμόζονται εύκολα

στα δεδομένα, χρησιμοποιώντας τα χαρακτηριστικά που δίνουν τη μέγιστη διακριτική ικανότητα. Αν είναι μικρό, τα διάφορα δέντρα είναι πολύ διαφορετικά αλλά μπορεί να απαιτεί να αναπτυχθούν σε μεγάλο βάθος για μην υποπροσαρμοστούν στα δεδομένα.

Για να πραγματοποιήθεί μια πρόβλεψη, κάθε δέντρο του Τυχαίου Δάσους κάνει τη δική του εκτίμηση και οι εκτιμήσεις αυτές συνδυάζονται. Στην περίπτωση της Παλινδρόμησης, λαμβάνεται απλώς ο μέσος όρος των προβλέψεων όλων των δέντρων. Στην Ταξινόμηση, εφαρμόζεται μια λογική *ασθενούς ψηφοφορίας* (*soft voting*). Αυτό σημαίνει πως κάθε δέντρο εκτιμά την πιθανότητα κάθε τάξης, λαμβάνεται ο μέσος όρος των πιθανοτήτων ανά τάξη και επιλέγεται η τάξη με τη μεγαλύτερη πιθανότητα.[1]

4.3.8 Ερμηνεία Τυχαίων Δασών

Όπως είναι ίσως φανερό, τα Τυχαία Δάση δεν μπορούν να οπτικοποιηθούν όπως ένα Δέντρο Αποφάσεων. Επιπλέον, δεν μπορεί να εφαρμοστεί ούτε η λύση ανάγκης του να παρακολουθήσει κανείς τη διαδρομή που συνδέει τους κόμβους με τις περισσότερες παρατηρήσεις για κάθε επίπεδο, γιατί το μοντέλο αποτελείται από πολλά, ίσως εκατοντάδες ή και χιλιάδες δέντρα. Είναι όμως δυνατό να υπολογιστούν οι σημαντικότητες των χαρακτηριστικών, οι οποίες μάλιστα είναι πιο αξιόπιστες σε σχέση με την περίπτωση ενός μόνο δέντρου, αφού έχουν προκύψει ως συνδυασμός των τιμών για τα διάφορα δέντρα του δάσους.

Σε γενικές γραμμές, τα Τυχαία Δάση αποδίδουν λιγότερες τιμές σημαντικότητας κοντά στο μηδέν σε σχέση με ένα Δέντρο Αποφάσεων. Ο λόγος είναι ότι η τυχαιότητα που σχετίζεται με την κατασκευή των διάφορων δέντρων αναγκάζει τον αλγόριθμο να σκεφτεί περισσότερες από μία λογικές, με αποτέλεσμα το δάσος να σχηματίζει μια πολύ ευρύτερη εικόνα της πληροφορίας που περιέχεται στα δεδομένα.

4.3.9 Πλεονεκτήματα και Μειονεκτήματα

Όπως προαναφέρθηκε, τα Τυχαία Δάση συγκαταλέγονται μεταξύ των ευρύτερα χρησιμοποιούμενων μοντέλων Μηχανικής Μάθησης. Αυτό οφείλεται στο ότι είναι πολύ ισχυρά, παρουσιάζουν ικανοποιητική συμπεριφορά χωρίς ιδιαίτερη ρύθμιση των παραμέτρων του αλγορίθμου και δεν απαιτούν αναγωγή των τιμών των χαρακτηριστικών στην ίδια κλίμακα. Επιπλέον, το γεγονός ότι βασίζονται σε Δέντρα Αποφάσεων αφαιρούν την ανάγκη προκαταβολικής επιλογής των κατάλληλων χαρακτηριστικών,

[1] Η άλλη λογική που μπορεί να εφαρμοστεί είναι αυτή της *ισχυρής ψηφοφορίας* (*hard voting*), δηλαδή κάθε δέντρο να κάνει τη δική του πρόβλεψη και η τελική πρόβλεψη να προκύψει πλειοψηφικά.

αφού κατά τη δημιουργία κάθε δέντρου πραγματοποιείται μια αυτόματη επιλογή των καταλληλότερων χαρακτηριστικών.

Στην ουσία, τα Τυχαία Δάση διατηρούν τα πλεονεκτήματα των Δέντρων Αποφάσεων, ενώ ταυτόχρονα αντιδιαστέλλουν και μερικά από τα μειονεκτήματά τους. Ο μόνος ίσως λόγος για να χρησιμοποιήσει κανείς ένα Δέντρο Αποφάσεων είναι αν επιθυμεί να έχει μια συμπαγή και επεξηγήσιμη απεικόνιση της διαδικασίας με την οποία πραγματοποιούνται οι προβλέψεις. Προφανώς, είναι πρακτικά αδύνατο να αναλύσει κανείς σε βάθος εκατοντάδες ή χιλιάδες δέντρα. Αυτό οφείλεται και στο γεγονός ότι τα δέντρα ενός Τυχαίου Δάσους τείνουν να είναι βαθύτερα από ένα Δέντρο Αποφάσεων, λόγω του ότι είναι αναγκασμένα να αξιοποιήσουν υποσύνολα του συνόλου των χαρακτηριστικών.

Ένα άλλο μειονέκτημα των Τυχαίων Δασών που όμως πλέον αντιμετωπίζεται σχετικά εύκολα είναι η απαίτηση σε υπολογιστική ισχύ για την κατασκευή πολλών δέντρων. Φυσικά, η κατασκευή αυτή μπορεί να παραλληλοποιηθεί άμεσα στους σύγχρονους επεξεργαστές, οι οποίοι αποτελούνται από πολλούς πυρήνες, κάθε ένας από τους οποίους μπορεί ταυτόχρονα να κατασκευάζει ένα διαφορετικό δέντρο. Έτσι επιτυγχάνεται γραμμική επιτάχυνση, π.χ. η χρήση δύο πυρήνων έχει ως αποτέλεσμα την κατασκευή ενός Τυχαίου Δάσους στο μισό χρόνο σε σχέση με τη χρήση ενός πυρήνα.

Τέλος, τα Τυχαία Δάση χρησιμοποιούν πολλή μνήμη και απαιτούν περισσότερο χρόνο τόσο για να εκπαιδευτούν, όσο και για να πραγματοποιήσουν τις προβλέψεις τους. Ενώ οι απαιτήσεις κατά την εκπαίδευση μπορούν να αγνοηθούν, αν υποτεθεί ότι τα διαθέσιμα δεδομένα είναι επαρκή για να εκπαιδευτεί απευθείας ένα αποτελεσματικό μοντέλο, οι απαιτήσεις κατά την πρόβλεψη μπορεί να είναι πολύ σημαντικές ανάλογα με την εφαρμογή. Στην παρούσα μελέτη, το αρχικό ζητούμενο είναι ένα σύστημα που θα προβλέπει την παρουσία μανιταριών σε επίπεδο βάρδιας, οπότε δεν υπάρχουν ιδιαίτερες απαιτήσεις από πλευράς ταχύτητας. Αν η λειτουργία του συστήματος απαιτηθεί να είναι σε πραγματικό χρόνο, θα υπάρχει κάποιος (μάλλον μη απαγορευτικός) περιορισμός στο χρόνο που το μοντέλο θα είναι ικανό να δώσει την πρόβλεψή του.

4.4 Αξιολόγηση των Μοντέλων

Η Δυαδική Ταξινόμηση, στην οποία εμπίπτει το πρόβλημα που εξετάζεται, είναι ίσως η πιο συνήθης και μάλλον η πιο απτή στην κατανόηση περίπτωση Μηχανικής Μάθησης. Ακόμη και έτσι, η αξιολόγηση της επιτυχίας της είναι συχνά δύσκολη, με αποτέλεσμα να καθίσταται αναγκαία η χρήση μιας ποικιλίας *δεικτών αξιολόγησης*. Ορισμένοι από αυτούς περιγράφονται στη συνέχεια, αρχίζοντας από τον πιο απλό, την *ευστοχία*.

Στόχος είναι το περιεχόμενο τους να γίνει κατά το δυνατόν πιο κατανοητό ώστε να μπορούν να συγκριθούν και μεταξύ τους να επιλεγεί τελικά ο καταλληλότερος.

Για την περιγραφή που ακολουθεί υπενθυμίζεται ότι στη Δυαδική Ταξινόμηση πραγματοποιείται διάκριση μεταξύ δύο τάξεων, της θετικής και της αρνητικής, όπου η θετική - εδώ η παρουσία μανιταριού - είναι αυτή που έχει σημασία να προσδιοριστεί.

4.4.1 Ευστοχία και Τύποι Σφαλμάτων

Η *ευστοχία* (*accuracy*) ορίζεται ως το πλήθος των ορθών προβλέψεων δια το συνολικό πλήθος των προβλέψεων. Ο δείκτης αυτός συχνά δεν αποτελεί έγκυρο μέτρο της αξιοπιστίας των προβλέψεων επειδή το πλήθος των όρθων προβλέψεων δεν εμπεριέχει από μόνο του όλη τη χρήσιμη πληροφορία. Αυτό ίσως μπορεί να γίνει κατανοητό μέσα από ένα επίκαιρο παράδειγμα: το τεστ διάγνωσης του νέου κορονοϊού SARS-CoV-2. Το αποτέλεσμα του τεστ είναι αρνητικό ή θετικό· στην πρώτη περίπτωση το άτομο θεωρείται υγιές, ενώ στη δεύτερη θεωρείται ότι είναι πιθανό να νοσεί από κορονοϊό. Με την παραδοχή ότι το τεστ κάνει οπωσδήποτε κάποια λάθη, είναι ενδιαφέρον να εξεταστούν οι επιπτώσεις αυτών των λαθών στην πράξη.

Το πρώτο πιθανό λάθος είναι το τεστ ενός υγιούς ατόμου να βγει θετικό, οπότε το άτομο να θεωρηθεί λανθασμένα ότι νοσεί. Τότε το άτομο τίθεται υπό περιορισμό, παρακολουθείται προληπτικά και ίσως υποβάλλεται σε συμπληρωματικές εξετάσεις. Αυτή η κατάσταση συνεπάγεται ορισμένες πρακτικές δυσκολίες, πιθανόν σε συνδυασμό με συναισθηματική φόρτιση και άγχος, αν το άτομο ή κάποιος οικείος του ανήκει σε κάποια ευπαθή ομάδα, καθώς και μια πρόσθετη επιβάρυνση για το σύστημα υγείας. Μια πρόβλεψη σαν την παραπάνω ονομάζεται *ψευδές θετικό* (*false positive, FP*) και στο εξής θα αναφέρεται απλά ως FP για συντομία.

Το δεύτερο πιθανό λάθος είναι το τεστ ενός ατόμου που νοσεί να βγει αρνητικό, οπότε το άτομο να θεωρηθεί λανθασμένα υγιές. Αυτό μπορεί να έχει ως συνέπεια την μετάδοση του ιού από αυτό το άτομο σε άλλα άτομα, καθώς ίσως και τη ραγδαία επιδείνωση της υγείας του ατόμου, ειδικά αν αυτό ανήκει σε κάποια ευπαθή ομάδα. Αυτή η εξέλιξη, εκτός από τις πολύ δυσάρεστες συνέπειες που μπορεί να έχει για το άτομο, μπορεί να επιβαρύνει σημαντικά το σύστημα υγείας, ρισκάροντας μια πιθανή αδυναμία του συστήματος να ανταποκριθεί σε μελλοντικές ανάγκες, αν π.χ. το άτομο εισαχθεί τελικά σε μία από τις περιορισμένες σε πλήθος μονάδες εντατικής θεραπείας. Μια πρόβλεψη σαν την παραπάνω ονομάζεται *ψευδές αρνητικό* (*false negative, FN*)

και στο εξής για συντομία θα αναφέρεται απλά ως FN. [2]

Με βάση το παραπάνω παράδειγμα είναι μάλλον εμφανές ότι οι συνέπειες των δύο τύπων σφαλμάτων μπορεί να διαφέρουν σημαντικά σε σοβαρότητα. Για την ακρίβεια, αυτό είναι περισσότερο ο κανόνας παρά η εξαίρεση στα προβλήματα Δυαδικής Ταξινόμησης. Δημιουργείται λοιπόν η ανάγκη να αποτιμηθούν οι συνέπειες κάθε τύπου σφάλματος και να δοθεί συνειδητά προτεραιότητα στο να αποφευχθεί το ένα έναντι του άλλου, σε μικρότερο ή μεγαλύτερο βαθμό.

Φυσικά, οι προβλέψεις ενός τεστ ή ενός μοντέλου Μηχανικής Μάθησης δεν είναι πάντα λανθασμένες. Εκτός από FP ή FN, το αποτέλεσμα της πρόβλεψης μπορεί να είναι ένα *αληθές θετικό* (*true positive, TP*) ή *αληθές αρνητικό* (*true negative, TN*). Χρησιμοποιώντας όλους τους όρους που εισήχθησαν μπορεί πλέον να δοθεί η σχέση για τον υπολογισμό της ευστοχίας,

$$\text{ευστοχία} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.6}$$

Είναι ίσως φανερό ότι, εκτός του ότι αδυνατεί να διακρίνει μεταξύ των δύο τύπων λαθών, η ευστοχία δεν είναι ικανή να χρησιμοποιηθεί στην περίπτωση δεδομένων με Ανισοκατανεμημένες Τάξεις, όπως είναι σε γενικές γραμμές και τα δεδομένα που αξιοποιούνται σε αυτή τη μελέτη. Αυτό συμβαίνει επειδή ένα ακατάλληλο μοντέλο μπορεί να πετύχει παραπληνητικά υψηλή ευστοχία σε ένα τέτοιο πρόβλημα, απλά προβλέποντας πάντα την συχνότερη τάξη, εδώ την απουσία μανιταριού. Οι παραπάνω σημαντικές αδυναμίες της ευστοχίας την καθιστούν ακατάλληλη για τη μελέτη και οδηγούν στην αναζήτηση εναλλακτικών δεικτών αξιολόγησης.

4.4.2 Πίνακας Σύγχυσης, Ακρίβεια και Ανάκτηση

Ένας εποπτικός τρόπος έκφρασης των αποτελεσμάτων της Ταξινόμησης με στόχο την αξιολόγησή τους είναι ο *Πίνακας Σύγχυσης* (*Confusion Matrix*). Αυτός είναι απλά ένας $\text{n} \times \text{n}$ πίνακας $\mathbf{C}$, όπου n το πλήθος των διαφορετικών τάξεων, του οποίου το στοιχείο C_{ij} εκφράζει το πλήθος των παρατηρήσεων που ανήκουν στην i τάξη και προβλέφθηκε ότι ανήκουν στην j τάξη. Είναι ίσως εμφανές ότι τα στοιχεία της διαγωνίου του πίνακα αποτελούν ορθές προβλέψεις, ενώ τα υπόλοιπα λανθασμένες. Στην περίπτωση της Δυαδικής Ταξινόμησης είναι $\text{n} = 2$ και τα στοιχεία του Πίνακα Σύγχυσης αντιστοιχίζονται στα FP, FN, TP και TN όπως φαίνεται στο Σχ. 4.3.

[2] Στο πεδίο της Στατιστικής, ένα FP αναφέρεται ως *σφάλμα τύπου I* (*type I error*) και ένα FN ως *σφάλμα τύπου II* (*type II error*). Εδώ χρησιμοποιούνται αποκλειστικά οι όροι FP και FN τόσο ως πιο συνήθεις στο πεδίο της Μηχανικής Μάθησης, όσο και ως πολύ πιο περιγραφικοί του πραγματικού περιεχομένου των αντίστοιχων εννοιών.

Παρότι ο Πίνακας Σύγχυσης εκθέτει το βαθμό στον οποίο το μοντέλο υποπίπτει σε κάθε τύπο σφάλματος, η ανάγνωση και κατανόηση ενός πίνακα, καθώς και η σύγκριση μεταξύ πινάκων που αντιστοιχούν σε διαφορετικά μοντέλα, είναι μια διαδικασία ποιοτική, χρονοβόρα και μη αυτοματοποιήσιμη. Για το λόγο αυτό, αντί του πίνακα αξιοποιούνται δείκτες που συνοψίζουν μέρος της πληροφορίας που αυτός περιέχει σε έναν μόνο αριθμό. Δύο τέτοιοι δείκτες, εκτός από την ευστοχία που έχει ήδη αναφερθεί, είναι η *ακρίβεια* και η *ανάκτηση*.

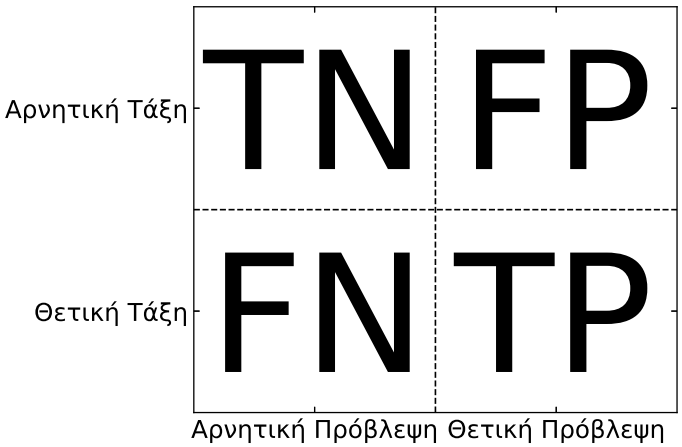
Η *ακρίβεια* (precision) ορίζεται ως το ποσοστό των παρατηρήσεων που προβλέφθηκε ότι ανήκουν στη θετική τάξη και όντως ανήκουν σε αυτή, σύμφωνα με τη σχέση

$$\text{ακρίβεια} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.7}$$

Επιλέγεται ως δείκτης όταν ο στόχος είναι η ελαχιστοποίηση των FP αφού, όπως φαίνεται από τη σχέση, όταν αυτά τείνουν να μηδενιστούν η ακρίβεια τείνει στη μονάδα. Στην παρούσα μελέτη ναι μεν ζητείται να μην προβλέπεται πολλές φορές λανθασμένα η παρουσία μανιταριού, προτεραιότητα όμως δίνεται στο να εντοπιστούν όσο το δυνατόν περισσότερα μανιτάρια, ακόμα και με κόστος μερικά παραπάνω FP, αφού οι έλεγχοι για την επιβεβαίωση της παρουσίας μανιταριού δεν είναι απαγορευτικά δύσκολοι. Σε αυτή την κατεύθυνση μπορεί να φανεί πιο χρήσιμη η *ανάκτηση*.

Η *ανάκτηση* (recall) ορίζεται ώς το ποσοστό των παρατηρήσεων που ανήκουν στη θετική τάξη και όντως προβλέφθηκε ότι ανήκουν σε αυτή, δηλαδή

$$\text{ανάκτηση} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.8}$$



Σχήμα 4.3: Πίνακας Σύγχυσης στην περίπτωση της Δυαδικής Ταξινόμησης.

Όπως αναφέρθηκε και προηγουμένως, αυτός ο δείκτης επιλέγεται όταν στόχος είναι η ελαχιστοποίηση των FN αφού, όπως φαίνεται από τη σχέση, όταν αυτά τείνουν να μηδενιστούν η ανάκτηση τείνει στη μονάδα. Ένα χαρακτηριστικό παράδειγμα μιας τέτοιας περίπτωσης είναι αυτό που αναφέρθηκε παραπάνω για το τεστ διάγνωσης του κορονοϊού ή άλλα αντίστοιχα τεστ: το να εντοπιστούν όλοι οι ασθενείς είναι πιο σημαντικό από το να μην θεωρηθεί λανθασμένα ότι κάποιος νοσεί.

Είναι εμφανές ότι όταν η ακρίβεια αυξάνεται, η ανάκτηση μειώνεται και αντίστροφα. Από τη μία, η απόλυτη ανάκτηση μπορεί να επιτευχθεί αν όλες οι παρατηρήσεις αποδοθούν στην θετική τάξη. Έτσι δε θα υπάρχουν ψευδή αρνητικά, αλλά και καθόλου αρνητικά γενικότερα. Αντίθετα, σχεδόν απόλυτη ακρίβεια μπορεί να επιτευχθεί αν στη θετική τάξη αποδοθούν οι ελάχιστες παρατηρήσεις για τις οποίες το μοντέλο είναι απολύτως σίγουρο ότι ανήκουν σε αυτήν. Σε αυτήν την περίπτωση όμως η ανάκτηση θα είναι πάρα πολύ χαμηλή.

Από τα παραπάνω φαίνεται πως κανένα από τα δύο μεγέθη δεν είναι ικανό να δώσει μια σφαιρική εικόνα. Αυτό μπορεί σε ένα βαθμό να το πετύχει το *f-σκορ* (*f-score*), η πιο κοινή εκδοχή του οποίου ονομάζεται f_1-score και δίνεται από τη σχέση

$$f_1 = 2 \cdot \frac{\text{ακρίβεια} \cdot \text{ανάκτηση}}{\text{ακρίβεια} + \text{ανάκτηση}} \tag{4.9}$$

Από τη σχέση φαίνεται ότι το f_1-score αποτελεί τον αρμονικό μέσο της ακρίβειας και της ανάκτησης. Επειδή λαμβάνει υπόψη και τους δύο αυτούς δείκτες, αποτελεί καταλληλότερο δείκτη σε σχέση με την ευστοχία για ειδικές περιπτώσεις, όπως αυτή των Ανισοκατανεμημένων Τάξεων. Μάλιστα, είναι δυνατό να δοθεί προτεραιότητα στο ένα από τα δύο μεγέθη, χρησιμοποιώντας τη γενική εκδοχή του δείκτη, που ονομάζεται f_β-score και δίνεται από τη σχέση

$$f_\beta = (1 + \beta^2) \cdot \frac{\text{ακρίβεια} \cdot \text{ανάκτηση}}{\beta^2 \cdot \text{ακρίβεια} + \text{ανάκτηση}} \tag{4.10}$$

Για παράδειγμα, για $\beta = 2$ και $\beta = 0.5$ προκύπτουν οι δείκτες

$$f_2 = 5 \cdot \frac{\text{ακρίβεια} \cdot \text{ανάκτηση}}{4 \cdot \text{ακρίβεια} + \text{ανάκτηση}} \tag{4.11}$$

και

$$f_{0.5} = 1.25 \cdot \frac{\text{ακρίβεια} \cdot \text{ανάκτηση}}{0.25 \cdot \text{ακρίβεια} + \text{ανάκτηση}} \tag{4.12}$$

που δίνουν προτεραιότητα στην ανάκτηση και την ακρίβεια, αντίστοιχα.

Οι δείκτες που έχουν ήδη περιγραφεί (ευστοχία, ακρίβεια και ανάκτηση) σε συνδυασμό με τον Πίνακα Σύγχυσης είναι εργαλεία ικανά να παρέχουν μια λεπτομερή

και σφαιρική εικόνα που επιτρέπει την αξιολόγηση ενός συνόλου προβλέψεων. Οι ίδιες οι προβλέψεις όμως έχουν προκύψει χωρίς να ληφθεί υπόψη σημαντικός όγκος πληροφορίας που περιλαμβάνεται στο μοντέλο. Πολλοί ταξινομητές, στους οποίους περιλαμβάνονται τα Decision Trees και τα παράγωγά τους, πρώτα υπολογίζουν την πιθανότητα μία παρατήρηση να ανήκει σε κάθε μία από τις τάξεις και στη συνέχεια αποδίδουν την παρατήρηση στην τάξη με την υψηλότερη πιθανότητα.

4.4.3 Καμπύλη Ακρίβειας-Ανάκτησης

Προφανώς, στην περίπτωση της Δυαδικής Ταξινόμησης, για να αποδοθεί μια παρατήρηση στη θετική τάξη αρκεί η αντίστοιχη πιθανότητα να υπερβαίνει το 0.5. Η τιμή αυτή ονομάζεται *όριο απόφασης* (*decision threshold*). Μεταβάλλοντας την τιμή του ορίου απόφασης είναι δυνατό να προκύψουν διαφορετικές προβλέψεις, άρα και διαφορετικές τιμές των δεικτών, π.χ. της ακρίβειας και της ανάκτησης, οι οποίες είναι και αντιστρόφως ανάλογες. Για παράδειγμα, αν το όριο είναι ίσο με 0.4, περισσότερες παρατηρήσεις θα αποδοθούν στην θετική τάξη, με αποτέλεσμα την αύξηση της ανάκτησης σε βάρος της ακρίβειας. Το μοντέλο δηλαδή αποδίδει την παρατήρηση στη θετική τάξη ακόμα και αν δεν είναι τόσο βέβαιο για την εκτίμησή του.

Το όριο απόφασης εξαρτάται από το ποιο είναι το ζητούμενο στην εκάστοτε εφαρμογή. Στην παρούσα μελέτη, η ανάκτηση είναι σημαντικότερη από την ακρίβεια. Ακόμη και έτσι όμως, το ακριβές όριο, το οποίο ονομάζεται και *σημείο λειτουργίας*, δεν είναι γνωστό προκαταβολικά, ενώ μπορεί στην πορεία να αλλάξει. Παρόλα αυτά, τα διάφορα μοντέλα που παράγονται πρέπει σε πρώτη φάση να συγκριθούν και μεταξύ τους να επιλεγεί αυτό που γενικά παρουσιάζει την καλύτερη συμπεριφορά. Αυτό είναι δυνατό αν υπολογιστούν τα ζεύγη ακρίβειας και ανάκτησης για τις διάφορες τιμές που μπορεί να λάβει το όριο απόφασης. Έτσι προκύπτει ένας ακόμη τρόπος αξιολόγησης ενός μοντέλου, η *καμπύλη ακρίβειας-ανάκτησης*.

Η *καμπύλη ακρίβειας-ανάκτησης* (*precision-recall curve*) αποτελείται, όπως υποδεικνύει και το όνομα της, από σημεία με συντεταγμένες τα ζεύγη τιμών ακρίβειας και ανάκτησης που προκύπτουν για διαφορετικές τιμές του ορίου απόφασης. Η καμπύλη αυτή απεικονίζεται σε ένα διάγραμμα οι άξονες του οποίου έχουν ως όρια τις τιμές 0 και 1, αφού τόσο η ακρίβεια, όσο και η ανάκτηση κυμαίνονται μεταξύ αυτών των τιμών. Προφανώς, το επιθυμητό για ένα μοντέλο είναι να συνδυάζει υψηλή ακρίβεια και ανάκτηση, δηλαδή να παράγει μια καμπύλη ακρίβειας-ανάκτησης με σημεία στην πάνω δεξία γωνία ενός τέτοιου διαγράμματος. Με αυτό το σκεπτικό μπορούν να συγκριθούν οπτικά καμπύλες που αντιστοιχούν σε διαφορετικά μοντέλα, ώστε να επιλεγεί το καλύτερο από αυτά. Αυτός ο τρόπος σύγκρισης παρουσιάζει τις ίδιες αδυναμίες με τον Πίνακα Σύγχυσης, οπότε θα ήταν ιδιαίτερα χρήσιμο να μπορούσε να περιγραφεί

από έναν μόνο αριθμό.

Ένας τρόπος να συνοψίσει κανείς την πληροφορία της καμπύλης ακρίβειας-ανάκτησης για ευκολότερη και ενδεχομένως αυτόματη σύγκριση μεταξύ μοντέλων είναι υπολογίζοντας το εμβαδό κάτω από την καμπύλη, το οποίο ορισμένες φορές αναφέρεται και ως *μέση ακρίβεια* (*average precision*). Επειδή η μέση ακρίβεια είναι το εμβαδόν μιας καμπύλης που πηγαίνει από το 1 στο 0, κυμαίνεται πάντα μεταξύ 0 και 1. Με βάση τα όσα έχουν ειπωθεί για την παρούσα μελέτη, το μέγεθος αυτό θα μπορούσε να είναι ικανοποιητικό για την αξιολόγηση και τη σύγκριση των διάφορων μοντέλων. Υπάρχει όμως ένα παρόμοιο μέγεθος το οποίο ίσως είναι ακόμη καταλληλότερο.

4.4.4 Χαρακτηριστική Καμπύλη Λειτουργίας Δέκτη

Ένα ακόμη εργαλείο που αξιοποιείται για την ανάλυση της συμπεριφοράς ταξινομητών σε διαφορετικά όρια απόφασης είναι η *χαρακτηριστική καμπύλη λειτουργίας δέκτη* (*receiver operating characteristic, ROC*). Αντίστοιχα με την καμπύλη ακρίβειας-ανάκτησης, η ROC αποτελείται από σημεία που προκύπτουν για διαφορετικές τιμές του ορίου απόφασης, με συντεταγμένες όμως τα ζεύγη τιμών του *βαθμού ψευδών θετικών* και του *βαθμού αληθών θετικών*, αντί των ζευγών ακρίβειας-ανάκτησης. Ο *βαθμός αληθών θετικών* (*true positive rate, TPR*) δεν είναι τίποτα άλλο παρά μια εναλλακτική ονομασία για την ανάκτηση, υπολογίζεται δηλαδή ως εξής:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.13}$$

Συνεπώς, η διαφορά της ROC από την καμπύλη ακρίβειας-ανάκτησης έγκειται στη δεύτερη συντεταγμένη, το *βαθμό ψευδών θετικών* (*false positive rate, FPR*), που ορίζεται ως το ποσοστό των παρατηρήσεων που ανήκουν στην αρνητική τάξη αλλά προβλέφθηκε ότι ανήκουν στη θετική, δηλαδή

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4.14}$$

Στην περίπτωση της ROC, η ιδανική θέση της καμπύλης είναι πάνω αριστερά· το ζητούμενο είναι ένα μοντέλο με υψηλή ανάκτηση/TPR και χαμηλό FPR. Μια ενδεικτική καμπύλη, η οποία μάλιστα προέρχεται από τα αποτελέσματα της παρούσας μελέτης παρουσιάζεται στο Σχ. 5.1. Όπως και η καμπύλη ακρίβειας-ανάκτησης όμως, η ROC πηγαίνει από το 0 στο 1, οπότε το εμβαδόν της κυμαίνεται μεταξύ 0 και 1. Λόγω της φύσης των TPR και FPR, ένα μοντέλο που πραγματοποιεί εντελώς τυχαίες προβλέψεις παράγει μια ROC που ταυτίζεται με την ευθεία $y = x$, ανεξάρτητα από

την κατανομή των παρατηρήσεων στις δύο τάξεις, ακόμη δηλαδή και αν οι τάξεις είναι ανισοκατανεμημένες. Αυτό καθιστά τόσο την καμπύλη, όσο και το εμβαδόν κάτω από αυτήν, το οποίο είναι γνωστό ως *area under the curve (AUC)* καταλληλότερους δείκτες αξιολόγησης για προβλήματα με Ανισοκατανεμημένες Τάξεις. Αυτός είναι και ο λόγος που το AUC είναι ο δείκτης με βάση τον οποίο αξιολογούνται και συγκρίνονται όλα τα μοντέλα σε αυτή τη μελέτη.

Τέλος, υπενθυμίζεται ότι τόσο το AUC όσο και το εμβαδόν κάτω από την καμπύλη ακρίβειας-ανάκτησης υποδεικνύουν το μοντέλο με την γενικά καλύτερη συμπεριφορά. Συνεπώς, η επιλογή του κατάλληλου σημείου λειτουργίας, δηλαδή του ορίου απόφασης που δίνει τα επιθυμητά αποτελέσματα για την παρούσα εφαρμογή, εξακολουθεί να είναι απαραίτητη μετά την επιλογή του καταλληλότερου μοντέλου.

4.5 Ρύθμιση Παραμέτρων

Έχοντας ήδη αναφερθεί σε όλα τα επιμέρους στοιχεία της μεθοδολογίας Μηχανικής Μάθησης που ακολουθείται, αυτό που μένει είναι η περιγραφή της σύνθεσής τους σε μία ενιαία ροή. Αυτό δεν είναι αρκετό, καθώς για την εύρεση του μοντέλου με τη βέλτιστη γενίκευση απαιτείται η ορθή επιλογή των κρίσιμων παραμέτρων των αλγορίθμων που υπεισέρχονται στη ροή αυτή, συγκεκριμένα των Δέντρων Αποφάσεων και των Τυχαίων Δασών. Η *Ρύθμιση Παραμέτρων* (*Parameter Tuning*) αποτελεί μια απαιτητική διαδικασία, απαραίτητη για όλα σχεδόν τα μοντέλα.

Η ρύθμιση αυτή θα μπορούσε να πραγματοποιηθεί δοκιμάζοντας διαφορετικές παραμέτρους χρησιμοποιώντας απλά τα σύνολα εκπαίδευσης και ελέγχου. Η προσέγγιση αυτή ονομάζεται *Αναζήτηση Πλέγματος* (*Grid Search*) και έγκειται στην δοκιμή όλων των πιθανών συνδυασμών των κρίσιμων παραμέτρων και της σύγκρισης των αποτελεσμάτων για την εύρεση του καλύτερου δυνατού συνδυασμού. Ο όρος «πλέγμα» στην ονομασία αναφέρεται σε ένα πλέγμα στο ν-διάστατο χώρο των παραμέτρων, οι κορυφές του οποίου αποτελούν τους παραπάνω πιθανούς συνδυασμούς.

Για μια καλύτερη, όμως, εκτίμηση της ικανότητας των μοντέλων να γενικεύουν είναι προτιμότερο να χρησιμοποιηθεί η τεχνική του cross-validation. Αυτο σημαίνει πως κάθε συνδυασμός δε θα εκπαιδευτεί από ένα μόνο σύνολο εκπαίδευσης και θα ελεγχθεί αποκλειστικά σε ένα σύνολο ελέγχου, αλλά αντίθετα θα ακολουθηθεί η επαναληπτική διαδικασία που περιγράφεται στην Εν. 4.2. Επειδή η ρύθμιση παραμέτρων με αυτόν τον τρόπο είναι μια πολύ κοινή διαδικασία, οι βιβλιοθήκες που αφορούν στη Μηχανική Μάθηση περιλαμβάνουν έτοιμες σχετικές μεθόδους, με αυτή που περιλαμβάνεται στο scikit-learn να είναι η `GridSearchCV`. Το όνομα της μεθόδου είναι

πλήρως περιγραφικό, Αναζήτηση Πλέγματος με cross-validation.

Πριν περιγραφεί ο τρόπος χρήσης της μεθόδου, υπενθυμίζεται ότι τα δεδομένα που τροφοδοτούνται στη μέθοδο δε μπορούν να χρησιμοιηθούν για την τελική αξιολόγηση του μοντέλου. Από τη στιγμή που έχουν αξιοποιηθεί για την επιλογή των παραμέτρων, θεωρείται ότι πληροφορία έχει «διαρρεύσει» από αυτά στον αλγόριθμο, οπότε εδώ απαιτείται η χρήση των συνόλων `X_valid` και `y_valid` που παράγονται από τη μέθοδο `train_test_split`. Με βάση αυτά, που μαζί αποτελούν το σύνολο διακρίβωσης, χρησιμοποιώντας τον επιλεγμένο δείκτη αξιολόγησης, εκτιμάται έγκυρα η ικανότητα του μοντέλου να προβλέψει με βάση άγνωστα δεδομένα. Όταν αυτή η αξιολόγηση ολοκληρωθεί, το μοντέλο εκπαιδεύεται χρησιμοποιώντας όλα τα δεδομένα και είναι έτοιμο για χρήση.

Πέρα από την εκτίμηση της γενίκευσης του μοντέλου, στην παρούσα μελέτη το βήμα αυτό αξιοποιείται και για τη σύγκριση μεταξύ μοντέλων που έχουν προκύψει από τα διαφορετικά σύνολα ασφαλών και μη ασφαλών δεδομένων. Αυτό γίνεται πιο κατανοητό στο Κεφ. 5 όπου παρουσιάζονται τα αποτελέσματα των υπολογισμών και πραγματοποιείται αυτή ακριβώς η σύγκριση.

Τέλος, περιγράφονται οι λεπτομέρειες της χρήσης της μεθόδου `GridSearchCV`. Στη μέθοδο παρέχεται η ακολουθία των εργαλείων Μηχανικής Μάθησης που ζητείται να ελεγχθεί. Υπό μορφή λίστας παρέχονται οι συνδυασμοί των παραμέτρων που ζητείται να διερευνήσει η Αναζήτηση Πλέγματος. Επίσης υπό μορφή λίστας παρέχονται και οι Δείκτες Αξιολόγησης που ζητείται να χρησιμοποιηθούν. Πέραν αυτών, οι σημαντικές παράμετροι αυτής της μεθόδου είναι οι εξής:

- `cv`: Το πλήθος των folds του cross-validation. Επειδή το πλήθος αυτό είναι ανάλογο του πλήθους των υπολογισμών που θα πραγματοποιηθούν, επιλέγεται μία μικρή αλλά επαρκής τιμή, `cv` $= 5$.
- `refit`: Η παράμετρος αυτή καθορίζει αν μετά την ολοκλήρωση της διερεύνησης όλα τα δεδομένα που χρησιμοποιούνται για τη μοντελοποίηση θα αξιοποιηθούν για μια πλήρη εκπαίδευση του βέλτιστου μοντέλου. Σε περίπτωση που αυτό είναι επιθυμητό, λαμβάνει την τιμή `True` αν χρησιμοποιείται μόνο ένας Δείκτης Αξιολόγησης ή το όνομα του δείκτη με βάση τον οποίο θεωρείται το ποιο είναι το βέλτιστο μοντέλο. Αν δεν είναι επιθυμητό προφανώς λαμβάνει την τιμή `False`. Στη συγκεκριμένη περίπτωση, επειδή εξετάζονται περισσότεροι του ενός δείκτες λαμβάνει την τιμή `'roc_auc'`, που είναι ο δείκτης με βάση τον οποίο αξιολογούνται τελικά τα μοντέλα (βλ. 4.4).
- `return_train_scores`: Καθορίζει το κάτα πόσο στον πίνακα των αποτελεσμάτων περιλαμβάνονται μόνο οι επιδόσεις των μοντέλων στα σύνολα ελέγχου ή και οι επιδόσεις στα σύνολα εκπαίδευσης. Όλα αυτά τα στοιχεία χρησιμο-

ποιούνται σε ορισμένες καμπύλες του Κεφ. 5 που αφορά στα Αποτελέσματα, για την εκτίμηση της εξάρτησης της γενίκευσης από την περιπλοκότητα των μοντέλων. Για το λόγο αυτό η παράμετρος ορίζεται ως `True`.

- `n_jobs`: Το πλήθος των υπολογισμών που εκτελούνται παράλληλα, ο καθένας σε διαφορετικό επεξεργαστή του Η/Υ που χρησιμοποιείται. Κάθε υπολογισμός αντιστοιχεί σε ένα μοντέλο προς δοκιμή. Επειδή ο υπολογιστής που χρησιμοποιείται για τη μελέτη είναι 8-πύρηνος, κατά περίπτωση η παράμετρος αυτή ορίζεται ως 4 ή 6. Η ύπαρξη αυτής της δυνατότητας αποτελεί και ένα παράδειγμα παραλληλοποίησης που παρέχεται με απλό τρόπο από την *Python* και συγκεκριμένα τη βιβλιοθήκη *scikit-learn*.

Κεφάλαιο 5

Υπολογισμοί και Αποτελέσματα

5.1 Ασφαλή Δεδομένα

5.1.1 Δέντρα Αποφάσεων

Η πρώτη οικογένεια μοντέλων που αναπτύχθηκε περιέχει μοντέλα βασισμένα σε Δέντρα Αποφάσεων. Η εφαρμογή του σχετικού αλγορίθμου που αξιοποιείται είναι ο `DecisionTreeClassifier` της βιβλιοθήκης scikit-learn. Εξετάζονται τα μοντέλα που προκύπτουν μεταβάλλοντας δύο παραμέτρους του αλγορίθμου, που είναι οι εξής:

- `max_depth`: Το μέγιστο βάθος του δέντρου, δηλαδή το μέγιστο πλήθος των επιπέδων του. Οι διαθέσιμες επιλογές είναι κάποιος θετικός ακέραιος ή `None`, τιμή της παραμέτρου για την οποιά το δέντρο αναπτύσσεται έως ότου όλα τα φύλλα να είναι καθαρά ή να περιέχουν λιγότερες παρατηρήσεις από όσες ορίζει η παράμετρος του αλγορίθμου `min_samples_split` (βλ. παρακάτω).
- `class_weight`: Τα βάρη που αποδίδονται σε κάθε μία από τις δύο τάξεις, False (απουσία) και True (παρουσία μανιταριού). Τα βάρη είναι δυνατό να οριστούν επακριβώς από το χρήστη, π.χ. ορίζοντας την παράμετρο {False: 1, True: 50}, τιμή για την οποία οι παρατηρήσεις της τάξης True έχουν 50 φορές μεγαλύτερη βαρύτητα όταν καθορίζουν το σε ποια τάξη αντιστοιχεί ένα φύλλο. Μια τόσο συγκεκριμένη επιλογή των βαρών μπορεί να γίνει αν είναι γνωστή με βεβαιότητα η συχνότητα των δύο τάξεων στην πραγματικότητα. Στην παρούσα μελέτη, όπου δε συμβαίνει κάτι τέτοιο, είναι προτιμότερο να χρησιμοποιηθούν οι άλλες δύο διαθέσιμες τιμές της παραμέτρου, `None` και `balanced`. Η μεν πρώτη αποδίδει την ίδια βαρύτητα σε όλες τις παρατηρήσεις, ενώ η δεύτερη αποδίδει στις παρατηρήσεις κάθε τάξης βαρύτητα αντιστρόφως ανάλογη της συχνότητας με

την οποία η τάξη αυτή απαντάται **στα διαθέσιμα δεδομένα**.

Είναι πολύ σημαντικό να σημειωθεί ότι, λόγω της ιδιαιτερότητας περί αβεβαιότητας για τις ετικέτες, η οποία επιτάσσει τη δοκιμή διαφορετικών συνόλων δεδομένων, οι συχνότητες των δύο τάξεων είναι διαφορετικές για τα διάφορα σύνολα. Άρα, η επιλογή `class_weight = balanced` αποδίδει στις δύο τάξεις βαρύτητες συναρτήσει της συχνότητας των δύο τάξεων στο εκάστοτε σύνολο δεδομένων.

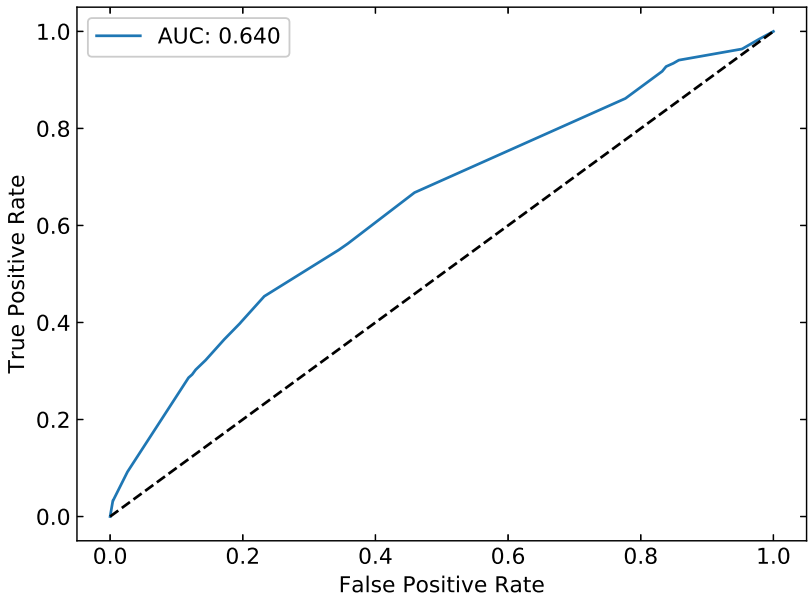
Οι τιμές των δύο παραμέτρων που τελικά εξετάζονται είναι

- `max_depth`: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]
- `class_weight`: [`None, balanced`]

Στη συνέχεια παρουσιάζονται τα αποτελέσματα που αφορούν την οικογένεια των *ασφαλών* δεδομένων. Υπενθυμίζεται πως στα δεδομένα αυτά έχουν αφαιρεθεί παρατηρήσεις που αφορούν 15 μέρες πριν και 1 έως 9 μέρες μετά τον εντοπισμό του μανιταριού, αφού για αυτές που προηγούνται είναι αβέβαιο το αν υπήρχε ή όχι μανιτάρι, ενώ για εκείνες που ακολουθούν είναι άγνωστο το αν η λεκάνη λειτουργούσε υπό φυσιολογικές συνθήκες. Πρώτα παρουσιάζονται ενδεικτικά διαγράμματα, τα οποία συγκεκριμένα προέρχονται από τη διερεύνηση που έγινε για τα δεδομένα S3, δηλαδή αυτά στα οποία για κάθε λεκάνη έχουν αγνοηθεί τα δεδομένα 15 ημερών πριν τον εντοπισμό του μανιταριού και 3 ημερών μετά, εξού και S3.

Στο Σχ. 5.1 απεικονίζεται η καμπύλη ROC που αντιστοιχεί στην εφαρμογή του καλύτερου μοντέλου στα δεδομένα του συνόλου διακρίβωσης. Υπενθυμίζεται ότι το καλύτερο μοντέλο προκύπτει από μία διαδικασία Αναζήτησης Πλέγματος. Επειδή μια τέτοια καμπύλη παρουσιάζεται για πρώτη φορά, αυτή περιγράφεται αναλυτικά ώστε να γίνει πλήρως κατανοητή. Ο οριζόντιος άξονας αφορά στο False Positive Rate, δηλαδή το ποσοστό των παρατηρήσεων που ανήκουν στην αρνητική τάξη αλλά προβλέφθηκε ότι ανήκουν στη θετική, ενώ ο κάθετος στο True Positive Rate, δηλαδή το ποσοστό των παρατηρήσεων που προβλέφθηκε ότι ανήκουν στη θετική τάξη από όλες όσες όντως ανήκουν σε αυτή. Η ευθεία $y = x$, με μαύρη διακεκομμένη γραμμή, αποτελεί την καμπύλη ROC που προκύπτει για εντελώς τυχαίες προβλέψεις και χρησιμοποιείται ως μέτρο σύγκρισης. Τέλος, στο σχήμα σημειώνεται και το AUC, δηλαδή το εμβαδόν κάτω από την καμπύλη ROC του σχήματος.

Όσον αφορά την ίδια την καμπύλη ROC, με κουκίδες σημειώνονται τα σημεία που αντιστοιχούν στις διάφορες τιμές του ορίου απόφασης για τις οποίες υπολογίστηκαν τα ζεύγη FPR-TPR. Τα σημεία ενώνονται αυθαίρετα με τεθλασμένη γραμμή, το εμβαδόν κάτω από την οποία είναι το AUC. Αυτό ζητείται να είναι κατά το δυνατόν μεγαλύτερο από 0.5, που αντιστοιχεί σε τυχαίες προβλέψεις. Η τιμή 0.64 αυτής της καμπύλης δεν είναι ιδιαίτερα ικανοποιητική, απέχει όμως από το να είναι τυχαία οπότε αποτελεί ένα βήμα στη σωστή κατεύθυνση. Η καμπύλη ROC ζητείται να έχει μεγάλη



Σχήμα 5.1: Καμπύλη ROC του βέλτιστου Δέντρου Αποφάσεων που προέκυψε για τα δεδομένα S3, εφαρμοσμένου στο σύνολο διακρίβωσης. Τιμές κρίσιμων παραμέτρων: `max_depth` = 5, `class_weight = balanced`.

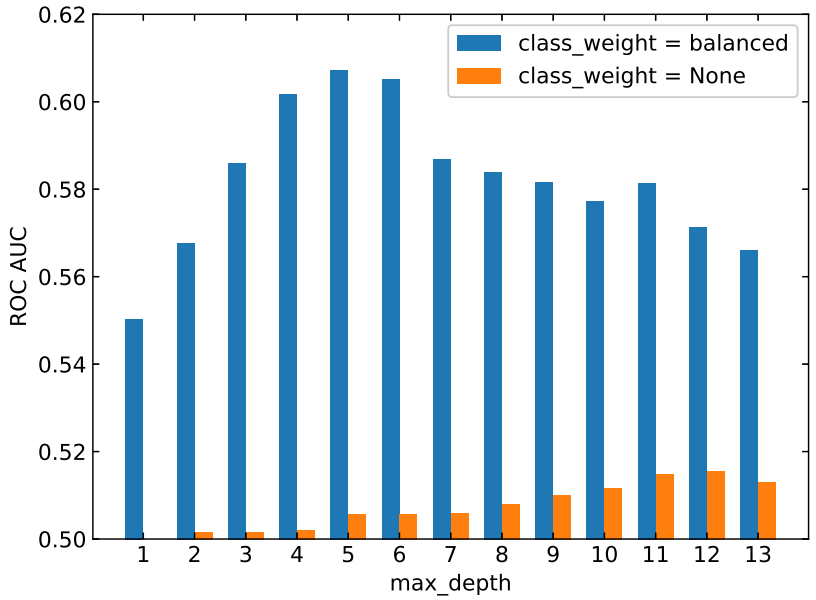
κλίση σε μικρές τιμές FPR και να συγκλίνει γρήγορα σε μεγάλες τιμές TPR, δηλαδή τα σημεία της να περιγράφουν χαμηλό FPR σε συνδυασμό με υψηλό TPR.

Στο Σχ. 5.2 απεικονίζεται η εξάρτηση του AUC των μοντέλων που εξετάστηκαν κατά την Αναζήτηση Πλέγματος από τις τιμές των δύο παραμέτρων του αλγορίθμου, `max_depth` και `class_weight`. Οι τιμές της πρώτης παραμέτρου αντιστοιχούν στις διάφορες θέσεις στον άξονα x, ενώ της δεύτερης που είναι μόλις δύο απεικονίζονται με διαφορετικό χρώμα, όπως φαίνεται στο υπόμνημα. Καταρχάς, είναι εμφανές ότι το μοντέλο με την υψηλότερη επίδοση, πάντα όσον αφορά το δείκτη αξιολόγησης που χρησιμοποιείται, είναι αυτό με `max_depth` = 5 και `class_weight = balanced`, η καμπύλη ROC του οποίου παρουσιάστηκε στο Σχ. 5.1. Πολύ κοντά του θα μπορούσε να θεωρηθεί ότι βρίσκονται και τα μοντέλα με `max_depth` = 4 ή 6. Αν έπρεπε να επιλεγεί ένα από τα τρία αυτά μοντέλα, και η διαφορά στην επίδοση τους θεωρούνταν αμελητέα, θα μπορούσε ίσως να επιλεγεί αυτό για το οποίο `max_depth` = 4, με το σκεπτικό ότι είναι ένα πιο απλό μοντέλο και κατά συνέπεια λιγότερο πιθανό να υπερπροσαρμόζεται στα δεδομένα. Μια συζήτηση με θέμα τη γενίκευση που παρέχουν αυτά τα μοντέλα πραγματοποιείται στη συνέχεια.

Από το Σχ. 5.2 εξάγονται και κάποια γενικά συμπεράσματα που αφορούν τις δύο

χρίσιμες παραμέτρους. Συγκεκριμένα, είναι φανερό πως οι επιδόσεις των μοντέλων για τα οποία `class_weight = balanced` είναι συντριπτικά καλύτερες από αυτών για τα οποία `class_weight = None`. Μάλιστα, οι επιδόσεις των τελευταίων είναι κοντινές με εκείνες που θα προέκυπταν για τυχαίες προβλέψεις. Αυτό σημαίνει πως τα δεδομένα του συνόλου εκπαίδευσης, και γενικότερα τα δεδομένα S3 είναι ανισοκατανεμημένα σε τέτοιο βαθμό που δεν μπορεί να προκύψει ένα ικανοποιητικό μοντέλο χωρίς η ιδιότητά τους αυτή να ληφθεί υπόψη. Πληροφοριακά αναφέρεται πως το 98% των παρατηρήσεων του συνόλου S3 αντιστοιχούν σε απουσία μανιταριών και μόλις το 2% σε παρουσία. Από τα παραπάνω γίνεται προφανές ότι η ανάγκη χρήσης της τιμής της παραμέτρου `class_weight = balanced` στη μελέτη είναι μάλλον επιτακτική, με αποτέλεσμα η μόνη παράμετρος που να έχει νόημα να εξεταστεί για τα Δέντρα Αποφάσεων να είναι η `max_depth`.

Σημειώνεται πως η επίδραση της επιλογής `class_weight = balanced` στο υπολογιστικό κόστος, σε αντίθεση με την τεράστια επίδραση της στην αξιοπιστία του μοντέλου, είναι αμελητέα. Συγκεκριμένα, στον υπολογιστή όπου εκτελέστηκαν οι υπολογισμοί, ο μέσος χρόνος που απαιτείται για την εκπαίδευση του μοντέλου με την επιλογή αυτή είναι 1.13 s, έναντι 1.08 s με την επιλογή `None`. Ο λόγος για τον οποίο δεν υπάρχει ουσιαστική μεταβολή είναι ότι η επιλογή `balanced` συνίσταται απλά στον υπολογισμό της συχνότητας κάθε τάξης στο σύνολο εκπαίδευσης και της απόδοσης

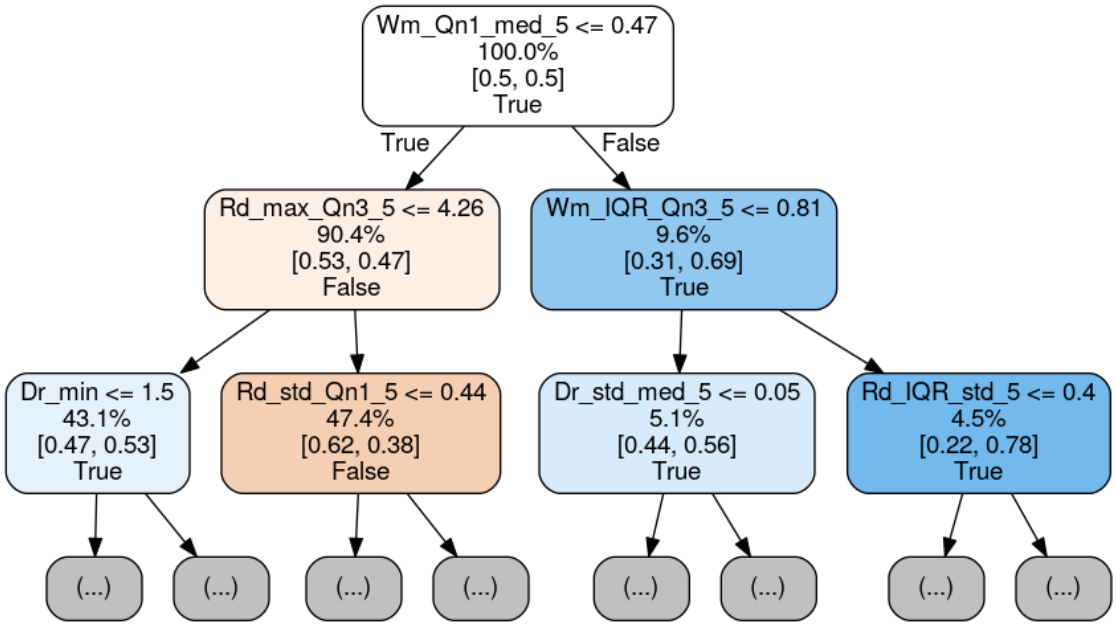


Σχήμα 5.2: Εξάρτηση του AUC των μοντέλων που εξέταστηκαν κατά την Αναζήτηση Πλέγματος για τα δεδομένα S3 από τις δύο παραμέτρους που ελέγχθηκαν.

του αντιστρόφου των δύο συχνοτήτων ως βαρύτητες στις παρατηρήσεις κάθε τάξης.

Στη συνέχεια αναλύεται συνοπτικά ο τρόπος λειτουργίας του βέλτιστου δέντρου που προέκυψε για τα δεδομένα S3. Όπως έχει ήδη αναφερθεί, ο πιο βασικός και εποπτικός τρόπος ανάλυσης ενός δέντρου είναι η οπτικοποίησή του. Μια τέτοια οπτικοποίηση, για τα 2 πρώτα από τα συνολικά 5 επίπεδα που αποτελούν το βέλτιστο δέντρο, παρουσιάζεται στο Σχ. 5.3. Ο λόγος που παρουσιάζεται ένα μόνο τμήμα του συνολικού δέντρου είναι για να κατανοηθεί η λογική πίσω από αυτό, χωρίς όμως η ανάλυση να ξεφύγει σε λεπτομέρεια. Όπως φαίνεται στο σχήμα, ξεκινώντας από το 100% των παρατηρήσεων, πραγματοποιείται η πρώτη δοκιμή, δηλαδή $\text{Wm\_Qn1\_med\_5} < 0.47$ μΩ. Το ερώτημα που τίθεται δηλαδή είναι αν η διάμεσος των τιμών 1ου τεταρτημορίου της αστάθειας για τα τελευταία 5 πόστα ήταν μικρότερη από 0.47 μΩ. Με βάση τη δοκιμή αυτή δημιουργούνται δύο υποσύνολα των δεδομένων: στο πρώτο, το οποίο περιέχει το 90.4% των αρχικών παρατηρήσεων, υπερισχύει οριακά η απουσία μανιταριού (False), με πιθανότητα 0.53 ή 53%, έναντι 0.47 της παρουσίας. Στο δεύτερο υπερισχύει η παρουσία μανιταριού (True), με πιθανότητα 0.69 έναντι 0.31 της απουσίας.

Περνώντας στο δεύτερο επίπεδο του δέντρου, η δοκιμή που πραγματοποιείται για να διαχωρίσει τις παρατηρήσεις του αριστερού συνόλου είναι η $\text{Rd\_max\_Qn3\_5} \leq 4.26$ μΩ, ενώ του δεξιού είναι η $\text{Wm\_IQR\_Qn3\_5} \leq 0.81$ μΩ. Το ερώτημα που θέτει η



Σχήμα 5.3: Οπτικοποίηση των 2 πρώτων επιπέδων του βέλτιστου Δέντρου Αποφάσεων που προέκυψε για τα δεδομένα S3.

πρώτη είναι αν η τιμή 3ου τεταρτημορίου των μεγίστων της διαφοράς αντίστασης για τα τελευταία 5 πόστα ήταν μικρότερη ή ίση του 4.26 μΩ. Με βάση τη δοκιμή αυτή, το αριστερό υποσύνολο χωρίζεται σε δύο νέα: στο πρώτο, το οποίο περιέχει πλέον το 43.1% των αρχικών παρατηρήσεων, υπερισχύει πλέον οριακά η παρουσία μανιταριού, με πιθανότητα 0.53 έναντι 0.47, ενώ στο δεύτερο, που περιέχει το 47.4% των παρατηρήσεων, υπερισχύει με διαφορά η απουσία, με πιθανότητα 0.62 έναντι 0.38.

Το ερώτημα που θέτει η δεύτερη δοκιμή του δεύτερου επιπέδου είναι αν η τιμή 3ου τεταρτημορίου των διατεταρτημοριακών ευρών της αστάθειας για τα τελευταία 5 πόστα είναι μικρότερη ή ίση του 0.81 μΩ. Με βάση τη δοκιμή αυτή, το δεξί υποσύνολο χωρίζεται σε δύο νέα: στο πρώτο, το οποίο περιέχει το 5.1% των αρχικών παρατηρήσεων, υπερισχύει η παρουσία μανιταριού, με πιθανότητα 0.56 έναντι 0.44, ενώ στο δεύτερο, που περιέχει το 4.5% των παρατηρήσεων, υπερισχύει επίσης η παρουσία με 0.78 έναντι 0.22. Στο σχήμα φαίνονται και οι δοκιμές που οδηγούν στο τρίτο επίπεδο του δέντρου, η περαιτέρω ανάλυση τους όμως μάλλον δεν προσφέρει κάτι στη συζήτηση.

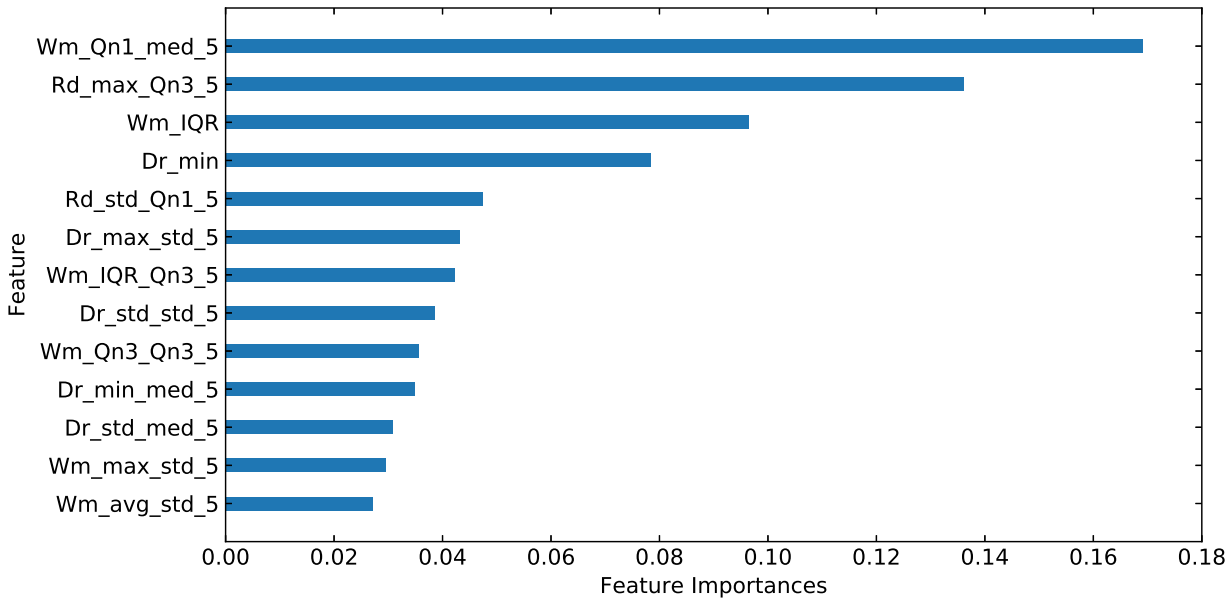
Αν και εκ πρώτης όψεως τα χαρακτηριστικά των παρατηρήσεων, άρα και οι δοκιμές, μοιάζουν σχετικά περίπλοκα, στην πραγματικότητα είναι μάλλον απλά. Για παράδειγμα, η πρώτη από όλες μπορεί να συνοψιστεί στο ερώτημα αν οι τιμές της αστάθειας στα τελευταία 5 πόστα ήταν υψηλές ή χαμηλές. Παρόλα αυτά, το πλήθος των επιπέδων του δέντρου καθιστά τη λογική πίσω από αυτό μη προσιτή σε έναν άνθρωπο. Γι΄ αυτό, για την κατανόηση του τρόπου με τον οποίο λειτουργεί το δέντρο, έναντι της οπτικοποίησής του προκρίνεται ο υπολογισμός των τιμών σημαντικότητας κάθε χαρακτηριστικού και η εύρεση των πιο σημαντικών χαρακτηριστικών.

Στο Σχ. 5.4 απεικονίζονται οι τιμές σημαντικότητας για τα 13 πιο σημαντικά, σύμφωνα με αυτό το κριτήριο, χαρακτηριστικά. Υπενθυμίζεται ότι η κάθε μεμονωμένη τιμή κυμαίνεται από 0 έως 1, ενώ το άθροισμα όλων των τιμών είναι επίσης το 1, οι τιμές δηλαδή είναι κανονικοποιημένες. Επίσης τονίζεται εκ νέου ότι η σημαντικότητα δεν υποδηλώνει μια σχέση μεταξύ κάποιων τιμών ενός χαρακτηριστικού και της μίας ή της άλλης τάξης, όπως υποδηλώνει η θετική ή αρνητική τιμή κάποιου εκ των συντελεστών ενός γραμμικού μοντέλου. Η σημαντικότητα δείχνει μόνο το σε ποιο βαθμό αξιοποιείται το κάθε χαρακτηριστικό για το διαχωρισμό των παρατηρήσεων μεταξύ των δύο τάξεων μέσα στο δέντρο.

Πιο συγκεκριμένα, τα δύο πιο σημαντικά χαρακτηριστικά είναι αυτά πάνω στα οποία βασίστηκαν 2 από τις 3 δοκιμές των δύο πρώτων επιπέδων του δέντρου. Το χαρακτηριστικό πάνω στο οποίο βασίστηκε η άλλη από τις 3 δοκιμές, όμως, βρίσκεται στην 7η θέση της σχετικής κατάταξης, πιθανότατα επειδή κατά μήκος του δέντρου συνεισφέρει λιγότερο στο διαχωρισμό των παρατηρήσεων σε σχέση με αυτά που βρίσκονται

στις θέσεις 2-6. Παρατηρώντας τα 4 πρώτα χαρακτηριστικά, τα οποία φαίνεται να ξεχωρίζουν από τα υπόλοιπα, προκύπτει το συμπέρασμα ότι μερικά από τα στοιχεία που καθορίζουν αν σε μια λεκάνη υπάρχει ή όχι μανιτάρι είναι το αν η αστάθεια είναι γενικά χαμηλή, αν η διαφορά μεταξύ της ζητούμενης και της εφαρμοζόμενης αντίστασης είναι μεγάλη, αν η αστάθεια εμφανίζει σημαντικές διακυμάνσεις και αν η εκάστοτε τιμή του τεστ είναι χαμηλή, αντίστοιχα. Επίσης, ενδεικτικά αναφέρεται ότι 97 από τα 120 χαρακτηριστικα παρουσιάζουν μηδενικές τιμές σημαντικότητας, δηλαδή δεν αξιοποιούνται καθόλου από το μοντέλο. Η παρατήρηση αυτή αναλύεται περισσότερο στη συνέχεια, στη σύγκριση μεταξύ Δέντρων Αποφάσεων και Τυχαίων Δασών.

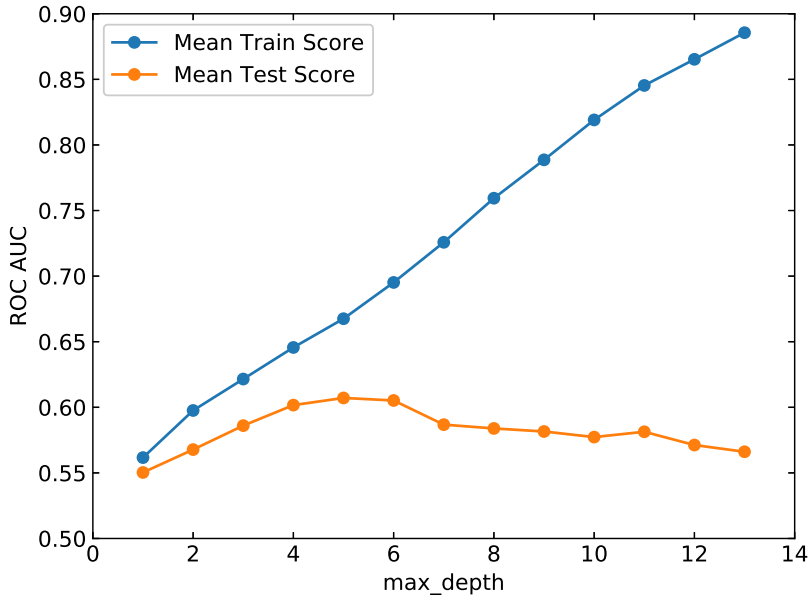
Μια προσέγγιση που θα μπορούσε να ακολουθηθεί με βάση τα παραπάνω είναι τα χαρακτηριστικά τα οποία περιέχουν σημαντική πληροφορία, σύμφωνα με ένα Δέντρο Αποφάσεων ή ένα Τυχαίο Δάσος, να τροφοδοτηθούν σε κάποιον άλλο αλγόριθμο, π.χ. για την παραγωγή ενός μοντέλου *Εγγύτερων Γειτόνων* (*Nearest Neighbors*). Για την ακρίβεια, μια τέτοια προσέγγιση υποστηρίζεται και από τη βιβλιογραφία, καθώς οι αλγόριθμοι που βασίζονται σε Δέντρα Αποφάσεων χρησιμοποιούνται ως μέθοδος *Επιλογής Χαρακτηριστικών* (*Feature Selection*), πριν από την ουσιαστική εκπαίδευση κάποιου άλλου μοντέλου. Αυτό δοκιμάστηκε και στην περίπτωση της παρούσας μελέτης, στο πλαίσιο της διερεύνησης άλλων μοντέλων, τα οποία δεν παρουσιάζονται στο τελικό κείμενο. Σε εκείνες τις δοκιμές, οι επιδόσεις των άλλων μοντέλων ήταν



Σχήμα 5.4: Υπολογισμένες τιμές της σημαντικότητας των διάφορων χαρακτηριστικών για το βέλτιστο Δέντρο Αποφάσεων που προέκυψε για τα δεδομένα S3.

υποδεέστερες έναντι αυτών που αξιοποιούν αποκλειστικά και απευθείας μοντέλα βασισμένα στα Δέντρα Αποφάσεων, αφού τα άλλα μοντέλα, όπως π.χ. και οι Εγγύτεροι Γείτονες, υποφέρουν από διάφορες αδυναμίες που δεν απαντώνται στους αλγορίθμους που εξετάζονται εδώ. Μια από αυτές είναι η αδυναμία στα δεδομένα με Ανισοκατανεμημένες Τάξεις, η οποία επιτάσσει τη χρήση πρόσθετων τεχνικών, όπως υπερ- ή υποδειγματοληψίας.

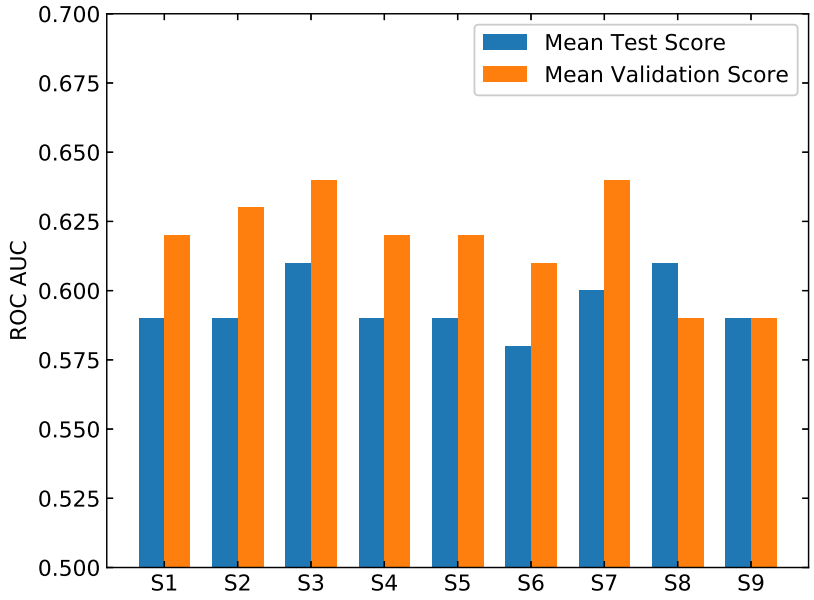
Το τελευταίο ενδεικτικό σχήμα που παρουσιάζεται με βάση τα δεδομένα S3 είναι το Σχ. 5.5. Με αφορμή το σχήμα αυτό πραγματοποιείται μια σύντομη συζήτηση περί γενίκευσης και προσαρμογής. Όπως έχει ήδη αναφερθεί σε προηγούμενο σημείο της εργασίας, το ζητούμενο της Επιτηρούμενης Μάθησης είναι η εύρεση ενός μοντέλου που να *γενικεύει* ικανοποιητικά από το σύνολο εκπαίδευσης στο σύνολο ελέγχου. Αυτό σημαίνει ότι το μοντέλο δεν είναι τόσο σύνθετο που να υπερπροσαρμόζεται στα δεδομένα, δηλαδή να μαθαίνει τόσο πολύ τις ιδιαιτερότητες του συνόλου εκπαίδευσης που να αδυνατεί να προβλέψει το σύνολο ελέγχου. Από την άλλη, δεν είναι ούτε τόσο απλό που να υποπροσαρμόζεται, δηλαδή να μαθαίνει τόσο λίγο που να αδυνατεί να προβλέψει ικανοποιητικά οποιοδήποτε από τα δύο σύνολα. Είχε μάλιστα αναφερθεί πως το σημείο ισορροπίας για το οποίο επιτυγχάνεται η βέλτιστη γενίκευση καθώς αυξάνεται η περιπλοκότητα ενός μοντέλου μπορεί να βρεθεί διαγραμματικά.



Σχήμα 5.5: Εξάρτηση του AUC που υπολογίστηκε για τα σύνολα εκπαίδευσης και ελέγχου των δεδομένων S3 από την περιπλοκότητα των μοντέλων που εξετάστηκαν, εκφρασμένης μέσω της τιμής της παραμέτρου max_depth.

Στο Σχ. 5.5 είναι δυνατό να φανούν όλα τα παραπάνω. Συγκεκριμένα, για τα μοντέλα που εξετάζονται εδώ, η μοναδική κρίσιμη παράμετρος είναι το μέγιστο πλήθος επιπέδων του δέντρου, `max_depth`. Η τιμή της παραμέτρου αυτής προφανώς καθορίζει και την περιπλοκότητα του μοντέλου, με μεγαλύτερες τιμές να αντιστοιχούν σε σύνθετα μοντέλα και μικρότερες σε απλά. Όπως φαίνεται στο σχήμα, η επίδοση του μοντέλου στα δεδομένα του συνόλου εκπαίδευσης, εκφρασμένη μέσω του AUC, αυξάνεται συνεχώς καθώς αυξάνεται και η περιπλοκότητα του μοντέλου. Αντίθετα, η επίδοση στα δεδομένα του συνόλου ελέγχου αρχικά αυξάνεται έως ότου λάβει τη μέγιστη τιμή της για $\text{max\_depth} = 5$, ενώ στη συνέχεια μειώνεται. Σύμφωνα με τα παραπάνω, τα μοντέλα για τα οποία $\text{max\_depth} < 5$ υποπροσαρμόζονται στα δεδομένα, ενώ τα μοντέλα για τα οποία $\text{max\_depth} > 5$ υπερπροσαρμόζονται σε αυτά. Για την οριακή τιμή της παραμέτρου προκύπτει το μοντέλο που γενικεύει κατά το δυνατόν πιο ικανοποιητικά, οπότε και επιλέγεται ως το βέλτιστο.

Ολοκληρώνοντας, πραγματοποιείται μια σύγκριση μεταξύ των βέλτιστων μοντέλων που προέκυψαν για τα διάφορα *ασφαλή* σύνολα δεδομένων. Η σύγκριση αυτή πραγματοποιείται στη βάση των AUC που υπολογίστηκαν από την εφαρμογή κάθε ενός από τα μοντέλα σε ένα σύνολο διακρίβωσης και απεικονίζεται στο Σχ. 5.6. Αυτό το σύνολο περιέχει παρατηρήσεις που δεν χρησιμοποιήθηκαν ούτε για την εκπαίδευση,



Σχήμα 5.6: Σύγκριση των AUC που προέκυψαν από την εφαρμογή των βέλτιστων μοντέλων για τα Δέντρα Αποφάσεων σε κάθε ένα από τα *ασφαλή* σύνολα δεδομένων, στα σύνολα ελέγχου και διακρίβωσης.

ούτε και για την επιλογή του καλύτερου μοντέλου από την Αναζήτηση Πλέγματος που πραγματοποιήθηκε για κάθε σύνολο δεδομένων. Σε αντίθετη περίπτωση μια αξιολόγηση με βάση αυτό δε θα ήταν έγκυρη, αφού τα δεδομένα αυτά θα είχαν ήδη καθορίσει την επιλογή των βέλτιστων παραμέτρων. Φυσικά, το σύνολο στο οποίο εφαρμόζεται κάθε μοντέλο είναι διαφορετικό και προέρχεται από την αντίστοιχη εκδοχή των *ασφαλών* δεδομένων, οπότε η σύγκριση δεν είναι απόλυτη αλλά ενδεικτική του αν κάποιο σύνολο οδήγησε στην κατασκευή ενός δραματικά καλύτερου ή χειρότερου μοντέλου σε σχέση με τα υπόλοιπα.

Από το Σχ. 5.6 φαίνεται πως τόσο τα AUC επί των συνόλων ελέγχου, τα οποία περιλαμβάνονται στο σχήμα για λόγους σύγκρισης, όσο και αυτά που υπολογίστηκαν επί των συνόλων διακρίβωσης είναι κοντινά μεταξύ τους. Επίσης, δεν παρατηρείται κάποια κανονικότητα μεταβαίνοντας από το S1 στο S9, π.χ. τα αποτελέσματα είναι σχετικά κοντά είτε αγνοούνται 3 μέρες μετά την αφαίρεση των μανιταριών, είτε 7. Αυτό σημαίνει πως οι τιμές των χαρακτηριστικών επανέρχονται στο να αντιστοιχούν σε απουσία μανιταριού μάλλον γρήγορα μετά την αφαίρεσή του. Αν έπρεπε να επιλεγεί ένα σύνολο δεδομένων, αυτό θα ήταν μάλλον το S3, το οποίο παρουσιάζει την καλύτερη επίδοση επί του συνόλου διακρίβωσης, $\mathrm{AUC} = 0.64$, ενώ ταυτόχρονα δεν προκύπτει από κάποια ακραία υπόθεση, όπως π.χ. το S7 που υποθέτει ότι απαιτούνται 7 μέρες για να επανέλθει η λεκάνη μετά την αφαίρεση του μανιταριού. Παρόλα αυτά, η σχετική επιλογή θα γίνει με βάση τα αποτελέσματα που προκύπτουν από τα μοντέλα Τυχαίων Δασών.

Τέλος, εντύπωση προκαλεί η διαφορά που παρατηρείται στις επιδόσεις του μοντέλου όταν αυτό εφαρμόζεται στα σύνολα ελέγχου κατά την Αναζήτηση Πλέγματος και στα σύνολα διακρίβωσης. Σε όλες τις περιπτώσεις εκτός του S8, το μοντέλο παρουσιάζει καλύτερη επίδοση στο άγνωστο σύνολο διακρίβωσης. Το γεγονός αυτό, αν και δεν προκαλεί ανησυχία ότι το μοντέλο έχει υπερπροσαρμοστεί στα δεδομένα εκπαίδευσης (γιατί σε αυτήν την περίπτωση θα συνέβαινε το ακριβώς αντίθετο) είναι πιθανό να οφείλεται απλά στο ότι τόσο η επίδοση στα σύνολα ελέγχου, όσο και στα σύνολα διακρίβωσης είναι σχετικά χαμηλές.

5.1.2 Τυχαία Δάση

Η δεύτερη οικογένεια μοντέλων που αναπτύχθηκε περιέχει μοντέλα βασισμένα στα Τυχαία Δάση. Η εφαρμογή του σχετικού αλγορίθμου που αξιοποιείται σε αυτήν την περίπτωση είναι ο `RandomForestClassifier` της βιβλιοθήκης scikit-learn. Κάθε δάσος αποτελείται από 100 τυχαιοποιημένα Δέντρα Αποφάσεων. Εξετάζονται τα μοντέλα που προκύπτουν μεταβάλλοντας τρεις παραμέτρους του αλγορίθμου:

- `max_depth`: Το μέγιστο βάθος του δέντρου. Ισχύουν όλα όσα έχουν ήδη αναφερθεί στην Υποενότητα 5.1.1 για τα Δέντρα Αποφάσεων.
- `class_weight`: Τα βάρη που αποδίδονται σε κάθε μία από τις δύο τάξεις. Στην περίπτωση των Τυχαίων Δασών δεν κρίνεται σκόπιμο να χρησιμοποιηθεί η τιμή της παραμέτρου `class_weight` = `None`, η οποία στην περίπτωση των Δέντρων Αποφάσεων οδήγησε σε επιδόσεις παραπλήσιες με αυτές των τυχαίων προβλέψεων, υστερώντας σημαντικά έναντι της εναλλακτικής επιλογής, `balanced`. Αυτή είναι και η πρώτη μεταξύ των δύο τιμών της παραμέτρου `class_weight` που εξετάζονται σε αυτή την περίπτωση. Η δεύτερη, `balanced_subsample` έχει την ίδια λειτουργία με την `balanced`, δηλαδή αποδίδει στις δύο τάξεις βαρύτητες αντιστρόφως ανάλογες τις συχνότητας τους στα δεδομένα. Αντί όμως να λαμβάνει υπόψη τις συχνότητες στα αρχικά δεδομένα, υπολογίζει και χρησιμοποιεί τις συχνότητες που προκύπτουν από τα δεδομένα του bootstrap sample.[1]
- `max_features`: Το πλήθος των τυχαίων χαρακτηριστικών που εξετάζονται από την αναζήτηση για την εκάστοτε βέλτιστη δοκιμή κατά την κατασκευή των δέντρων του δάσους. Μεταξύ των διαθέσιμων επιλογών, αυτές που αξιοποιούνται στην παρούσα μελέτη είναι `max_features = sqrt` ή `log2`, με την πρώτη να αντιστοιχεί στην τετραγωνική ρίζα και την δεύτερη στο λογάριθμο με βάση το 2 του πλήθους των χαρακτηριστικών. Σημειώνεται ότι εδώ το πλήθος αυτό ισούται με 120, οπότε οι δύο τιμές είναι 7 και 11, αντίστοιχα.

Οι τιμές των τριών παραμέτρων που τελικά εξετάζονται είναι:

- `max_depth`: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]
- `class_weight`: [`balanced, balanced_subsample`]
- `max_features`: [`sqrt, log2`]

Στη συνέχεια παρουσιάζονται τα αποτελέσματα που αφορούν την οικογένεια των *ασφαλών* δεδομένων. Όπως και για τα Δέντρα Αποφάσεων, πρώτα παρουσιάζονται ενδεικτικά διαγράμματα, τα οποία συγκεκριμένα προέρχονται από τη διερεύνηση που έγινε για τα δεδομένα S2, δηλαδή αυτά στα οποία για κάθε λεκάνη έχουν αγνοηθεί τα δεδομένα 15 ημερών πριν τον εντοπισμό του μανιταριού και 2 ημερών μετά, εξού και S2.

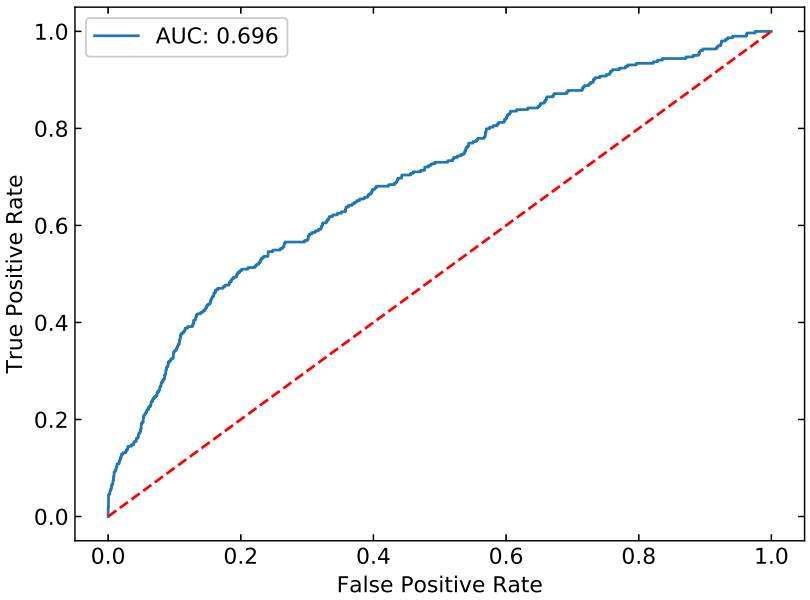
Στο Σχ. 5.7 απεικονίζεται η καμπύλη ROC που αντιστοιχεί στην εφαρμογή του καλύτερου μοντέλου στα δεδομένα του συνόλου διακρίβωσης. Η τιμή του εμβαδού κάτω από αυτήν την καμπύλη, AUC = 0.7 δεν είναι πολύ υψηλή, θα μπορούσε όμως να

[1] Υπενθυμίζεται ότι αυτό είναι το δείγμα πάνω στο οποίο εκπαιδεύεται κάθε δέντρο του Τυχαίου Δάσους και αποτελεί ένα σύνολο ίσου μεγέθους με το σύνολο εκπαίδευσης αλλά που έχει προκύψει από αυτό με τυχαία δειγματοληψία με αντικατάσταση.

θεωρηθεί ικανοποιητική. Άλλωστε, δεν υπάρχει κάποιος γενικός κανόνας που να ορίζει το από ποια τιμή και πάνω η γενίκευση που προσφέρει ένα μοντέλο είναι ικανοποιητική για κάθε είδος προβλήματος. Επίσης, η τιμή αυτή είναι πολύ υψηλότερη του 0.64, που υπολογίστηκε ως η βέλτιστη για τα μοντέλα βασισμένα σε Δέντρα Αποφάσεων. Επιβεβαιώνεται δηλαδή η θεωρητική προσδοκία ότι τα Τυχαία Δάση υπερισχύουν των Δέντρων Αποφάσεων, επειδή με την τυχαιότητα που εισάγουν δεν υπερπροσαρμόζονται στο σύνολο εκπαίδευσης και γενικεύουν καλύτερα.

Ένας προσεκτικός αναγνώστης ενδεχομένως παρατηρήσει πως η καμπύλη ROC του Σχ. 5.7 είναι ομαλότερη εκείνης του Σχ. 5.1, αποτελείται δηλαδή από αρκετά περισσότερα σημεία. Αυτό δεν είναι απόλυτα ακριβές: στην πραγματικότητα, και οι δύο καμπύλες αποτελούνται από το ίδιο πλήθος σημείων. Αυτό που συμβαίνει είναι ότι στην περίπτωση της καμπύλης του Σχ. 5.1, που φαινομενικά αποτελείται από λιγότερα σημεία, ορισμένες από τις διαφορετικές τιμές του ορίου απόφασης που χρησιμοποιούνται για το σχηματισμό της καμπύλης δίνουν ακριβώς το ίδιο αποτέλεσμα σε όρους FPR και TPR.

Στο Σχ. 5.8 απεικονίζεται η εξάρτηση του AUC των μοντέλων που εξετάστηκαν κατά την Αναζήτηση Πλέγματος από τις τιμές και των τριών κρίσιμων παραμέτρων του αλ-

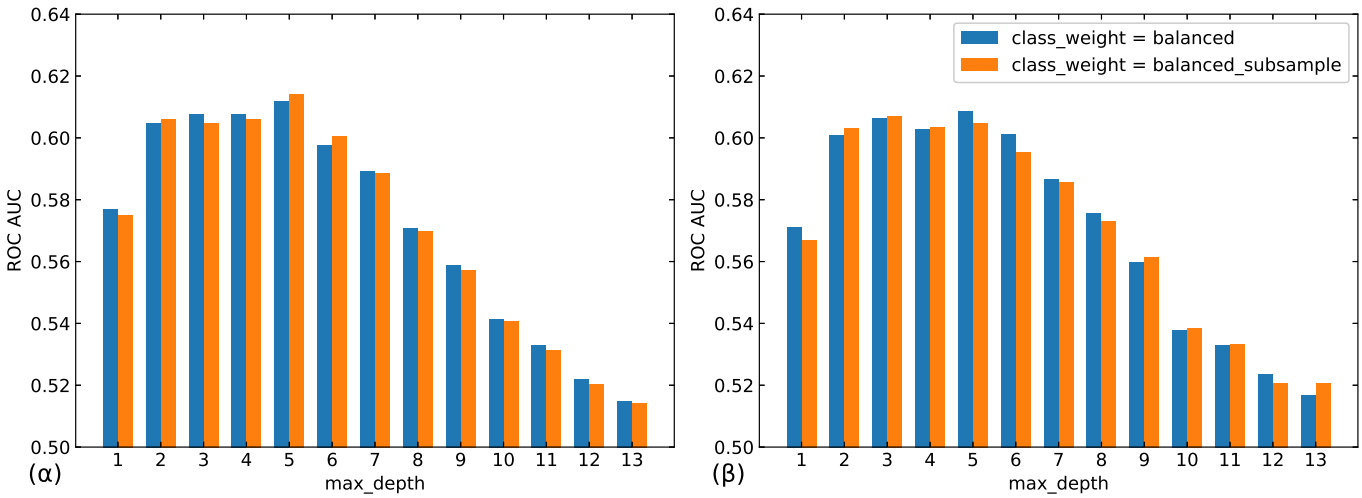


Σχήμα 5.7: Καμπύλη ROC του βέλτιστου Τυχαίου Δάσους που προέκυψε για τα δεδομένα S2, εφαρμοσμένου στο σύνολο διακρίβωσης. Τιμές κρίσιμων παραμέτρων: `max_depth` = 5, `class_weight = balanced_subsample`, `max_features = log2`.

γορίθμού. Το αριστερό και το δεξί διάγραμμα αντιστοιχούν στις τιμές `log2` και `sqrt` της παραμέτρου `max_features`, αντιστοιχα. Σε καθε ένα από τα δύο διαγράμματα, στις διάφορες θέσεις στον άξονα x αντιστοιχούν οι τιμές της παραμέτρου `max_depth` και οι δύο τιμές της παραμέτρου `class_weight` απεικονίζονται με διαφορετικό χρώμα, όπως αναφέρεται στο υπόμνημα.

Από το Σχ. 5.8 εξάγονται κάποια γενικά συμπεράσματα που αφορούν τις κρίσιμες παραμέτρους. Συγκεκριμένα, είναι φανερό πως οι επιδόσεις των μοντέλων δεν εξαρτώνται ιδιαίτερα από την τιμή της παραμέτρου `class_weight`. Αυτό σημαίνει ότι και οι δύο επιλογές αποκαθιστούν επαρκώς το πρόβλημα που προκύπτει από την άνιση κατανομή των δεδομένων στις δύο τάξεις, σε αντίθεση με την επιλογή `None`, η οποία δοκιμάστηκε στα Δέντρα Αποφάσεων και πρακτικά οδηγούσε σε τυχαίες προβλέψεις. Η επιλογή αυτή προφανώς είναι διαθέσιμη και σε αυτήν την περίπτωση, απλά δεν κρίθηκε σκόπιμο να διερευνηθεί περαιτέρω. Οι δύο τιμές της παραμέτρου που έχουν αποτέλεσμα υπερισχύουν κατά περίπτωση, οι επιδόσεις τους όμως είναι τόσο κοντινές που φαίνεται ότι για αυτά τα δεδομένα μπορεί να χρησιμοποιηθεί απευθείας οποιαδήποτε από τις δύο.

Κάτι αντίστοιχο συμβαίνει και για την παράμετρο `max_features`. Οι τιμές AUC που υπολογίζονται για `max_features = log2` (αριστερά) και `sqrt` (δεξιά) είναι παραπλήσιες, ενώ δεν παρατηρείται και κάποια διαφορά στην τάση που έχουν συναρτήσει και των τιμών των άλλων δύο παραμέτρων. Αυτό φαίνεται ακόμη πιο ξεκάθαρα στο

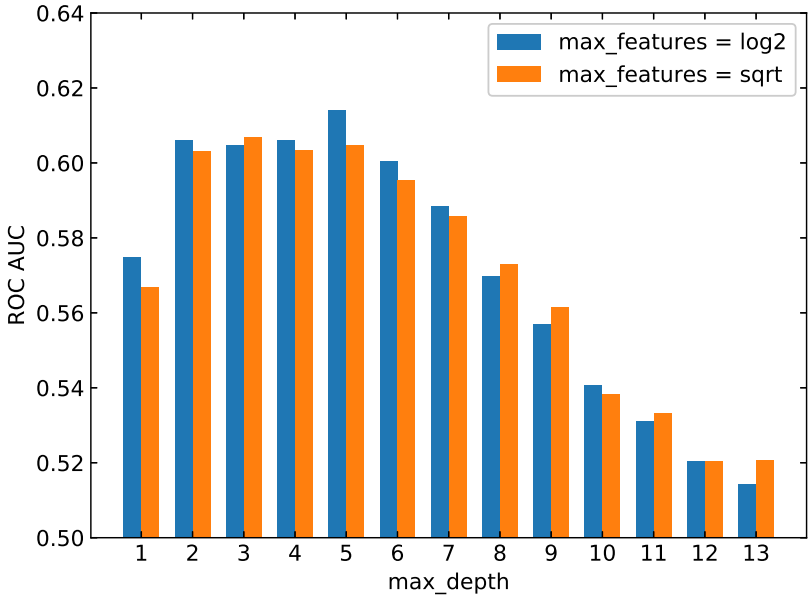


Σχήμα 5.8: Εξάρτηση του AUC των μοντέλων που εξέταστηκαν κατά την Αναζήτηση Πλέγματος για τα δεδομένα S2 από τις τρεις παραμέτρους που ελέγχθηκαν: (α) `max_features = log2`, (β) `max_features = sqrt`.

Σχ. 5.9, όπου συγκρίνονται οι επιδόσεις για τα τις δύο τιμές αυτής της παραμέτρου, με την τιμή της παραμέτρου σταθερή και ίση με `class_weight = balanced_subsample`.

Το Σχ. 5.9 μπορεί να αξιοποιηθεί για να οπτικοποιηθεί η εξάρτηση από την παράμετρο `max_depth` αφού, όπως εξηγήθηκε παραπάνω, οι άλλες δύο παράμετροι δεν επηρεάζουν ιδιαίτερα το αποτέλεσμα. Σε αυτό φαίνεται ότι η εξάρτηση της επίδοσης από αυτή την παράμετρο είναι ακριβώς η ίδια που παρατηρήθηκε και για τα Δέντρα Αποφάσεων, χωρίς αυτό να προκαλεί απορία αφού στην ουσία ένα Τυχαίο Δάσος δεν είναι τίποτα άλλο παρά πολλά, παρόμοια μεταξύ τους, Δέντρα Αποφάσεων. Πιο συγκεκριμένα, τα Τυχαία Δάση με μικρό πλήθος επιπέδων συμπεριφέρονται αρκετά καλύτερα, ενώ για πλήθος επιπέδων μεγαλύτερο του 6 η επίδοση των δασών χειροτερεύει ραγδαία. Παρότι το βέλτιστο δέντρο προέκυψε για `max_depth` $= 5$, παρατηρείται ότι για τιμές της παραμέτρου ίσες με 4, 3, αλλά ακόμη και 2, τα αποτελέσματα επί του συνόλου ελέγχου δεν διαφέρουν ιδιαίτερα. Αυτό σημαίνει πως θα μπορούσε ίσως να επιλεγεί και κάποιο από αυτά τα μοντέλα, με την ελπίδα να γενικεύει καλύτερα, λόγω του χαμηλότερου πλήθους των επιπέδων του.

Στη γενική περίπτωση, ένα παράπλευρο όφελος από την επιλογή ενός μοντέλου με λιγότερα επίπεδα είναι η μείωση του υπολογιστικού κόστους τόσο της εκπαίδευσης, όσο και των προβλέψεων που αυτό πραγματοποιεί. Το όφελος αυτό, όμως, είναι α-

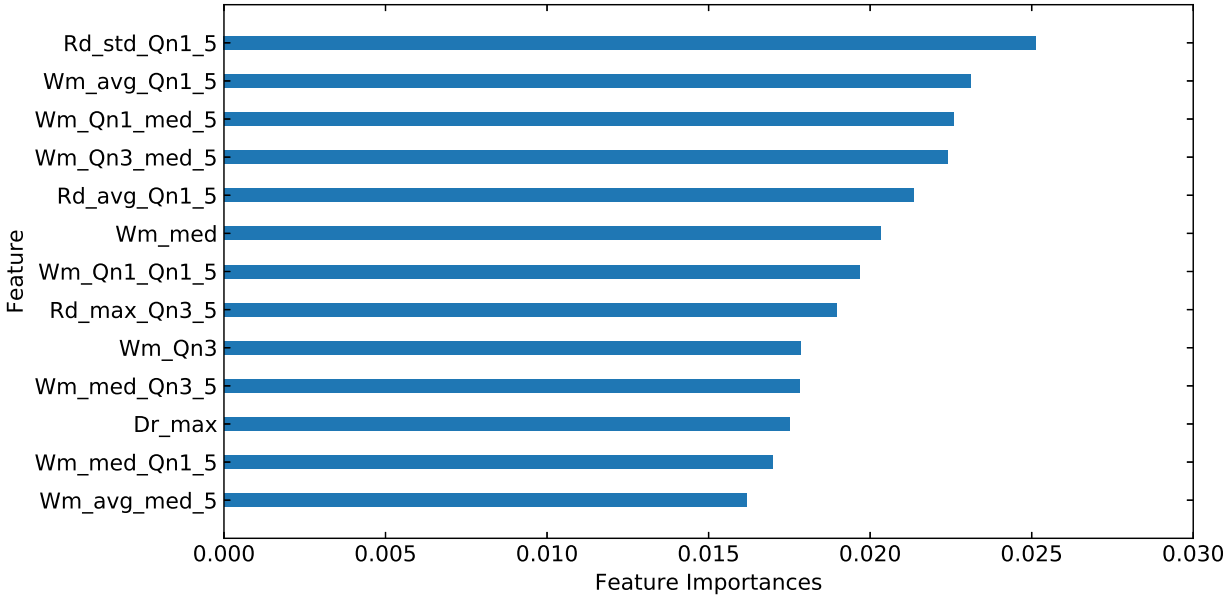


Σχήμα 5.9: Εξάρτηση του AUC των μοντέλων που εξέταστηκαν κατά την Αναζήτηση Πλέγματος για τα δεδομένα S2 από τις παραμέτρους `max_depth` και `max_features`.

μελητέο στην παρούσα περίπτωση, αφού η εκπαίδευση θα πραγματοποιηθεί μία φορά πριν από την έναρξη χρήσης του συστήματος πρόβλεψης και θα επαναλαμβάνεται σε πολύ αραιά χρονικά διαστήματα στη συνέχεια. Ενδεχομένως ακόμη σημαντικότερο να είναι το γεγονός ότι ο χρόνος που απαιτείται για την εκπαίδευση αυτή είναι πολύ μικρός. Ενδεικτικά αναφέρεται ότι, στον υπολογιστή όπου εκτελέστηκαν οι υπολογισμοί, ο μέσος χρόνος εκπαίδευσης για το μοντέλο με τα 5 επίπεδα ήταν 6.72 s, έναντι 4.29 s για αυτό με τα μόλις 2 επίπεδα. Ο δε χρόνος που απαιτείται για μία πρόβλεψη είναι αμελητέος. Συνεπώς, το υπολογιστικό κόστος δεν είναι ένα από τα κριτήρια που απαιτείται να συνυπολογιστεί στην επιλογή του μοντέλου που θα αξιοποιηθεί.

Φυσικά, η επιλογή ενός δάσους με λιγότερα επίπεδα δεν προσφέρει κάτι στην οπτικοποίηση του μοντέλου, όπως στην περίπτωση των Δέντρων Αποφάσεων, αφού τα Τυχαία Δάση αποτελούνται από δεκάδες διαφορετικά δέντρα και ως εκ τούτου δεν οπτικοποιούνται. Ο μόνος τρόπος με τον οποίο μπορεί να κατανοηθεί ο τρόπος λειτουργίας του Τυχαίου Δάσους είναι μέσω του υπολογισμού των τιμών σημαντικότητας κάθε χαρακτηριστικού και η εύρεση των πιο σημαντικών χαρακτηριστικών. Οι τιμές σημαντικότητας για τα 13 πιο σημαντικά χαρακτηριστικά του βέλτιστου Τυχαίου Δάσους για τα δεδομένα S2 απεικονίζονται στο Σχ. 5.10.

Με την πρώτη ματιά παρατηρείται ότι σε αυτήν την περίπτωση δεν υπάρχουν κάποια



Σχήμα 5.10: Υπολογισμένες τιμές της σημαντικότητας των διάφορων χαρακτηριστικών για το βέλτιστο Τυχαίο Δάσος που προέκυψε για τα δεδομένα S2.

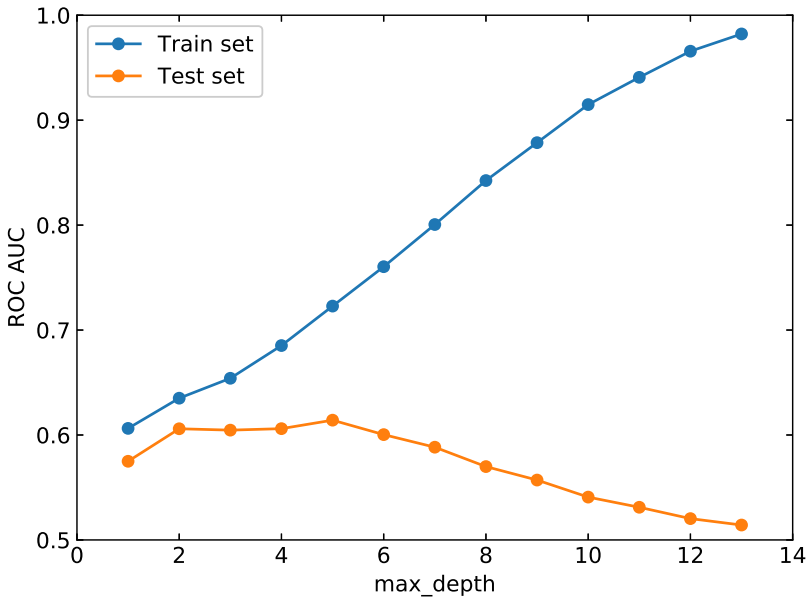
χαρακτηριστικά που να ξεχωρίζουν δραστικά από τα υπόλοιπα, όπως στην περίπτωση των Δέντρων Αποφάσεων (βλ. Σχ. 5.4). Μια πιο προσεκτική παρατήρηση των τιμών δείχνει ότι η μέγιστη σημαντικότητα κάποιου χαρακτηριστικού είναι περίπου 0.025, σε αντιδιαστολή με την τιμή 0.170 για το βέλτιστο Δέντρο Αποφάσεων. Αυτό υποδηλώνει πως το Τυχαίο Δάσος λαμβάνει υπόψη πολύ περισσότερα χαρακτηριστικά, με χαμηλότερη σημαντικότητα το καθένα, χωρίς δηλαδή να βασίζει τις προβλέψεις του εξ΄ ολοκλήρου σε μια μικρή ομάδα χαρακτηριστικών. Πηγαίνοντας την ανάλυση ένα βήμα παραπέρα, διαπιστώνεται ότι το πλήθος των χαρακτηριστικών που παρουσιάζουν μηδενικές τιμές σημαντικότητας για το Τυχαίο Δάσος είναι μόλις 4, έναντι 97 (!) για το βέλτιστο Δέντρο Αποφάσεων. Αυτό σημαίνει πως η τυχαιότητα στην κατασκευή των δέντρων του δάσους, η οποία πηγάζει τόσο από τα bootstrap samples, όσο και από την επιλογή κάθε δοκιμής από μικρά υποσύνολα του συνόλου των χαρακτηριστικών, οδηγεί σε μεγαλύτερη αξιοποίηση της διαθέσιμης πληροφορίας. Η μεταβολή αυτή πιθανότατα επιφέρει και τη σημαντική βελτίωση στις επιδόσεις των μοντέλων που βασίζονται σε Τυχαία Δάση.

Όλα τα παραπάνω εξηγούν το γιατί τα κοινά χαρακτηριστικά μεταξύ των πιο σημαντικών για τα δύο είδη μοντέλων είναι μόλις 3 στα 13. Τα χαρακτηριστικά αυτά είναι τα Rd_std_Qn1_5, Wm_Qn1_med_5 και Rd_max_Qn3_5, δηλαδή η τιμή 1ου τεταρτημορίου των τιμών τυπικής απόκλισης της διαφοράς αντίστασης, η διάμεσος των τιμών 1ου τεταρτημορίου της αστάθειας και η τιμή 3ου τεταρτημορίου των μεγίστων της διαφοράς αντίστασης, όλες για τα τελευταία 5 πόστα, αντίστοιχα. Φυσικά, πολλά χαρακτηριστικά από τα 120 που τροφοδοτούνται στους αλγορίθμους έχουν παρόμοιο περιεχόμενο, οπότε μπορεί και κάποια άλλα από τις λίστες πιο σημαντικών χαρακτηριστικών να είναι κοντινά μεταξύ τους. Υπενθυμίζεται ότι τόσο τα Δέντρα Αποφάσεων, όσο και τα Τυχαία Δάση πραγματοποιούν αυτόματη Επιλογή Χαρακτηριστικών, λόγω του τρόπου με τον οποίο λειτουργούν. Αυτή είναι και η ιδιότητα τους που επιτρέπει να τροφοδοτηθούν με όλα τα πιθανά χαρακτηριστικά, χωρίς αυτό να επηρεάσει την ικανότητα του αλγορίθμου να προσεγγίσει τη βέλτιστη λύση.

Το τελευταίο ενδεικτικό σχήμα που παρουσιάζεται για τα Τυχαία Δάση με βάση τα δεδομένα S2 είναι το Σχ. 5.11. Στο σχήμα αυτό είναι δυνατό να επιβεβαιωθούν εκ νέου όλα όσα έχουν αναφερθεί περί γενίκευσης και προσαρμογής των μοντέλων στα δεδομένα, ανάλογα με το βαθμό περιπλοκότητας τους. Η τιμή του AUC αυξάνεται καθώς το μέγιστο βάθος του δέντρου αυξάνεται από το 1 έως το 5, που είναι και η οριακή τιμή, ενώ περαιτέρω αύξηση της παραμέτρου αυτής οδηγεί σε υποβάθμιση της επίδοσης του μοντέλου λόγω υπερπροσαρμογής. Καθώς τα δέντρα του Τυχαίου Δάσους χρησιμοποιούν περισσότερες δοκιμές για να πραγματοποιήσουν τις προβλέψεις τους, γίνονται δηλαδή πιο σύνθετα, τα μοντέλα που προκύπτουν *απομνημονεύουν* το σύνολο εκπαίδευσης, οπότε αδυνατούν να προβλέψουν σωστά όταν αντιμετωπίζουν τα άγνωστα δεδομένα του συνόλου διακρίβωσης.

Ολοκληρώνοντας, πραγματοποιείται μια σύγκριση μεταξύ των βέλτιστων μοντέλων που προέκυψαν για τα διάφορα *ασφαλή* σύνολα δεδομένων, χρησιμοποιώντας τα Τυχαία Δάση. Η σύγκριση αυτή, όπως για την περίπτωση των Δέντρων Αποφάσεων, γίνεται στη βάση των AUC που υπολογίζονται από την εφαρμογή κάθε ενός εκ των μοντέλων σε ένα σύνολο διακρίβωσης και απεικονίζεται στο Σχ. 5.12. Υπενθυμίζεται πως το σύνολο στο οποίο εφαρμόζεται κάθε μοντέλο είναι διαφορετικό και προέρχεται από την αντίστοιχη εκδοχή των *ασφαλών* δεδομένων, οπότε η σύγκριση δεν είναι απόλυτη αλλά ενδεικτική του αν κάποιο σύνολο, δηλαδή η λογική πίσω από αυτό, οδήγησε στην κατασκευή ενός δραματικά καλύτερου ή χειρότερου μοντέλου σε σχέση με τα υπόλοιπα.

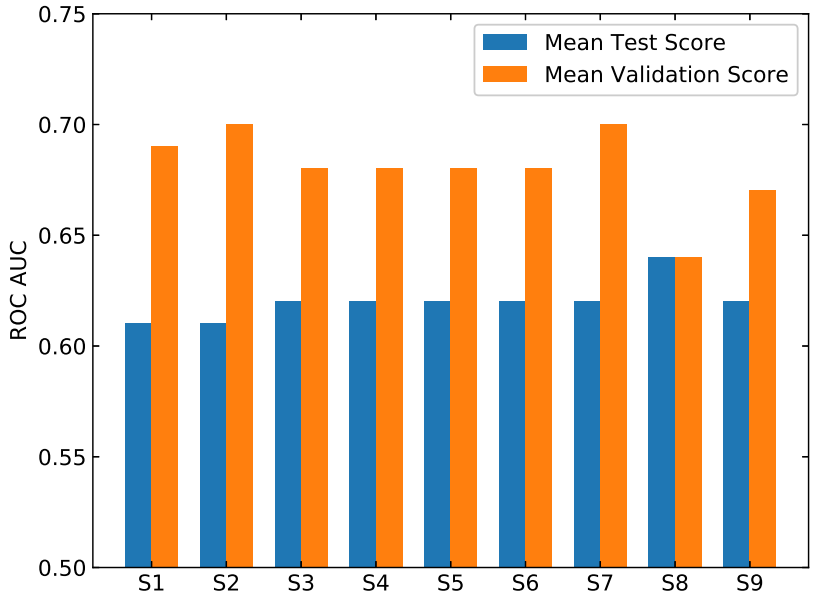
Από το Σχ. 5.12 φαίνεται ακριβώς ότι φάνηκε και για τα αποτελέσματα που παρουσιάστηκαν στο Σχ. 5.6. Συγκεκριμένα, παρατηρείται ότι τα AUC που υπολογίστηκαν επί των συνόλων ελέγου και επί των συνόλων διακρίβωσης είναι ανά κατηγορία κοντινά μεταξύ τους. Επίσης, απουσιάζει οποιαδήποτε κανονικότητα μεταβαίνοντας από το S1 στο S9. Αυτό επιβεβαιώνει πως τα αποτελέσματα είναι κοντινά, ανεξάρτητα από το αν αγνοούνται 2 ή 7 ημέρες μετά την αφαίρεση των μανιταριών. Αυτό σημαίνει πως οι τιμές των χαρακτηριστικών επανέρχονται στο να αντιστοιχούν σε απουσία



Σχήμα 5.11: Εξάρτηση του AUC που υπολογίστηκε για τα σύνολα εκπαίδευσης και ελέγχου των δεδομένων S2 από την περιπλοκότητα των μοντέλων που εξετάστηκαν, εκφρασμένης μέσω της τιμής της παραμέτρου max_depth. Λοιπές παράμετροι: `class_weight = balanced_subsample`, `max_features = log2`.

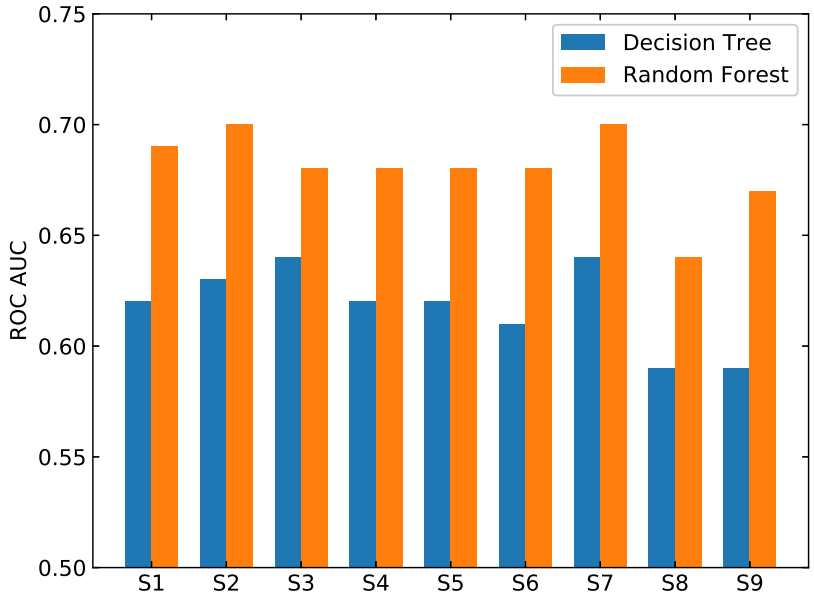
μανιταριού σύντομα μετά την αφαίρεση του, ενώ, όπως αναφέρεται και στην Εν. 3.4, η δημιουργία των συνόλων S1-S9 συνίσταται απλά σε μια υποστηριζόμενη από τη λογική υποδειγματοληψία της πιο κοινής τάξης, όχι όμως σε βαθμό που να χάνεται πληροφορία. Αν έπρεπε να επιλεγεί ένα σύνολο δεδομένων, αυτό θα ήταν μάλλον το S2, το οποίο παρουσιάζει την καλύτερη επίδοση επί του συνόλου διακρίβωσης, δηλαδή AUC $= 0.7$, ενώ ταυτόχρονα δεν προκύπτει από κάποια ακραία υπόθεση, όπως π.χ. το S7 που υποθέτει ότι απαιτούνται 7 μέρες για να επανέλθει η λεκάνη μετά την αφαίρεση του μανιταριού. Το αποτέλεσμα που προέκυψε για τα Δέντρα Αποφάσεων, ότι δηλαδή το S3 είναι το σύνολο για το οποίο προκύπτουν τα καλύτερα αποτελέσματα, δεν διαφέρει πολύ από το συμπέρασμα που βγαίνει και για τα Τυχαία Δάση.

Η διαφορά στις επιδόσεις, όταν τα μοντέλα εφαρμόζονται στα σύνολα ελέγχου της Αναζήτησης Πλέγματος και όταν εφαρμόζονται στα σύνολα διακρίβωσης είναι παρούσα και για τα Τυχαία Δάση. Για ακόμη μια φορά δεν υποδεικνύει υπερπροσαρμογή στα δεδομένα αλλά μάλλον το αντίθετο, δηλαδή καλύτερη συμπεριφορά στα άγνωστα δεδομένα. Σε κάθε περίπτωση, όμως, η μεθοδολογία που έχει ακολουθηθεί παρέχει την αυτοπεποίθηση ότι σε κανένα σημείο της υπολογιστικής διαδικασίας δεν έχει συμβεί διαρροή πληροφορίας για τα δεδομένα πάνω στα οποία αξιολογείται τελικά ο αλγόριθμος.



Σχήμα 5.12: Σύγκριση των AUC που προέκυψαν από την εφαρμογή των βέλτιστων μοντέλων για τα Τυχαία Δάση σε κάθε ένα από τα *ασφαλή* σύνολα δεδομένων, στα σύνολα ελέγχου και διακρίβωσης.

Σε αυτό το σημείο μπορεί πλέον να γίνει και η σύγκριση μεταξύ μοντέλων βασισμένων στα Δέντρα Αποφάσεων και στα Τυχαία Δάση, όταν αυτά εφαρμόζονται στα *ασφαλή* δεδομένα. Η σύγκριση αυτή γίνεται στη βάση των επιδόσεων επί των συνόλων διακρίβωσης και παρουσιάζεται στο Σχ. 5.13. Από το σχήμα φαίνεται πως για όλα τα σύνολα δεδομένων, η επίδοση του Τυχαίου Δάσους είναι σημαντικά καλύτερη. Αυτό είναι αναμενόμενο με βάση όσα έχουν ήδη ειπωθεί για τα δύο είδη μοντέλων στην Εν. 4.3, όπου εξηγήθηκε ότι ένα Τυχαίο Δάσος έχει μια πιο ευρεία οπτική των δεδομένων, με αποτέλεσμα να παρουσιάζει πάντα έστω και λίγο καλύτερη επίδοση από το αντίστοιχο Δέντρο Αποφάσεων. Τέλος, προκύπτει το συμπέρασμα ότι αν το μοντέλο που τελικά θα χρησιμοποιηθεί στην πράξη προέρχεται από τα *ασφαλή* σύνολα δεδομένων, αυτό θα πρέπει να είναι το μοντέλο που έχει βασιστεί σε Τυχαία Δάση, εκπαιδεύοντάς τα πάνω στο σύνολο S2. Στη συνέχεια εξετάζονται τα μοντέλα που προκύπτουν με βάση τα *μη ασφαλή* σύνολα δεδομένων.

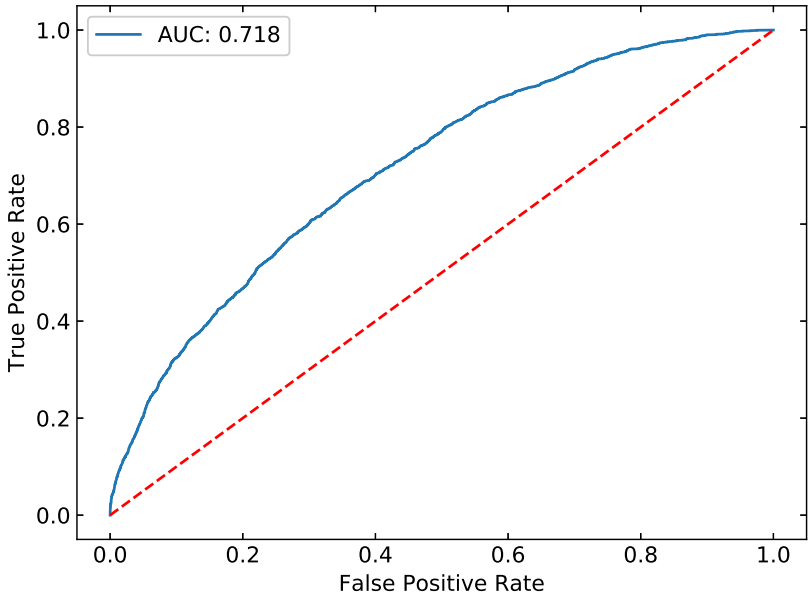


Σχήμα 5.13: Σύγκριση των AUC που προέκυψαν από την εφαρμογή των βέλτιστων μοντέλων για τα Δέντρα Αποφάσεων και τα Τυχαία Δάση, για κάθε ένα από τα *ασφαλή* σύνολα δεδομένων, επί των συνόλων διακρίβωσης.

5.2 Μη Ασφαλή Δεδομένα

Μετά τα αποτελέσματα που αφορούν την οικογένεια των *ασφαλών* συνόλων δεδομένων, σειρά έχουν αυτά που αφορούν στα *μη ασφαλή* δεδομένα. Και τα δεδομένα αυτά αντιμετωπίζονται με μοντέλα που βασίζονται τόσο σε Δέντρα Αποφάσεων, όσο και σε Τυχαία Δάση. Παρόλα αυτά, για λόγους συντομίας παρουσιάζονται μόνο τα μοντέλα που βασίζονται στα Τυχαία Δάση. Άλλωστε, οι επιδόσεις των μοντέλων αυτών είναι σημαντικά καλύτερες, όπως αποδείχθηκε για τα *ασφαλή* δεδομένα. Αρχικά παρουσιάζονται ενδεικτικά διαγράμματα, τα οποία συγκεκριμένα προέρχονται από τη διερεύνηση που έγινε για τα δεδομένα U3, δηλαδή αυτά στα οποία για κάθε λεκάνη έχουν μετατραπεί σε True οι ετικέτες 3 ημερών πριν τον εντοπισμό του μανιταριού, ενώ έχουν αφαιρεθεί οι παρατηρήσεις των υπόλοιπων ημερών μεχρι τις 15 ημέρες πριν, καθώς και 3 ημερών μετά τον εντοπισμό του μανιταριού.

Στο Σχ. 5.14 απεικονίζεται η καμπύλη ROC που αντιστοιχεί στην εφαρμογή του καλύτερου μοντέλου στα δεδομένα του συνόλου διακρίβωσης. Η τιμή του εμβαδού κάτω από αυτήν την καμπύλη, AUC = 0.72 θα μπορούσε να θεωρηθεί ικανοποιητική, ειδικά συγκρινόμενη με τη σχεδόν τυχαία πρόβλεψη, η οποία αποτελεί την τρέχουσα

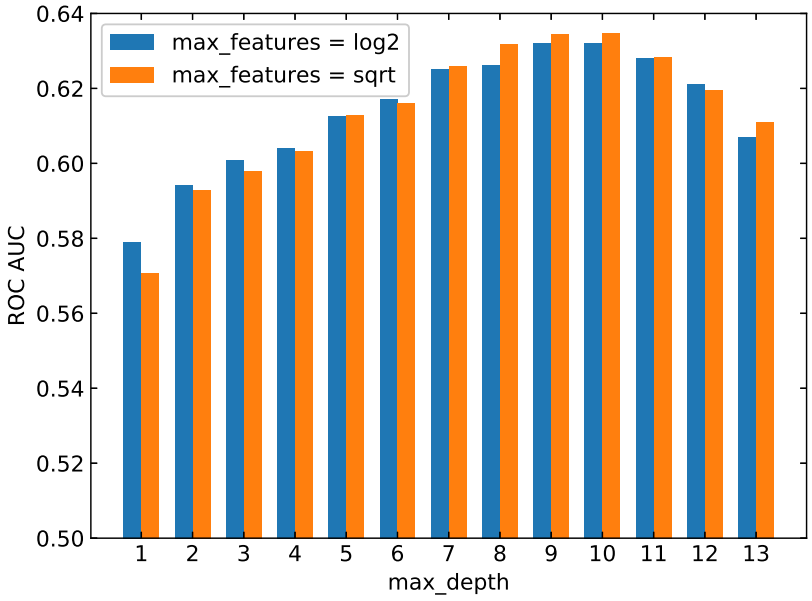


Σχήμα 5.14: Καμπύλη ROC του βέλτιστου Τυχαίου Δάσους που προέκυψε για τα δεδομένα U2, εφαρμοσμένου στο σύνολο διακρίβωσης. Τιμές κρίσιμων παραμέτρων: `max_depth` = 10, `class_weight = balanced_subsample`, `max_features = sqrt`.

αντιμετώπιση του προβλήματος. Επίσης, η τιμή αυτή είναι πολύ υψηλότερη του 0.70, που υπολογίστηκε ως η βέλτιστη για τα μοντέλα βασισμένα στα *ασφαλή* δεδομένα. Φυσικά, όλα τα παραπάνω έχουν νοήμα μόνο αν υπάρχει η εμπιστοσύνη ότι τα δεδομένα του συνόλου U3 ανταποκρίνονται στην πραγματικότητα, δηλαδή ότι για να παραχθούν δεν έχει γίνει κάποια ακραία υπόθεση.

Το Σχ. 5.15 μπορεί να αξιοποιηθεί για να οπτικοποιηθεί η εξάρτηση από την παράμετρο `max_depth` αφού η παράμετρος που επίσης παρουσιάζεται στο διάγραμμα, `max_features`, δεν επηρεάζει ιδιαίτερα την επίδοση των μοντέλων. Στο σχήμα φαίνεται ότι η εξάρτηση της επίδοσης από αυτή την παράμετρο διαφέρει λίγο από τις προηγούμενες περιπτώσεις. Πιο συγκεκριμένα, τα Τυχαία Δάση με μικρό πλήθος επιπέδων δε συμπεριφέρονται καλά, γεγονός που ενδεχομένως υποδεικνύει ότι υποπροσαρμόζονται στα δεδομένα. Αυτό είναι εν μέρει αναμενόμενο, αφού το σύνολο U3 περιέχει περισσότερη πληροφορία σε σχέση με τα σύνολα *ασφαλών* δεδομένων, αφού σε αυτό αρκετές ετικέτες έχουν μετατραπεί σε True, ενώ για τα *ασφαλή* δεδομένα είχαν αφαιρεθεί.

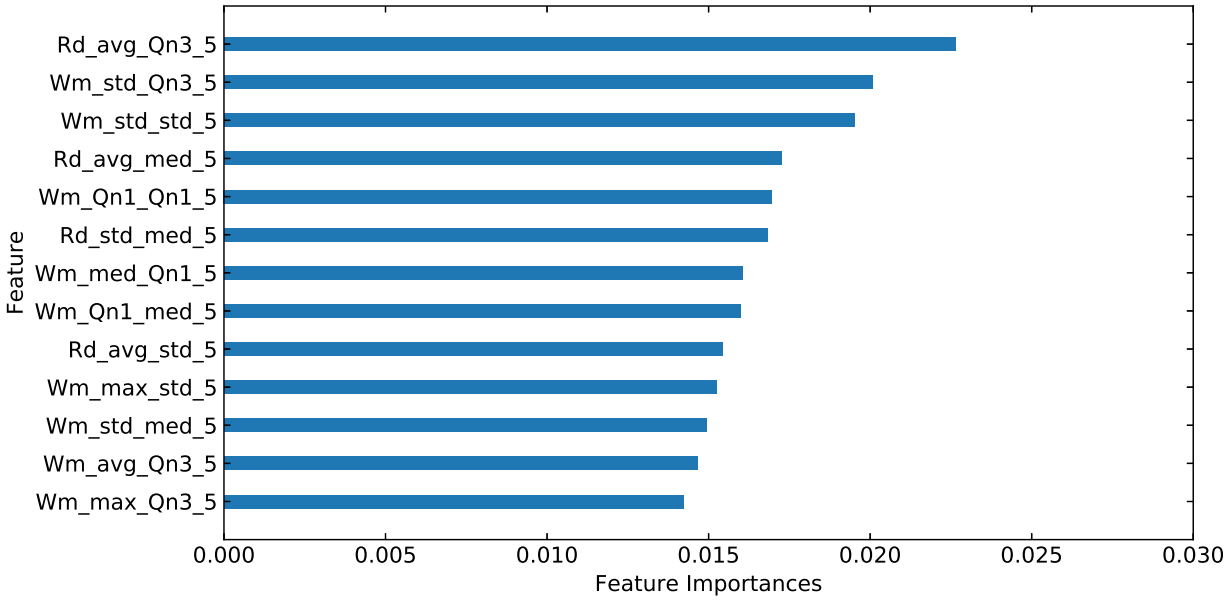
Ο μόνος τρόπος με τον οποίο μπορεί να κατανοηθεί ο τρόπος λειτουργίας του Τυχαίου Δάσους είναι μέσω του υπολογισμού των τιμών σημαντικότητας κάθε χαρακτηριστι-



Σχήμα 5.15: Εξάρτηση του AUC των μοντέλων που εξέταστηκαν κατά την Αναζήτηση Πλέγματος για τα δεδομένα U3 από τις παραμέτρους `max_depth` και `max_features`.

κού και η εύρεση των πιο σημαντικών χαρακτηριστικών. Οι τιμές σημαντικότητας για τα 13 πιο σημαντικά χαρακτηριστικά του βέλτιστου Τυχαίου Δάσους για τα δεδομένα U3 απεικονίζονται στο Σχ. 5.16. Τα αποτελέσματα στην περίπτωση αυτή είναι παρόμοια με εκείνα που προέκυψαν από την εφαρμογή των Τυχαίων Δασών στα *ασφαλή*. Η μέγιστη σημαντικότητα κάποιου χαρακτηριστικού είναι λίγο χαμηλότερη από 0.025, σε αντιδιαστολή με την τιμή 0.170 για το βέλτιστο Δέντρο Αποφάσεων. Αυτό υποδηλώνει πως το Τυχαίο Δάσος λαμβάνει υπόψη πολλά χαρακτηριστικά, με χαμηλότερη σημαντικότητα το καθένα, χωρίς δηλαδή να βασίζει τις προβλέψεις του εξ΄ ολοκλήρου σε μια μικρή ομάδα χαρακτηριστικών. Κανένα χαρακτηριστικό δεν παρουσιάζει μηδενική τιμή σημαντικότητας για το Τυχαίο Δάσος, οπότε όλα έχουν ληφθεί υπόψη, σε μικρότερο ή μεγαλύτερο βαθμό. Αυτό σημαίνει πως η τυχαιότητα στην κατασκευή των δέντρων του δάσους, η οποία πηγάζει τόσο από τα bootstrap samples, όσο και από την επιλογή κάθε δοκιμής από μικρά υποσύνολα του συνόλου των χαρακτηριστικών, οδηγεί σε μεγαλύτερη αξιοποίηση της διαθέσιμης πληροφορίας.

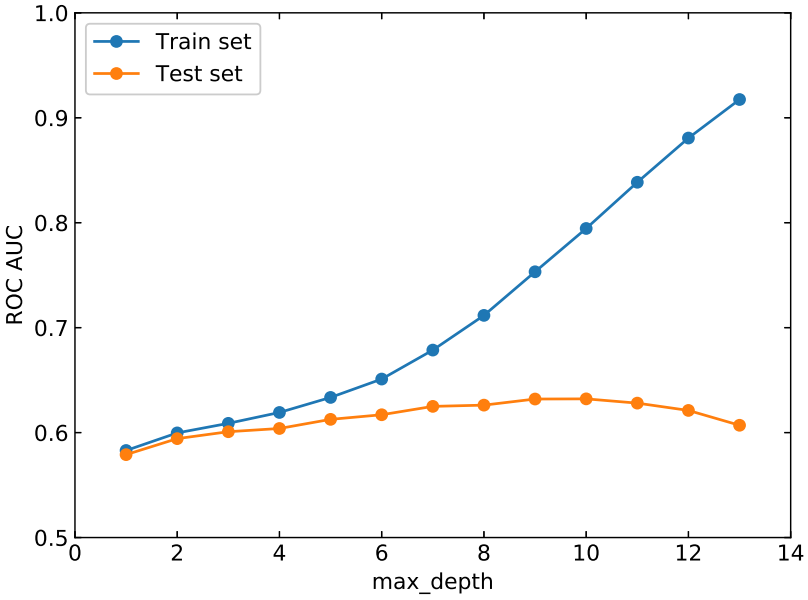
Το τελευταίο ενδεικτικό σχήμα που παρουσιάζεται για τα Τυχαία Δάση με βάση τα δεδομένα U3 είναι το Σχ. 5.17. Στο σχήμα αυτό είναι δυνατό να επιβεβαιωθούν εκ νέου όλα όσα έχουν αναφερθεί περί γενίκευσης και προσαρμογής των μοντέλων στα δεδομένα, ανάλογα με το βαθμό περιπλοκότητας τους. Η τιμή του AUC αυξάνεται καθώς το μέγιστο βάθος του δέντρου αυξάνεται από το 1 έως το 10, που



Σχήμα 5.16: Υπολογισμένες τιμές της σημαντικότητας των διάφορων χαρακτηριστικών για το βέλτιστο Τυχαίο Δάσος που προέκυψε για τα δεδομένα U3.

είναι και η οριακή τιμή, ενώ περαιτέρω αύξηση της παραμέτρου αυτής οδηγεί σε υποβάθμιση της επίδοσης του μοντέλου λόγω υπερπροσαρμογής. Καθώς τα δέντρα του Τυχαίου Δάσους χρησιμοποιούν περισσότερες δοκιμές για να πραγματοποιήσουν τις προβλέψεις τους, γίνονται δηλαδή πιο σύνθετα, τα μοντέλα που προκύπτουν *απομνημονεύουν* το σύνολο εκπαίδευσης, οπότε αδυνατούν να προβλέψουν σωστά όταν αντιμετωπίζουν τα άγνωστα δεδομένα του συνόλου διακρίβωσης.

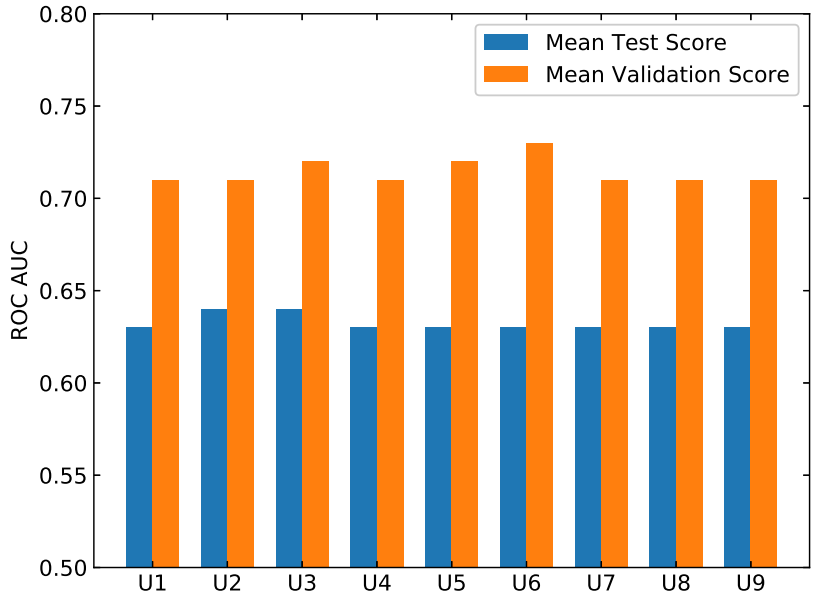
Ολοκληρώνοντας, πραγματοποιείται μια σύγκριση μεταξύ των βέλτιστων μοντέλων που προέκυψαν για τα διάφορα *μη ασφαλή* σύνολα δεδομένων, χρησιμοποιώντας τα Τυχαία Δάση. Η σύγκριση αυτή, όπως και στις προηγούμενες περιπτώσεις, γίνεται στη βάση των AUC που υπολογίζονται από την εφαρμογή κάθε ενός εκ των μοντέλων σε ένα σύνολο διακρίβωσης και απεικονίζεται στο Σχ. 5.18. Υπενθυμίζεται πως το σύνολο στο οποίο εφαρμόζεται κάθε μοντέλο είναι διαφορετικό και προέρχεται από την αντίστοιχη εκδοχή των *ασφαλών* δεδομένων, οπότε η σύγκριση δεν είναι απόλυτη αλλά ενδεικτική του αν κάποιο σύνολο, δηλαδή η λογική πίσω από αυτό, οδήγησε στην κατασκευή ενός δραματικά καλύτερου ή χειρότερου μοντέλου σε σχέση με τα υπόλοιπα.



Σχήμα 5.17: Εξάρτηση του AUC που υπολογίστηκε για τα σύνολα εκπαίδευσης και ελέγχου των δεδομένων U3 από την περιπλοκότητα των μοντέλων που εξετάστηκαν, εκφρασμένης μέσω της τιμής της παραμέτρου max_depth. Λοιπές παράμετροι: `class_weight = balanced_subsample`, `max_features = log2`.

Από το Σχ. 5.18 φαίνεται ακριβώς ότι φάνηκε και για τα αποτελέσματα που παρουσιάστηκαν στο Σχ. 5.12. Συγκεκριμένα, παρατηρείται ότι τα AUC που υπολογίστηκαν επί των συνόλων ελέγου και επί των συνόλων διακρίβωσης είναι ανά κατηγορία κοντινά μεταξύ τους. Επίσης, απουσιάζει οποιαδήποτε κανονικότητα μεταβαίνοντας από το U1 στο U9. Αυτό επιβεβαιώνει πως τα αποτελέσματα είναι κοντινά, ανεξάρτητα από το αν αγνοούνται 3 ή 5 ημέρες μετά την αφαίρεση των μανιταριών. Αυτό σημαίνει πως οι τιμές των χαρακτηριστικών ίσως αντιστοιχούν σε παρουσία μανιταριού σύντομα μετά την εμφάνισή του η οποία, σύμφωνα με την εμπειρία των μηχανικών της Δραστηριότητας, μπορεί να συμβαίνει έως και 10 μέρες πριν τον εντοπισμό του. Όπως αναφέρεται και στην Εν. 3.4, η δημιουργία των συνόλων U1-U9 συνίσταται σε μια υποστηριζόμενη από τη λογική υπερδειγματοληψία της πιο σπάνιας τάξης. Εκ των πραγμάτων όμως αποδεικνύεται πως οι πρόσθετες παρατηρήσεις που σημειώνονται ως True δεν προσφέρουν σημαντικά περισσότερη πληροφορία σε σχέση με εκείνες που ήταν σημειωμένες με True εξαρχής.

Η διαφορά στις επιδόσεις, όταν τα μοντέλα εφαρμόζονται στα σύνολα ελέγχου της Αναζήτησης Πλέγματος και όταν εφαρμόζονται στα σύνολα διακρίβωσης είναι παρούσα και για τα Τυχαία Δάση που αφορούν τα *μη ασφαλή* δεδομένα. Για ακόμη μια φορά δεν υποδεικνύει υπερπροσαρμογή στα δεδομένα αλλά μάλλον το αντίθετο,



Σχήμα 5.18: Σύγκριση των AUC που προέκυψαν από την εφαρμογή των βέλτιστων μοντέλων για τα Τυχαία Δάση σε κάθε ένα από τα *μη ασφαλή* σύνολα δεδομένων, στα σύνολα ελέγχου και διακρίβωσης.

δηλαδή καλύτερη συμπεριφορά στα άγνωστα δεδομένα. Σε κάθε περίπτωση, όμως, η μεθοδολογία που έχει ακολουθηθεί παρέχει τη διαβεβαίωση ότι σε κανένα σημείο της υπολογιστικής διαδικασίας δεν έχει συμβεί διαρροή πληροφορίας για τα δεδομένα πάνω στα οποία αξιολογείται τελικά ο αλγόριθμος.

5.3 Επιλογή του Καταλληλότερου Μοντέλου

Τέλος, πραγματοποιείται μια σύγκριση των βέλτιστων επιδόσεων για τα Τυχαία Δάση που βασίζονται στα *ασφαλή* και για αυτά που βασίζονται στα *μη ασφαλή* δεδομένα. Φαίνεται πως ένα μικρό προβάδισμα έχει το βέλτιστο μοντέλο για τα δεδομένα U3, με AUC = 0.72, έναντι του 0.7 που προέκυψε για τα δεδομένα S2. Φυσικά, πρέπει να αξιολογηθεί το κατά πόσο οι πρόσθετες παραδοχές που συνεπάγεται η χρήση των δεδομένων U3 είναι καταρχήν αποδεκτές, καθώς και κατά πόσο αξίζει να γίνουν για μια μεταβολή επίδοσης σαν αυτή που επιτυγχάνεται. Αυτό πρέπει να κριθεί από τους μηχανικούς της Δραστηριότητας Ηλεκτρόλυσης.

Στο έργο των μηχανικών που καλούνται να αποφασίσουν για την τύχη των μοντέλων, καθώς και να αντιληφθούν το νόημα πίσω από μια επίδοση της τάξης του 0.7 ή του 0.72 μπορεί να βοηθήσει ένα εργαλείο αξιολόγησης που περιγράφηκε στην Εν. 4.4, ο Πίνακας Σύγχυσης. Το εργαλείο αυτό, αν και δε συνοψίζει το αποτέλεσμα της εφαρμογής ενός μοντέλου σε έναν αριθμό, ώστε να μπορεί να χρησιμοποιηθεί σε μια διαδικασία επιλογής του καλύτερου μοντέλου όπως η Αναζήτηση Πλέγματος με cross-validation, δίνει μια πολύ απτή εικόνα της πρακτικής του εφαρμογής. Πριν από την ανάγνωση των παρακάτω, ο αναγνώστης παραπέμπεται στο Σχ. 4.3 ώστε να θυμηθεί το περιεχόμενο του κάθε στοιχείου του Πίνακα Σύγχυσης.

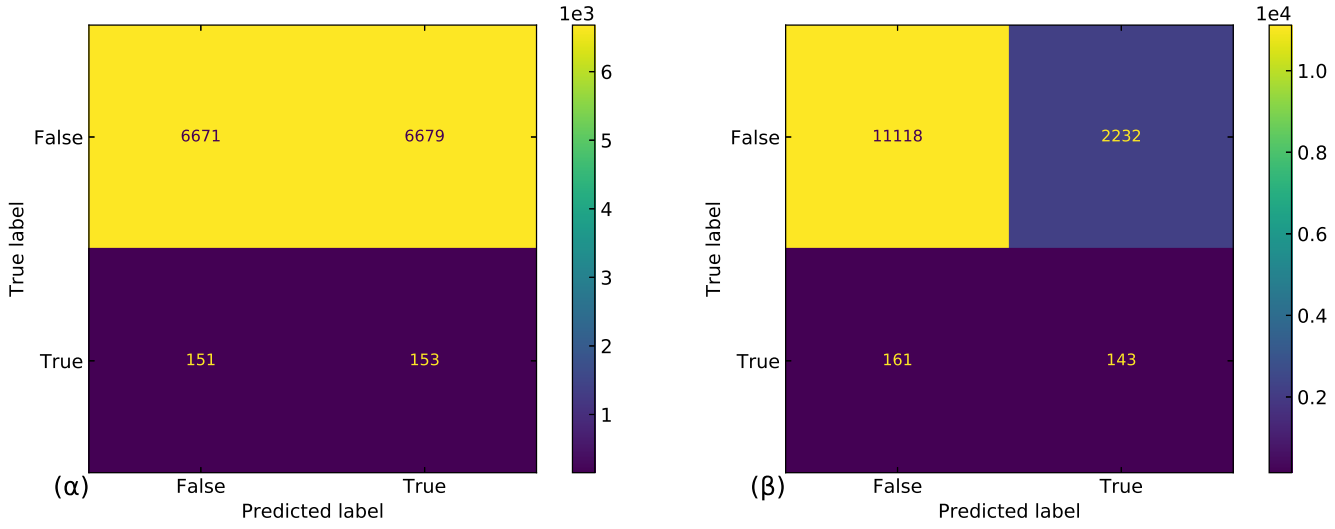
Αρχικά παρουσιάζονται οι Πίνακες Σύγχυσης που προκύπτουν από τη χρήση του Ψευδοταξινομητή και του βέλτιστου Τυχαίου Δάσους για τα *ασφαλή* δεδομένα S2. Οι πίνακες αυτοί φαίνονται στο Σχ. 5.19, αριστερά και δεξία, αντίστοιχα. Στο εσωτερικό κάθε τετραγώνου αναγράφεται το πλήθος των παρατηρήσεων που εντάσσονται σε κάθε μία από τις τέσσερις κατηγορίες, TN, FP, FN και TP, ενώ το άθροισμα όλων των παρατηρήσεων αποτελεί το μέγεθος του συνόλου διακρίβωσης, που για την περίπτωση αυτή ισούται με 13350 παρατηρήσεις. Επίσης, για λόγους σύγκρισης παρέχεται μία χρωματική κλίμακα ανά πίνακα. Υπενθυμίζεται ότι το AUC στην περίπτωση αυτή ισούται με 0.7.

Όσον αφορά τις παρατηρήσεις που αποτελούν όντως μανιτάρια, έχουν δηλαδή την έγκυρη ετικέτα True, ο Ψευδοταξινομητής και το Τυχαίο Δάσος δε φαίνεται να διαφέρουν ιδιαίτερα. Συγκεκριμένα, το Τυχαίο Δάσος υποπίπτει σε 10 περισσότερα

ψευδή αρνητικά, με αποτέλεσμα να χάσει 10 μανιτάρια, να έχει δηλαδή 10 λιγότερα αληθή θετικά. Το αποτέλεσμα αυτό είναι μάλλον απογοητευτικό, αν αναλογιστεί κανείς πως ο Ψευδοταξινομητής επιλέγει τυχαία σε ποιες παρατηρήσεις θα αποδώσει την ετικέτα του μανιταριού, εντελώς στην τύχη, δηλαδή με πιθανότητα 50-50. Όπως όμως είχε διευκρινιστεί και στην Εν. 4.4, το να περιοριστεί κανείς σε αυτή τη σύγκριση του παρέχει μια εικόνα που απέχει δραματικά από την πραγματικότητα.

Μια ματιά στα αποτελέσματα που αφορούν στις παρατηρήσεις με πραγματική ετικέτα False, δηλαδή σε αυτές που αντιστοιχούν σε απουσία μανιταριού, αρκεί για να κατανοηθεί το πως ο Ψευδοταξινομητής πετυχαίνει αυτή την επίδοση. Μα απλούστατα προβλέποντας με πιθανότητα 50-50, με συνέπεια όμως να έχει θεωρήσει ως θετικές και τις μισές παρατηρήσεις που ανήκουν στην αρνητική τάξη. Αντίθετα, ο αληθινός Ταξινομητής που έχει βασιστεί στα Τυχαία Δάση ταξινόμησε λανθασμένα μόλις το 17% των αρνητικών παρατηρήσεων, ταυτόχρονα πετυχαίνοντας ένα παραπλήσιο αποτέλεσμα όσον αφορά τις θετικές παρατηρήσεις. Πραγματοποιώντας την αντιστοίχιση με το πρακτικό πρόβλημα που αντιμετωπίζεται, το μοντέλο που έχει επιλεγεί ως το βέλτιστο για τα δεδομένα S2 βοήθησε στον εντοπισμό του ίδιου περίπου ποσοστού μανιταριών, στέλνοντας όμως τους εργαζομένους για έλεγχο μόλις το 17% των φορών σε σχέση με τους τυχαίους, δειγματοληπτικούς ελέγχους.

Ανάλογες παρατηρήσεις προκύπτουν και για το Σχ. 5.20, το οποίο αφορά στα α-

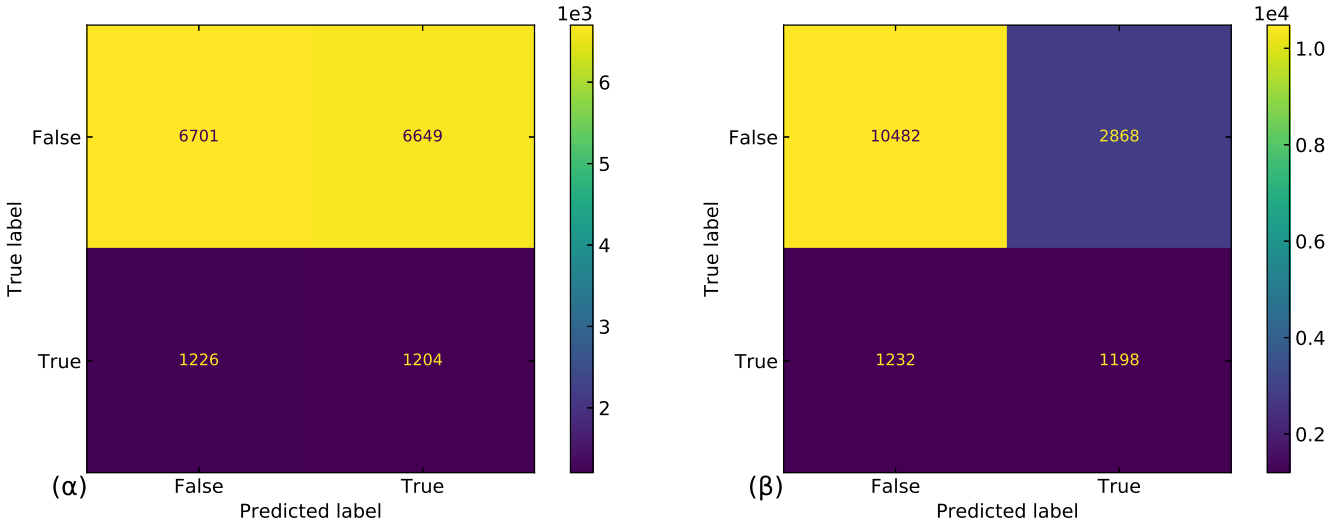


Σχήμα 5.19: Πίνακες Σύγχυσης που αξιολογούν την εφαρμογή (α) του Ψευδοταξινομητή και (β) του βέλτιστου μοντέλου βασισμένου στα Τυχαία Δάση για το σύνολο *ασφαλών* δεδομένων S2.

ντίστοιχα μοντέλα για το σύνολο *μη ασφαλών* δεδομένων U3. Ο Ψευδοταξινομητής και το Τυχαίο Δάσος έχουν παρόμοια επίδοση στην πρόβλεψη των παρατηρήσεων με πραγματική ετικέτα True, η διαφορά τους όμως στα ψευδή θετικά που πρέπει να υποπέσουν για να πετύχουν τις επιδόσεις αυτές είναι πολύ σημαντική. Συγκεκριμένα, το Τυχαίο Δάσος ταξινόμησε λανθασμένα μόλις το 21% των αρνητικών παρατηρήσεων, έναντι (προφανώς) του 50% που ταξινόμησε λανθασμένα ο Ψευδοταξινομητής.

Υπενθυμίζεται ότι ο Πίνακας Σύγχυσης αντιστοιχεί στην περίπτωση που ο αλγόριθμος χαρακτηρίζει ως θετική μια παρατήρηση για την οποία είναι πάνω από 50% σίγουρος ότι αντιστοιχεί σε μανιτάρι. Αντίθετα, η τιμή AUC παρουσιάζει μια πιο σφαιρική εικόνα για όλες τις τιμές αυτής της οριακής τιμής πιθανότητας. Η τιμή λειτουργίας επιλέγεται από τους μηχανικούς της Δραστηριότητας, με βάση το τι επιδιώκουν: να εντοπίζονται όλα τα μανιτάρια με αυξημένο πλήθος λεκανών που ελέγχονται, ή το πλήθος αυτό να είναι το επιτρεπτό από τις συνθήκες και να βρίσκονται κατά το δυνατόν περισσότερα μανιτάρια.

Ας δοθεί όμως ένα παράδειγμα για όριο ίσο με 50%. Αν, κατά τη διάρκεια ενός πόστου, από τις 780 λεκάνες υπάρχει μανιτάρι στις 10, ο βέλτιστος Ταξινομητής βασισμένος στα Τυχαία Δάση θα πετύχει την εξής επίδοση: θα εντοπίσει τα 5 μανιτάρια, στέλνοντας τους εργαζόμενους για έναν έλεγχο σε 156 λεκάνες. Το υπάρχον σύστημα, το οποίο στην ουσία συνίσταται σε δειγματοληπτικούς ελέγχους, αφού η ευ-



Σχήμα 5.20: Πίνακες Σύγχυσης που αξιολογούν την εφαρμογή (α) του Ψευδοταξινομητή και (β) του βέλτιστου μοντέλου βασισμένου στα Τυχαία Δάση για το σύνολο *μη ασφαλών* δεδομένων U3.

στοχία του εξετάστηκε και βρέθηκε ότι δε διαφέρει από εκείνη της τυχαίας επιλογής, θα απαιτούσε τον έλεγχο των μισών λεκανών, δηλαδή 390 λεκανων. Η επίδοση αυτή ίσως δε μοιάζει τόσο εντυπωσιακή, είναι αναντίρρητο όμως το γεγονός ότι αποτελεί σημαντική βελτίωση έναντι της τρέχουσας κατάστασης.

Κεφάλαιο 6

Συμπεράσματα

Σε αυτό το κεφάλαιο συνοψίζονται τα συμπεράσματα της παρούσας μελέτης. Υπενθυμίζεται πως το πρόβλημα που ζητήθηκε να επιλυθεί είναι η πρόβλεψη της ύπαρξης μανιταριών στις λεκάνες ηλεκτρόλυσης του εργοστασίου του Αλουμινίου της Ελλάδος στον Άγιο Νικόλαο Βοιωτίας. Η προσέγγιση που ακολουθήθηκε ήταν βασισμένη στη Μηχανική Μάθηση και επί της ουσίας αξιοποιούσε αλγορίθμους Δυαδικής Ταξινόμησης, αφού το ζητούμενο είναι να προβλεφθεί αποκλειστικά η παρουσία ή μη των μανιταριών στις λεκάνες, ώστε αυτά να αφαιρεθούν εγκαίρως από τους εργαζόμενους.

Το πρώτο βήμα της στρατηγικής που επελέγη ήταν η αναγωγή των δεδομένων χρονοσειράς που χαρακτηρίζουν την κατάσταση του συστήματος, δηλαδή των εν δυνάμει χαρακτηριστικών, από το λεπτό, που είναι η αρχική συχνότητας καταγραφής τους, στη βάρδια ή πόστο. Η αναγωγή αυτή πραγματοποιήθηκε λαμβάνοντας 8 στατιστικά μεγέθη σε επίπεδο πόστου για κάθε ένα από τα 3 εξεταζόμενα αρχικά μεγέθη, καθώς και 4 επιπλέον στατιστικά μεγέθη ανά παράγωγο μέγεθος για τις τελευταίες 5 ημέρες, με σκοπό να αποτυπωθεί με κάποιο τρόπο και η χρονική εξέλιξη της κατάστασης του συστήματος. Έτσι παρήχθησαν συνολικά 120 χαρακτηριστικά.

Το δεύτερο βήμα της στρατηγικής ήταν η επεξεργασία των ιστορικών δεδομένων για την παρουσία μανιταριών, δηλαδή των ετικετών, με δύο διαφορετικούς τρόπους. Ο πρώτος, ο οποίος οδήγησε στην παραγωγή των λεγόμενων *ασφαλών* συνόλων δεδομένων, συνίστατο στην αφαίρεση παρατηρήσεων που αντιστοιχούν σε 15 μέρες πριν από τον εντοπισμό του μανιταριού και 1 έως 9 μέρες μετά την αφαίρεση του. Έτσι παρήχθησαν τα *ασφαλή* σύνολα δεδομένων S1-S9. Ο δεύτερος, που με τη σειρά του οδήγησε στην παραγωγή των λεγόμενων *μη ασφαλών* συνόλων δεδομένων, συνίστατο στην τροποποίηση του συνόλου S2, το οποίο διαπιστώθηκε ότι παρουσίαζε τα καλύτερα αποτελέσματα από τα σύνολα S1-S9 (όπως αναλύεται και παρακάτω), επαναφέροντας τις παρατηρήσεις 1 έως 9 ημερών πριν από τον εντοπισμό του μανιταριού

και σημειώνοντας τες με την ετικέτα True, που αντιστοιχεί σε παρουσία μανιταριού. Έτσι παρήχθησαν τα *μη ασφαλή* σύνολα δεδομένων U1-U9.

Αξίζει να σημειωθεί εκ νέου ότι από κατασκευής τα σύνολα S1-S9 απλά αφορούν σε μία υποδειγματοληψία της πιο κοινής τάξης, η οποία είναι η παρουσία μανιταριού, χωρίς όμως να βασίζονται σε κάποια ακραία υπόθεση η οποία μπορεί να μην ισχύει. Αντίθετα, τα σύνολα U1-U9 βασίζονται σε μία υπόθεση η οποία όμως κρίνεται ως ρεαλιστική, συγκεκριμένα ότι το μανιτάρι μπορεί να προϋπήρχε έως και 9 μέρες του εντοπισμού του. Η υπόθεση αυτή υποδεικνύεται από την εμπειρία των ειδικών της διεργασίας και σε καμία περίπτωση δεν έγινε αυθαίρετα, πραγματοποιήθηκε δε στην προσπάθεια έμμεσης υπερδειγματοληψίας της σπάνιας τάξης, αυτής της απουσίας μανιταριού, αποσκοπώντας σε καλύτερα αποτελέσματα.

Όλα τα σύνολα δεδομένων που προαναφέρθηκαν εξετάστηκαν με την ίδια ακριβώς μεθοδολογία. Συγκεκριμένα, για κάθε ένα από τα σύνολα πραγματοποιήθηκε μια Αναζήτηση Πλέγματος ανά οικόγενεια αλγορίθμων που χρησιμοποιήθηκαν. Η Αναζήτηση Πλέγματος έγκειται στην εφαρμογή των αλγορίθμων για διαφορετικούς συνδυασμούς των υπό εξέταση παραμέτρων τους, την αξιολόγηση των αποτελεσμάτων με βάση ένα προεπιλεγμένο δείκτη και την εύρεση του βέλτιστου συνδυασμού παραμέτρων. Ο λόγος που επιλέγεται αυτή η τεχνική είναι επειδή είναι κατασκευασμένη με τρόπο τέτοιο ώστε να μην υπάρχει διαρροή πληροφορίας από τα δεδομένα που χρησιμοποιούνται για την τελική αξιολόγηση στον αλγόριθμο. Έτσι οι επιδόσεις που υπολογίζονται είναι έγκυρες.

Πιο συγκεκριμένα, οι οικογένειες αλγορίθμων για τις οποίες παρουσιάζονται αποτελέσματα είναι τα Δέντρα Αποφάσεων και τα Τυχαία Δάση. Οι παράμετροι που εξετάστηκαν είναι για τα μεν Δέντρα Αποφάσεων το πλήθος των επιπέδων τους και ο τρόπους υπολογισμού της βαρύτητας των τάξεων, ενώ για τα Τυχαία Δάση εξετάστηκε επιπλέον το μέγιστο πλήθος τυχαίων χαρακτηριστικών μεταξύ των οποίων επιλέγεται το βέλτιστο σε κάθε διαχωρισμό των δεδομένων. Ο δείκτης αξιολόγησης που χρησιμοποιήθηκε στις Αναζητήσεις Πλέγματος ήταν το AUC, δηλαδή το εμβαδό κάτω από την καμπύλη ROC.

Μετά την εύρεση του βέλτιστου μοντέλου για κάθε σύνολο δεδομένων, μεταξύ των S1-S9 και U1-U9, βρέθηκε τόσο το σύνολο που δίνει τα καλύτερα αποτελέσματα ανά οικογένεια αλγορίθμων, όσο αυτό που δίνει γενικότερα τα καλύτερα αποτελέσματα. Προφανώς, έγινε σαφές και το ποια οικογένεια αλγορίθμων πετυχαίνει τις κατά μέσο όρο μέγιστες επιδόσεις, καθώς και ο συνδυασμός των παραμέτρων που οδηγεί στα επιθυμητά αποτελέσματα.

Η καλύτερη επίδοση που πέτυχε κάποιο μοντέλο είναι AUC = 0.72. Το μοντέλο

αυτό προέκυψε για τα δεδομένα του συνόλου U3, από ένα Τυχαίο Δάσος με τιμές κρίσιμων παραμέτρων `max_depth` $= 10$, `class_weight = balanced_subsample` και `max_features = sqrt`. Το γεγονός ότι προέκυψε από αυτό το σύνολο και όχι από κάποιο με πιο ακραίες υποθέσεις, δηλαδή με περισσότερες παρατηρήσεις σημειωμένες ως True, σημαίνει πως η υπόθεση αυτή μπορεί να βελτιώσει την επίδοση, πάντα όμως εντός ορίων. Όσον αφορά την ίδια την επίδοση, το 0.72 δεν είναι ιδιαίτερα υψηλό, απέχει όμως πολύ από το 0.5 που αντιστοιχεί σε τυχαίες προβλέψεις. Είναι δηλαδή ικανό να βελτιώσει κατά πολύ την υπάρχουσα κατάσταση όσον αφορά την πρόβλεψη της ύπαρξης μανιταριών.

Ο τρόπος με τον οποίο αυτό μπορεί να πραγματοποιηθεί είναι ο παρακάτω. Το μοντέλο με επίδοση AUC $= 0.72$ τίθεται σε εφαρμογή για τον εντοπισμό των μανιταριών στο εργοστάσιο. Σταδιακά εντοπίζονται περισσότερα μανιτάρια πιο έγκαιρα από ότι θα εντοπίζονταν με τον υπάρχοντα τρόπο εντοπισμού, οπότε οι ετικέτες, δηλαδή η παρουσία ή μη μανιταριού για ένα συγκεκριμένο σετ συνθηκών, περιέχουν όλο και λιγότερες ανακρίβειες με την πάροδο του χρόνου. Σε τακτά διαστήματα πραγματοποιείται επανεκπαίδευση του μοντέλου με βάση τα νέα, πιο ακριβή δεδομένα, έως ότου η επίδοση του νέου μοντέλο σταθεροποιηθεί σε ένα υψηλότερο επίπεδο από το τρέχον 0.72. Όσον αφορά την πρακτική σημασία της τρέχουσας επίδοσης για τους ελέγχους, το μοντέλο που αναπτύχθηκε στο πλαίσιο της μελέτης μπορεί να βρει το ίδιο πλήθος μανιταριών που θα έβρισκαν οι δειγματοληπτικοί έλεγχου, απαιτώντας όμως μόλις το 20% των αρχικών ελέγχων.

Περνώντας στα γενικότερα συμπεράσματα της μελέτης, καταρχάς επιβεβαιώθηκε ότι τα μοντέλα που βασίζονται στα Τυχαία Δάση προφέρουν καλύτερη γενίκευση σε σχέση με εκείνα που βασίζονται στις δομικές τους μονάδες, τα Δέντρα Αποφάσεων. Όσον αφορά τις παραμέτρους, έγινε εμφανές ότι η επιλογή να αποδοθούν στις παρατηρήσεις των δύο τάξεων, της παρουσίας και της απουσίας μανιταριού, βαρύτητες που σχετίζονται με τη συχνότητα των τάξεων αυτών στο δείγμα, δηλαδή η επιλογή `class_weight = balanced`, ήταν απολύτως απαραίτητη. Για την ακρίβεια, τα μοντέλα που προέκυψαν χωρίς αυτήν την επιλογή πρακτικά πραγματοποιούσαν τυχαίες προβλέψεις. Ο λόγος για τον οποίο ισχύουν τα παραπάνω είναι ότι μόνο αν ληφθεί υπόψη η συχνότητα των τάξεων στο δείγμα μπορεί να αποφευχθεί η επίδραση των Ανισοκατανεμημένων Τάξεων στα δεδομένα. Η επίδραση αυτή είναι ένα από τα κύρια προβλήματα που αντιμετωπίστηκαν στη μελέτη και κατέστησε προτιμότερη την επιλογή και των συγκεκριμένων τύπων αλγορίθμων Μηχανικής Μάθησης.

Η δεύτερη παράμετρος που εξετάστηκε, δηλαδή το πλήθος των επιπέδων των δέντρων, τόσο στην περίπτωση των μεμονομένων Δέντρων Αποφάσεων όσο και σε εκείνη των Τυχαίων Δασών, είχε την αναμενόμενη συμπεριφορά. Υπήρχε δηλαδή ένα ενδιάμεσο πλήθος επιπέδων που οδηγούσε στη βέλτιστη επίδοση. Το πλήθος αυτό παρατηρήθη-

κε ότι ήταν γενικά μικρότερο στην περίπτωση των Τυχαίων Δασών, αφού, λόγω του ότι αποτελούνται από περισσότερα του ενός Δέντρα Αποφάσεων, τα οποία μάλιστα διαφέρουν μεταξύ τους, λαμβάνουν υπόψη τους μεγαλύτερο μέρος της πληροφορίας και γενικεύουν πολύ καλύτερα.

Η τελευταία παράμετρος, η οποία εξετάστηκε μόνο για τα Τυχαία Δάση, ήταν το μέγιστο πλήθος τυχαίων χαρακτηριστικών μεταξύ των οποίων τυχαία επιλέγεται η δοκιμή, δηλαδή η ερώτηση, με την οποία διαχωρίζονται οι κλάδοι σε κάθε επίπεδο των επιμέρους Δέντρων Αποφάσεων. Δεν παρατηρήθηκε κάποια ουσιαστική διαφορά μεταξύ των δύο επιλογών που εξετάστηκαν, δηλαδή `max_features = sqrt` και `log2`, οι οποίες υπενθυμίζεται ότι είχαν ως αποτέλεσμα την τυχαία επιλογή μεταξύ περίπου 7 και 11 χαρακτηριστικών, αντίστοιχα. Εκτός από τις επιλογές αυτές, για τις οποίες τα αποτελέσματα παρουσιάζονται στο παρόν κείμενο, εξετάστηκαν και άλλες τιμές της παραμέτρου, με ανάλογα αποτελέσματα.

Το συμπέρασμα που προκύπτει από τη χρήση διαφορετικών συνόλων δεδομένων, συγκεκριμένα εκείνων των ασφαλών (S1-S9) και των μη ασφαλών (U1-U9) δεδομένων, είναι ότι η λογική διαχείρισης των δεδομένων ετικετών είναι πολύ σημαντική. Στη γενικότερη αντιμετώπιση των μεθόδων Μηχανικής Μάθησης κάτι τέτοιο δεν ισχύει, αφού θεωρητικά υπάρχει μια βεβαιότητα για τα δεδομένα με βάση τα οποία εκπαιδεύεται ένα μοντέλο. Αποδεικνύεται όμως πως στην πράξη μπορεί να προκύψει αβεβαιότητα για την αλήθεια των δεδομένων, τόσο των χαρακτηριστικών όσο και των ετικετών. Ειδικά στην περίπτωση βιομηχανικών εφαρμογών, όπου τα δεδομένα συλλέγονται από αισθητήρες, οι οποίοι ενδέχεται να παρουσιάζουν κάποια βλάβη ή να μην είναι βαθμονομημένοι, καθώς και με μη αυτοματοποιημένο τρόπο, οι ανακρίβειες στα δεδομένα είναι μάλλον ο κανόνας παρά η εξαίρεση. Για το λόγο αυτό απαιτείται πολύ προσεκτική προεπεξεργασία των δεδομένων, καθώς και έξυπνες προσεγγίσεις στην αξιοποίησής τους. Αυτό επιχειρήθηκε να γίνει και στην παρούσα μελέτη κατά την αντιμετώπιση των χρονοσειρών από τις οποίες προέκυψαν τα χαρακτηριστικά, καθώς και των δεδομένων ετικετών.

Τα αποτελέσματα της παρούσας μελέτης κρίνονται ως πολύ ελπιδοφόρα για το μέλλον. Ακόμα και με τις εγγενείς πρακτικές δυσκολίες που συνδέονταν με τη φύση των δεδομένων, όπως π.χ. διαφορετικές συχνότητες καταγραφής, ανάγκη για αξιοποίηση παρελθοντικής πληροφορίας, ανακρίβειες και ελλείψεις στα δεδομένα και ανισοκατανομή των τάξεων, επετεύχθη μια πολύ σημαντική βελτίωση έναντι της υφιστάμενης κατάστασης. Πόσο μάλλον δε όταν αυτό πραγματοποιήθηκε από ένα άτομο, σε έναν προσωπικό υπολογιστή ικανοποιητικών επιδόσεων και σε ένα διάστημα περίπου 6 μηνών (πλήρους εργασίας). Είναι πολύ πιθανό αντίστοιχες προσεγγίσεις στη βιομηχανία να μπορούν να δώσουν λύσεις με σημαντικά πλεονεκτήματα έναντι των παραδοσιακών προσεγγίσεων που αξιοποιούνται μέχρι σήμερα.

Βιβλιογραφία

[1] Bard, Allen J., György Inzelt, and Fritz Scholz (editors): *Electrochemical Dictionary: 2nd, Revised and Extended Edition.* Springer, New York, 2012.

[2] Müller, Andreas C. and Sarah Guido: *Introduction to Machine Learning with Python.* O'Reilly, Boston, 2016.

[3] Van Rossum, Guido and Fred L. Drake: *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA, 2009.

[4] Python Software Foundation: *The Python Programming Language.* `https://www.python.org/`, June 2020.

[5] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay: *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12:2825–2830, 2011.

[6] scikit-learn Authors: *scikit-learn: Machine Learning in Python.* `https://scikit-learn.org/`, June 2020.

[7] NumPy Authors: *NumPy: The fundamental package for scientific computing with Python.* `https://numpy.org/`, June 2020.

[8] The pandas Development Team: *pandas: A fast, powerful, flexible and easy to use open source data analysis and manipulation tool.* `https://pandas.pydata.org/`, June 2020.

[9] Hunter, J. D.: *Matplotlib: A 2d graphics environment.* Computing in Science & Engineering, 9(3):90–95, 2007.

[10] Matplotlib Authors: *Matplotlib: Visualization with Python.* `https://matplotlib.org/`, June 2020.

[11] Michael Waskom: *seaborn: statistical data visualization*. `https://seaborn.pydata.org/`, June 2020.

[12] Jupyter Authors: *The Jupyter Project*. `https://jupyter.org/`, June 2020.

[13] Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas: *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning*. Journal of Machine Learning Research, 18(17):1–5, 2017.

[14] imbalanced-learn Authors: *Toolbox for Imbalanced Datasets in Machine Learning*. `https://imbalanced-learn.readthedocs.io/en/stable/`, June 2020.

[15] Rashka, Sebastian (editor): *Python Machine Learning*. Packt, Birmingham, 3rd edition, 2019.

[16] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (editors): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

[17] Breiman, L., J. Friedman, R. Olshen, and C. Stone (editors): *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

[18] Breiman, L.: *Arcing classifiers*. Annals of Statistics, 26:801–849, 1998.

[19] Breiman, L.: *Random forests*. Machine Learning, 45:5–32, 2001.