ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
Τμήμα Μηχανικών Βιομηχανικής Σχεδίασης & Παραγωγής

Τεχνητή Νοημοσύνη στη Ναυτιλία

Προεπεξεργασία και επιλογή δεδομένων στην ναυτιλία μέσω της μηχανικής μάθησης για την πρόβλεψη της ταχύτητας ενός πλοίου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

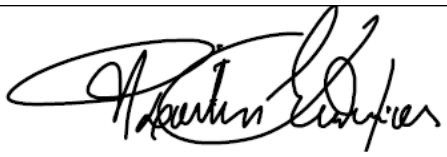
του

Αθανάσιου Α. Αρβανιτίδη

Επιβλέπουσα: Ελένη-Αικατερίνη Λελίγκου

Αθήνα, Ιούλιος 2021

ΜΕΛΗ ΕΞΕΤΑΣΤΙΚΗΣ ΕΠΙΤΡΟΠΗΣ

Επιβλέπουσα Καθηγήτρια: Ελένη – Αικατερίνη Λελίγκου	 Eleni Aikaterini Leligkou Digitally signed by Eleni Aikaterini Leligkou Date: 2021.08.03 12:16:01 +03'00'
Επ. Καθηγήτρια: Παρασκευή Ζαχαρία	
Συνεπιβλέπων- Αναπληρωτής Καθηγητής ΣΝΔ: Ευθύμιος Παριώτης	

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η **ΑΡΒΑΝΙΤΙΔΗΣ ΑΘΑΝΑΣΙΟΣ** του **ΓΕΩΡΓΙΟΥ**, με αριθμό μητρώου **71446168** φοιτητής/τρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής **Μηχανικών** του Τμήματος **Βιομηχανικής Σχεδίασης και Παραγωγής**, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών/ούσα

Αρβανιτίδης Αθανάσιος

Ευχαριστίες

Θέλω να ευχαριστήσω την οικογένεια μου για όλη την στήριξη τους κατά την διάρκεια των σπουδών μου, τον φίλο μου Γιάννη για την πολύτιμη βοήθεια του όλα αυτά τα χρόνια και την υπεύθυνη καθηγήτρια μου Κα. Λελίγκου Ελένη-Αικατερίνη για την βοήθεια, τις ευκαιρίες και τις γνώσεις που μου παρείχε μέχρι και την στιγμή της εκπόνησης αυτής της εργασίας.

Περίληψη

Σε αυτή τη διπλωματική εργασία μελετάται η εφαρμογή μηχανισμών τεχνητής νοημοσύνης στην μοντελοποίηση της λειτουργίας πλοίων με κύριο θέμα την προ-επεξεργασία των δεδομένων. Θα μελετηθούν διαφορετικές τεχνικές προ-επεξεργασίας και θα αξιολογηθούν με βάση την ακρίβεια των πρόβλεψης στην οποία οδηγούν με βάση τις απαιτήσεις που έχουν οριστεί (χρόνο σε σχέση με υπολογιστικούς πόρους).

Με την χρήση των κατάλληλων εργαλείων λογισμικού, πραγματοποιείται η αξιολόγηση των δεδομένων, η διαμόρφωση αυτών με σκοπό την πρόβλεψη της ταχύτητας του πλοίου στο νερό.

Λέξεις Κλειδιά

Μηχανική μάθηση, Δεδομένα, Features , Μοντέλο, Έξοδος

Abstract

In this paper the application of artificial intelligence mechanisms in the modeling of ship operation is studied with the main theme of data pre-processing. Different pre-processing techniques will be studied and evaluated based on the accuracy of the prediction they lead to, based on the requirements set (time relative to computational resources).

Using the appropriate software tools, the data is evaluated and configured to predict the ship's velocity in the water.

Keywords

Machine learning, Data, Features, Model, Output

Περιεχόμενα

Μέρος Α: Θεωρητικό υπόβαθρο

Κεφάλαιο 1: Εισαγωγή

1.1 Αντικείμενο διπλωματικής εργασίας

Αντικείμενο της παρούσας εργασίας, είναι το ευρύτερο πεδίο της εξόρυξης δεδομένων (Data Mining) με τα οποία μπορούμε να αποκτήσουμε μια πιο λεπτομερή εικόνα για την εφαρμογή που έχουμε να αντιμετωπίσουμε. Πιο συγκεκριμένα, ασχολείται με τα βήματα που πραγματοποιούνται για την προεπεξεργασία των δεδομένων που συλλέχτηκαν (Data preprocessing) και με την δημιουργία ενός μοντέλου μηχανικής μάθησης (Machine Learning), ώστε η διαδικασίες αυτές να γίνονται αυτόματα και αξιόπιστα.

1.2 Σκοπός διπλωματικής εργασίας

Στόχος αυτής της διπλωματικής εργασίας είναι, η παρουσίαση τεχνικών και μεθόδων που εφαρμόζονται για την προεπεξεργασία των δεδομένων καθώς και στην δημιουργία ενός μοντέλου κατάλληλου για την προετοιμασία αυτών. Αυτά γίνονται με σκοπό:

- Τον «καθαρισμό» και τον έλεγχο εγκυρότητας των δεδομένων
- Την καλύτερη επιλογή των δεδομένων, ανάλογα με το αποτέλεσμα που επιθυμούμε
- Την εκπαίδευση του κατάλληλου μοντέλου μηχανικής μάθησης ώστε να γίνεται η βέλτιστη επιλογή της ταχύτητας του πλοίου.

1.3 Μεθοδολογία

Για να υλοποιηθούν όλα τα παραπάνω, αρχικά ανακτήθηκε το απαιτούμενο σετ δεδομένων (dataset). Ακολούθησε, η μελέτη των δεδομένων αυτών ώστε να γίνει σωστά η διαλογή των «παραμέτρων» (features) των δεδομένων, που θα δοθούν στο μοντέλο ως είσοδοι. Τέλος, καθορίστηκε η έξοδος του μοντέλου, ο τύπος, η αρχική ρύθμιση και η εκπαίδευση αυτού με επιπλέον δεδομένα, με ανάλογη πλατφόρμα λογισμικού.

1.4 Δομή διπλωματικής εργασίας

Στο πρώτο μέρος της εργασίας, το οποίο αποτελεί και το θεωρητικό υπόβαθρο (κεφάλαια 1 έως 4), πραγματοποιείται μια επεξήγηση των αναφερόμενων εννοιών και των μεθόδων που έχουν ως τελικό αποτέλεσμα τον τελικό μας στόχο. Επίσης, αναλύονται και όροι του γενικότερου επιστημονικού πεδίου καθώς και τα συστήματα που χρησιμοποιούνται.

Στο δεύτερο μέρος (κεφάλαιο 5 έως 7), γίνεται η παρουσίαση και η αποσαφήνιση του πρακτικού τομέα της εργασίας. Δηλαδή, εξηγείται το κάθε βήμα που πραγματοποιήθηκε για την επεξεργασία των δεδομένων μέχρι να φτάσουμε στο τελικό αποτέλεσμα και να καταλήξουμε στα τελικά συμπεράσματα ενώ, συγχρόνως αναφέρονται και ενδεχόμενες επεκτάσεις της εφαρμογής αυτής.

Κεφάλαιο 2: Δεδομένα – Εξόρυξη δεδομένων (Data – Data mining)

2.1 Ορισμοί

Με τον όρο **δεδομένα (data)**, αναφερόμαστε στο σύνολο χαρακτηριστικών ή πληροφοριών, συνήθως υπό την μορφή αριθμών, που έχουν συλλεχθεί μέσω παρατήρησης. Πιο συγκεκριμένα, δεδομένα είναι ένα σύνολο τιμών οι οποίες μπορεί να είναι είτε ποσοτικές, είτε ποιοτικές και αφορούν ένα άτομο ή αντικείμενο.

Dataset, είναι ένα σύνολο (σετ) από δεδομένα.

Database (βάση δεδομένων), είναι μια οργανωμένη δομή δεδομένων αποθηκευμένη σε έναν ηλεκτρονικό υπολογιστή.

Η **εξόρυξη δεδομένων (data mining)**, είναι η διαδικασία με την οποία ακατέργαστα δεδομένα επεξεργάζονται με σκοπό την αποκόμιση πληροφοριών. Αυτό, επιτυγχάνεται αναζητώντας συσχετισμούς μέσα σε μεγάλους «όγκους» δεδομένων με την χρήση αλγοριθμικών και στατιστικών μοντέλων. Επίσης, πρέπει να σημειωθεί ότι η εξόρυξη δεδομένων εντάσσεται στο ευρύτερο πεδίο της επιστήμης των δεδομένων (data science).

2.1 Το data mining και η εξέλιξη της πληροφορίας

Το data mining, μπορεί να θεωρηθεί ως ένα μέρος της συνολικής εξέλιξης που έχει υποστεί η τεχνολογία της πληροφορίας. Η πληροφορία αντλείται, από την ανάλυση των δεδομένων που έχουν συλλεχθεί. Όμως, από τότε που ξεκίνησε η αποθήκευση δεδομένων και οι δημιουργία βάσεων δεδομένων (databases) μέχρι τώρα, ο τεράστιος όγκος των δεδομένων έχει καταστήσει πολύ δύσκολη την κατανόηση τους από τον άνθρωπο χωρίς την χρήση δυνατών εργαλείων. Συνεπώς, πολλές αποφάσεις παίρνονται έχοντας ως μοναδικό κριτήριο την αντίληψη εκείνου που αποφασίζει κι αυτό διότι, δεν έχει ό,τι χρειάζεται για να αποκτήσει την πολύτιμη γνώση από όλα τα δεδομένα που υπάρχουν. Σε αυτό το σημείο λοιπόν, το κενό ανάμεσα στα δεδομένα και την πληροφορία, έρχεται να συμπληρώσει το data mining που μπορεί να μετατρέψει αυτά τα «βουνά» δεδομένων σε πολύτιμη γνώση.

2.2 Η διαδικασία του data mining

1. **Καθορισμός του προβλήματος (State the problem):** Τα περισσότερα μοντέλα που υπάρχουν, χρησιμεύουν σε συγκεκριμένους γνωστικούς τομείς. Επομένως, πρέπει το πρόβλημα να καθοριστεί με ακρίβεια, για να γίνει η σωστή επιλογή του μοντέλου και των μεταβλητών που θα χρησιμοποιηθούν σε αυτό και να υπάρξει το επιθυμητό αποτέλεσμα.
2. **Συλλογή δεδομένων (Data collection):** Τα δεδομένα που θα χρησιμοποιηθούν μπορεί είτε να είναι «ελεγχόμενα» όταν ο ειδικός επιβλέπει την διαδικασία, είτε να είναι «τυχαία» καθώς ο ειδικός το μόνο που κάνει είναι να «παρατηρεί». Η διαδικασία της συλλογής δεδομένων είναι καθοριστική για το την εξόρυξη των δεδομένων αφού, αν αυτός που θα φτιάξει το μοντέλο γνωρίζει εξ' αρχής την κατανομή των δεδομένων, η δημιουργία του μοντέλου θα είναι πολύ ευκολότερη.
3. **Προετοιμασία των δεδομένων (Data preprocessing):** Τα περισσότερα δεδομένα που συλλέγονται δεν είναι «ελεγχόμενα». Προκειμένου, να εξασφαλιστεί η σωστή λειτουργία του μοντέλου πρέπει τα δεδομένα να ελεγχθούν για τυχόν παράταιρες σχετικά με την εφαρμογή τιμές, να διαμορφωθούν στην κατάλληλη μορφή και να γίνει η σωστή επιλογή μεταβλητών που θα χρησιμοποιηθούν για την εφαρμογή.
4. **Επιλογή μοντέλου (Model estimation):** Η επιλογή του κατάλληλου μοντέλου είναι, ίσως, το πιο σημαντικό μέρος της διαδικασίας αφού το μοντέλο θα καθορίσει το τελικό αποτέλεσμα στον μεγαλύτερο βαθμό.
5. **Δημιουργία του μοντέλου (Model interpretation):** Αφού υπάρχει η κεντρική ιδέα γύρω από το μοντέλο θα πρέπει μετά, αυτή η ιδέα να γίνει πράξη. Το πρόβλημα εδώ, είναι ότι όσο μεγαλύτερη ακρίβεια είναι η επιθυμητή ακρίβεια, τόσο πιο πολύπλοκο θα είναι το μοντέλο.
6. **Συμπεράσματα (Conclusions):** Αφού το μοντέλο που ολοκληρώσει τον σκοπό του θα πρέπει να εξαχθούν συμπεράσματα, για να γίνει, τελικά, η λήψη των αποφάσεων.

2.3 Η σπουδαιότητα του data mining

Το data mining χρησιμοποιεί διάφορους τύπους λογισμικού και εργαλεία ανάλυσης, ενώ η διαδικασία μπορεί να γίνει αυτόματα ή χειροκίνητα. Τι είναι αυτό που κάνει όμως το data mining τόσο χρήσιμο και σημαντικό;

- **Συσχέτιση των δεδομένων (Data association):** Όταν αντικρίζει κανείς ένα μεγάλο σετ δεδομένων (dataset), αρχικά δεν θα είναι εύκολο να καταλάβει τι μπορεί να σημαίνουν όλα αυτά τα στοιχεία. Το data mining, μπορεί να ανακαλύψει τις σχέσεις που έχουν όλες οι καταχωρήσεις μεταξύ τους και να ξεκαθαρίσει την κατάσταση για τον χρήστη, βοηθώντας τον να καταλάβει καλύτερα τα δεδομένα που έχει να διαχειριστεί.
- **Προβλέψεις (Predictions):** Ένα εργαλείο του data mining, που χρησιμοποιείται όλο και περισσότερο είναι τα μοντέλα πρόβλεψης. Με την χρήση αυτών μπορεί να βρεθεί η μελλοντική τιμή μιας μεταβλητής (feature) του dataset και να γίνει η κατηγοριοποίηση της έχοντας ως βάση τις προηγούμενες. Πάντα βέβαια με την μέγιστη δυνατή αποτελεσματικότητα και ακρίβεια.
- **Κατηγοριοποίηση (Clustering):** Με το εργαλείο της κατηγοριοποίησης, είναι δυνατή η ομαδοποίηση των δεδομένων ανάλογα με τα χαρακτηριστικά τους, δίνοντας έτσι μια πιο καθαρή και κατανοητή μορφή στο dataset.
- **Διαδοχικές σχέσεις (Sequential relationships):** Πολλές από τις μεταβλητές που υπάρχουν σε ένα dataset μπορεί να επηρεάζουν η μια την άλλη ταυτόχρονα ή η μια να είναι επακόλουθη της άλλης, όταν μιλάμε για χρονικά μεταβαλλόμενα γεγονότα. Το data mining μπορεί λοιπόν να ανακαλύψει τέτοιου είδους σχέσεις χρησιμοποιώντας τα κατάλληλα μοντέλα.

2.4 Εφαρμογές

Μια κοινή παραδοχή σήμερα είναι ότι, η εποχή μας χαρακτηρίζεται από το την σημαντικότητα της πληροφορίας. Έτσι λοιπόν, η εξόρυξη, η επεξεργασία δεδομένων και η μετατροπή αυτών σε χρήσιμες πληροφορίες γίνονται όλο και πιο απαραίτητα για τις επιχειρήσεις και την οικονομία, ώστε να μπορούν να εξελίσσονται. Μερικοί από τους κλάδους των επιχειρήσεων που εξαρτώνται σε μεγάλο βαθμό από την εξόρυξη δεδομένων είναι:

- **Μάρκετινγκ (Marketing):** Όλες οι επιχειρήσεις προκειμένου να μπορούν να ανταπεξέρχονται στις αλλαγές της αγοράς που προκύπτουν από τους καταναλωτές, προσπαθούν να αποκτούν όσο πιο πολλά δεδομένα για αυτούς, με σκοπό την ικανοποίηση τους και το υψηλότερο κέρδος για τις ίδιες.

- **Βιομηχανία (Industrial):** Οι γραμμές παραγωγής, μαζί με το αγαθό που παράγεται προσφέρουν και μεγάλο όγκο δεδομένων σχετικά με τον εξοπλισμό (π.χ. βλάβες, χρήσιμες πληροφορίες λειτουργίας). Επομένως απαιτείται η συλλογή και η επεξεργασία αυτών, για την επίτευξη του βέλτιστου αποτελέσματος.
- **Οικονομικά συστήματα (Financial systems):** Κάθε τραπεζικό σύστημα, αποτελείται πλέον από τεράστιους όγκους δεδομένων που προκύπτουν με ταχύτατους ρυθμούς. Η εξόρυξη δεδομένων μπορεί να συνεισφέρει έτσι ώστε οι ειδικοί να μπορούν να ενεργήσουν για το καλύτερο των πελατών και της οικονομίας.
- **Εντοπισμός απάτης (Fraud detection):** Η ψηφιακή εποχή χαρακτηρίζεται από την εκμετάλλευση των χρηστών. Μέσω του data mining, είναι πιο εύκολο και λιγότερο χρονοβόρο αναλύοντας δεδομένα, να εντοπίζεται μια απάτη.
- **Ναυτιλία (Maritime):** Στην ναυτιλία, τα δεδομένα που συλλέγονται μπορούν να δώσουν μια καλύτερη εικόνα για την μηχανική κατάσταση του πλοίου αλλά και το πώς μπορεί να γίνει η μεταφορά των αγαθών με μικρότερο κόστος και μεγαλύτερο κέρδος.

2.5 Προκλήσεις

Σε κάθε τεχνολογικό κλάδο υπάρχει πάντα η ανάγκη και επιθυμία της εξέλιξης, ειδικότερα όταν αυτός μπορεί να διευκολύνει και να εξυπηρετήσει και άλλα επιστημονικά πεδία. Αυτό ακριβώς συμβαίνει και στο πεδίο του data science και κατ' επέκταση στο data mining, καθώς οι απαιτήσεις προς αυτά γίνονται όλο και μεγαλύτερες. Πιο συγκεκριμένα κάποιες από τις προκλήσεις που καλείται να ξεπεράσει η διαδικασία της εξόρυξης δεδομένων είναι:

- Η **ταχύτητα** επεξεργασίας των δεδομένων από τα μοντέλα να είναι τέτοια ώστε να δίνουν το επιθυμητό αποτέλεσμα όσο το δυνατόν γρηγορότερα.
- Τα δεδομένα που προκύπτουν σήμερα, μπορεί να είναι όλο και πιο ασαφή και οι σχέσεις μεταξύ τους να είναι πιο **πολύπλοκες**. Επομένως, τα μοντέλα που θα χρησιμοποιηθούν πρέπει να μπορούν να ανταπεξέλθουν στις νέες καταστάσεις.
- Όπως αναφέρθηκε και παραπάνω τα πεδία που εφαρμόζεται το data mining είναι πολλά. Άρα υπάρχουν και πολλά μοντέλα που εξυπηρετούν τον ίδιο σκοπό αλλά με κάποιες διαφοροποιήσεις, που όμως μπορεί να προκαλέσουν σύγχυση στον χρήστη, διότι δεν θα

είναι σίγουρος για το πιο μοντέλο να διαλέξει για την εφαρμογή του. Έτσι, προκύπτει το πρόβλημα της **συμβατότητας** μοντέλων-εφαρμογών.

- Καθώς αυξάνονται οι απαιτήσεις των χρηστών, πρέπει να αυξάνεται και η **αποτελεσματικότητα-εγκυρότητα** των μοντέλων, εφόσον θα οι χρήστες θα βασίζονται σε ένα μοντέλο που θα καθορίζει τις ενέργειες τους.

Κεφάλαιο 3: Προετοιμασία των δεδομένων (Data Preparation)

3.1 Τι είναι η προετοιμασία δεδομένων (data preparation);

Σήμερα, οι βάσεις δεδομένων στην πλειοψηφία τους περιέχουν αλλοιωμένα, ελλιπή και ασυνεχή δεδομένα λόγω του γεγονότος ότι οι βάσεις αυτές στηρίζονται σε διάφορες πηγές και περιέχουν μεγάλο όγκο δεδομένων. Είναι εύκολο να καταλάβει κανείς, ότι χαμηλής ποιότητας δεδομένα θα οδηγήσουν σε ένα χαμηλής ακρίβειας αποτέλεσμα, ακόμα και αν το μοντέλο επεξεργασίας των δεδομένων είναι το καλύτερο δυνατόν. Η προετοιμασία των δεδομένων (data preparation ή data preprocessing), είναι το μέρος της συνολικής διαδικασίας του data mining που έχει ως επίκεντρο την λύση των παραπάνω προβλημάτων, μέσα από ένα σύνολο επιμέρους λειτουργιών που εν τέλει μας αποφέρουν ένα πιο ποιοτικό σύνολο δεδομένων.

3.2 Τι είναι η ποιότητα και γιατί είναι τόσο σημαντική;

Τα δεδομένα χαρακτηρίζονται ποιοτικά, όταν πληρούν συγκεκριμένα κριτήρια ανάλογα με την εφαρμογή που συσχετίζονται. Υπάρχουν αρκετοί παράγοντες που προσδιορίζουν την ποιότητα των δεδομένων. Πιο συγκεκριμένα είναι η ακρίβεια (accuracy), η πληρότητα(completeness), η συνέχεια (consistency), η πιστότητα (believability), η χρονική ακολουθία (timeliness) και το πόσο εύκολα μπορούν να ερμηνευθούν (interpretability).

Ατελή δεδομένα μπορεί να προκύψουν για διάφορους λόγους και παρατηρείται όλο και περισσότερο. Μπορεί να λείπουν δεδομένα επειδή δεν θεωρήθηκαν σημαντικά, να υπάρχει ανακρίβεια διότι οι χρήστες εσκεμμένα καταχώρησαν λάθος δεδομένα ή λόγω τεχνολογικών περιορισμών να μην έγινε σωστά η μεταφορά των δεδομένων ενώ, όταν τα δεδομένα δεν έχουν χρονική συνέχεια δημιουργείται σύγχυση όσον αφορά κατανόηση τους μειώνοντας έτσι την συνολική ποιότητα τους. Επιπλέον, όταν μια βάση δεδομένων αντιμετωπίζει προβλήματα, οι χρήστες θα αμφισβητήσουν την πιστότητα της και αν τα δεδομένα έχουν ερμηνευτεί σωστά, ακόμη και αν αυτά επιλυθούν αργότερα, δεν θα είναι σίγουροι για την ποιότητα των δεδομένων. Συλλογιζόμενοι, τα παραπάνω κριτήρια αποδεικνύεται ότι το data preprocessing έχει έναν πολύ σπουδαίο ρόλο, για την απόκτηση γνώσης μέσω του data mining.

3.3 «Μέσα» στο data preprocessing

Έχοντας, περιγράψει τον στόχο και την σπουδαιότητα του data preprocessing, θα αναλυθούν αναλυτικότερα οι διαδικασίες που γίνονται κατά το data preprocessing ώστε, τα δεδομένα να γίνονται πιο ποιοτικά. Ονομαστικά οι διαδικασίες αυτές είναι: **ο καθαρισμός δεδομένων (data cleaning)**, **η ενσωμάτωση δεδομένων (data integration)**, **η μείωση δεδομένων (data reduction)** και **η μετασχηματισμός δεδομένων (data transformation)**.

3.3.1 Καθαρισμός δεδομένων (Data Cleaning)

Όπως προαναφέρθηκε, τα πραγματικά δεδομένα τείνουν να είναι ατελή, ασυνεχή και αλλοιωμένα. Ο καθαρισμός των δεδομένων (data cleaning), έρχεται για να καλύψει τις τιμές που λείπουν, να διορθώσει ότι δεν είναι σχετικό με τα δεδομένα και να φέρει τα αλλοιωμένα δεδομένα σε μια καλύτερη κατάσταση.

Όσον αφορά τις **τιμές που λείπουν** από το dataset μπορούν να γίνουν τα εξής:

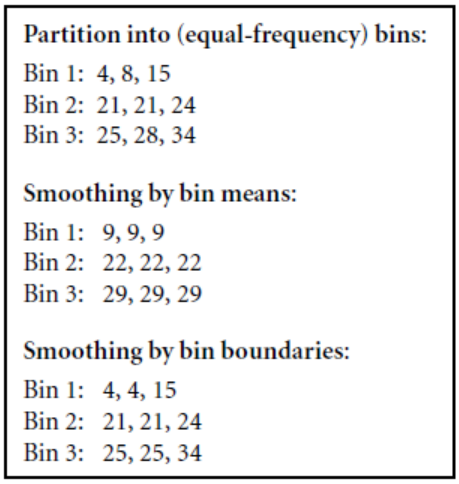
- **Παράλειψη της σειράς:** Αυτή η μέθοδος δεν είναι ιδιαίτερα αποτελεσματική εκτός αν η σειρά που παραλείπεται έχει αρκετές κατηγορίες από τις οποίες λείπουν τιμές. Επίσης, δεν αποδίδει καλά στην περίπτωση όπου η κάθε κατηγορία έχει διαφορετικό σύνολο τιμών που λείπουν. Αγνοώντας την σειρά, μπορεί να μην συμπεριληφθούν δεδομένα που μπορεί να είναι χρήσιμα για την εφαρμογή που πρέπει να υλοποιηθεί.
- **Συμπλήρωση των κενών χειροκίνητα:** Αυτή η μέθοδος είναι χρονοβόρα και ανέφικτη όταν πρόκειται για μεγάλο όγκο δεδομένων.
- **Χρήση μιας γενικής σταθεράς για την κάλυψη των κενών:** Αντικαθιστώντας όλα τα κενά με μια γενική σταθερά, μπορεί το πρόγραμμα που θα εξορύξει τα δεδομένα, να θεωρήσει ότι όλα τα στοιχεία με την σταθερά αποτελούν κάτι σημαντικό, αφού έχουν μια κοινή μεταβλητή. Επομένως, αν και απλή η μέθοδος αυτή δεν είναι αξιόπιστη.
- **Συμπλήρωση των κενών με την μέση ή την μεσαία τιμή (mean or median):** Σε ένα dataset μπορούμε να βρούμε την μέση τιμή και τα κενά να συμπληρωθούν με αυτή τα κενά (όταν πρόκειται για αριθμητικό περιεχόμενο) ή να βρούμε την μεσαία τιμή σε κάθε άλλη περίπτωση.

- **Χρήση της παραμετρικής μέσης ή μεσαίας τιμής για όλες τις καταχωρήσεις που ανήκουν στην ίδια κατηγορία, όπως η επιλεγμένη σειρά:** Εάν, θέλουμε να κατηγοριοποιήσουμε τα δεδομένα ανάλογα με ένα κριτήριο, μπορούμε να αντικαταστήσουμε με την μέση τιμή, τα κενά που ανήκουν στην ίδια κατηγορία με τα κενά της σειράς που έχει επιλεχθεί. Αν τα δεδομένα είναι αλλοιωμένα, μπορεί να χρησιμοποιηθεί η μεσαία τιμή.
- **Χρήση της πιο πιθανής τιμής:** Μπορεί να χρησιμοποιηθεί κάποιος αλγόριθμος πρόβλεψης, οποίος θα μας υποδεικνύει την καλύτερη τιμή με βάση τα χαρακτηριστικά (attributes ή features) του dataset.

Εκτός από κενά στα δεδομένα, ένα dataset μπορεί να έχει και δεδομένα με θόρυβο. Στον χώρο των δεδομένων, **θόρυβος (noise)** είναι ένα τυχαίο λάθος ή κάποια απόκλιση σε μία μεταβλητή. Πώς γίνεται να «καθαριστεί» ή να μειωθεί ο θόρυβος;

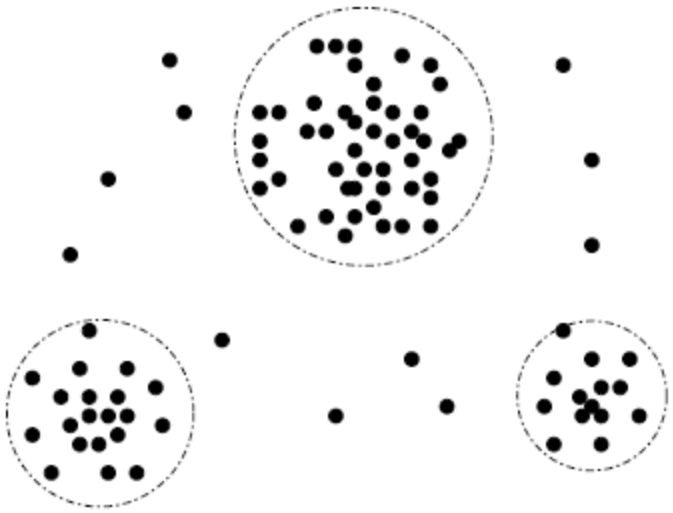
- **Δεδομένα σε καλάθια (Binning):** Οι μέθοδοι αυτοί, διορθώνουν μια ξεχωριστή τιμή από τα δεδομένα με βάση τις «γειτονικές» τιμές, γύρω από αυτή κι επειδή οι μέθοδοι που ακολουθούν αυτόν τον τρόπο λειτουργίας, χρησιμοποιούν τις κοντινές τιμές γύρω από μια τιμή, γίνεται τοπική βελτίωση των δεδομένων (local smoothing). Οι διορθωμένες τιμές, έπειτα πηγαίνουν σε καλάθια (bin – buckets). Επίσης, εφαρμόζονται και εδώ η μέση και η μεσαία τιμή, όπου κάθε τιμή που βρίσκεται στα καλάθια αντικαθίσταται με μια από τις δύο. Επιπλέον, μπορεί να γίνει βελτίωση των δεδομένων όταν στα καλάθια θέτονται όρια (smoothing by bin boundaries). Σε αυτήν την περίπτωση, χρησιμοποιείται η μέγιστη και η ελάχιστη τιμή μέσα σε έναν κάδο και έπειτα κάθε τιμή αντικαθίσταται με το όριο που βρίσκεται πιο κοντά. Μάλιστα, όσο πιο μεγάλα είναι τα όρια του κάδου, τόσο πιο αποτελεσματική θα είναι η συγκεκριμένη τεχνική. (Εικόνα 3.1)

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34



Εικόνα 3.1: Τεχνική κάδων – Binning [2]

- **Παλινδρόμηση (Regression):** Η βελτίωση των δεδομένων μπορεί να επιτευχθεί και με παλινδρόμηση. Παλινδρόμηση, είναι μια τεχνική που ενσωματώνει δεδομένα σε μια συνάρτηση. Μπορεί να είτε είναι γραμμική (linear regression) όπου, προσπαθούμε να βρούμε μια «γραμμή» που να ταιριάζει καλύτερα σε δύο μεταβλητές, ώστε να μπορούμε με βάση την μια να προβλέψουμε την άλλη, είτε να γίνεται με πολλές γραμμές (multiple linear regression), όπου χρησιμοποιούνται πολλές μεταβλητές και ο χώρος που τα δεδομένα θα αντιστοιχηθούν είναι πολυδιάστατος.
- **Ανάλυση διαφορετικών τιμών (Outlier Analysis):** Οι διαφορετικές τιμές σε ένα dataset μπορούν να βρεθούν με την τεχνική της ομαδοποίησης (clustering), όπου τα δεδομένα που έχουν παρόμοιες τιμές ομαδοποιούνται σε κλάδους (clusters). Επομένως, όποιες τιμές είναι εκτός κλάδων μπορούν να θεωρηθούν έκτοπες τιμές. (Εικόνα 3.2)



Εικόνα 3.2: Διαφορετικές τιμές εκτός ομάδων [2]

Μέχρι τώρα αναλύθηκαν οι τεχνικές που χρησιμοποιούνται κατά τον καθαρισμό των δεδομένων (data cleaning) ώστε, τα δεδομένα να είναι ολοκληρωμένα, συνεχή και χωρίς θόρυβο. Όμως, το data cleaning είναι ένα πολύ δύσκολο και ουσιαστικό βήμα για την ευρύτερη διαδικασία του data preprocessing και κατ' επέκταση και του data mining. Είναι ανάγκη λοιπόν, να αναδειχθούν τα εργαλεία και τα βήματα που μπορούν να εφαρμοστούν για να γίνει η συνολική διαδικασία πιο ανώδυνη και να αποφέρει καλύτερα αποτελέσματα.

Το data cleaning, ξεκινά με τον εντοπισμό «ανωμαλιών» στο σετ των δεδομένων. Οι ανωμαλίες αυτές μπορεί να έχουν προκληθεί από διάφορους παράγοντες, όπως αυτός του ανθρώπινου λάθους, τα δεδομένα να μην είναι ενημερωμένα, από τον κακό τρόπο που παρουσιάζονται τα δεδομένα ή ακόμα και από σφάλματα που προκύπτουν στα μηχανήματα που είναι υπεύθυνα για την καταγραφή των δεδομένων. Επίσης, αποκλίσεις στα δεδομένα δημιουργούνται, όταν τα δεδομένα χρησιμοποιούνται για άλλο σκοπό από αυτόν που προορίζονταν αρχικά ή κατά την διάρκεια της προεπεξεργασίας τους, στο στάδιο της ενσωμάτωσης των δεδομένων (data integration). Για να αντιμετωπιστούν όλα αυτά τα παράταιρα δεδομένα, το πρώτο βήμα που πρέπει να γίνει, είναι να χρησιμοποιηθεί κάθε πληροφορία που υπάρχει για τα δεδομένα. Για παράδειγμα, την μέση και την μεσαία τιμή, το εύρος των δεδομένων και αν υπάρχουν εξαρτήσεις μεταξύ των κατηγοριών του dataset. Οι πληροφορίες για τα δεδομένα ή τα «δεδομένα που αφορούν τα δεδομένα» ονομάζονται μετά-δεδομένα (metadata).

Στην ανάλυση δεδομένων, πρέπει να δοθεί μεγάλη προσοχή και σε λάθη που μπορεί να δημιουργηθούν και από τους προγραμματιστές των εφαρμογών διότι, πολλές φορές προσδίδουν επιπλέον καταχωρήσεις στα δεδομένα, ενώ είναι περιττό αφού έχει γίνει ήδη η απόδοση των χαρακτηριστικών των δεδομένων, προκαλώντας έτσι την λεγόμενη υπερκάλυψη του πεδίου (field overloading). Επιπροσθέτως, υπάρχουν κανόνες που τα δεδομένα θα πρέπει να ακολουθούν. Αυτοί είναι: ο κανόνας της μοναδικότητας (unique rule), ο κανόνας της διαδοχικότητας (consecutive rule) και κανόνας του κενού (null rule). Ο κανόνας της μοναδικότητας υποδείχνει, ότι κάθε καταχώρηση σε μια κατηγορία πρέπει να είναι διαφορετική από όλες τις άλλες καταχωρήσεις στην κατηγορία αυτή. Ο κανόνας της διαδοχικότητας, αναφέρει ότι δεν πρέπει να υπάρχουν κενά ανάμεσα στην μέγιστη και ελάχιστη τιμή μιας κατηγορίας και ότι κάθε τιμή πρέπει να είναι μοναδική. Τέλος, ο κανόνας του κενού διευκρινίζει τους ειδικούς χαρακτήρες (π.χ. ερωτηματικά) ή οποιαδήποτε άλλη ακολουθία χαρακτήρων, με τους οποίους μπορεί να υποδηλώνεται κάποιο κενό στα δεδομένα και πως αυτά τα κενά δεδομένα θα πρέπει να αντιμετωπίζονται.

Εκτός από τις παραπάνω μεθόδους υπάρχουν και προγραμματιστικά εργαλεία που εξυπηρετούν στην κάθαρση των δεδομένων. Αρχικά υπάρχουν εργαλεία καθαρισμού δεδομένων (data scrubbing tools), που χρησιμοποιούν απλή γνώση για να ανιχνεύσουν και να διορθώσουν δεδομένα, που συνήθως προέρχονται από διάφορες πηγές. Επιπλέον, υπάρχουν και εργαλεία ελέγχου για τα δεδομένα (data auditing tools), τα οποία αντιμετωπίζουν τα ασυνεχή δεδομένα αναλύοντας το σύνολο των δεδομένων για να ανακαλύψουν κανόνες και σχέσεις μεταξύ των δεδομένων κι έπειτα αναζητούν δεδομένα τα οποία πληρούν αυτές τις σχέσεις και τους κανόνες.

Καταλήγοντας, είναι πολύ σημαντικό να μοιραζόμαστε κάθε νέα πληροφορία που μαθαίνουμε για ένα συγκεκριμένο σετ δεδομένων, καθώς ο καθαρισμός του θα είναι πολύ πιο εύκολος αργότερα, όταν το ίδιο σετ εμπλουτισθεί με νέα δεδομένα.

3.3.2 Ενσωμάτωση δεδομένων (Data Integration)

Η εξόρυξη δεδομένων πολλές φορές απαιτεί την αξιοποίηση διάφορων πηγών, για την άντληση πληροφοριών. Επομένως, απαιτείται ο συνδυασμός και η εναρμόνιση όλων αυτών των δεδομένων που θα χρησιμοποιηθούν για αυτόν τον σκοπό. Η διαδικασία της ενσωμάτωσης δεδομένων (data

integration), αποσκοπεί στην μείωση των σφαλμάτων στο τελικό dataset που θα χρησιμοποιηθεί. Έτσι βελτιώνεται η απόδοση και μειώνεται ο χρόνος της διαδικασίας εξόρυξης. Αυτή είναι και η πραγματική πρόκληση του data integration. Παρακάτω θα αναλυθεί η διαδικασία του data integration, οι προκλήσεις που προκύπτουν κατά την διαδικασία και με ποιους τρόπους, μπορούν να αντιμετωπιστούν.

Ξεκινώντας, το ζήτημα της ενοποίησης των μεταβλητών και της αντιστοίχισης των δεδομένων από διαφορετικές πηγές μπορεί να γίνει αρκετά περίπλοκο. Αυτό το πρόβλημα είναι ευρέως γνωστό ως το «πρόβλημα αναγνώρισης του συνόλου» (entity identification problem). Για παράδειγμα, πως μπορεί ένας αναλυτής ή ένας αλγόριθμος, να είναι σίγουρος ότι δύο μεταβλητές σε δύο διαφορετικά databases αναφέρονται στο ίδιο πράγμα; Και σε αυτό, μπορεί να δοθεί λύση από τις γνώσεις που υπάρχουν για τα δεδομένα (metadata). Επίσης, όταν γίνεται η αντιστοίχιση μεταβλητών ανάμεσα σε δύο βάσεις δεδομένων πρέπει να δοθεί έμφαση και στην δομή των δεδομένων (data structure). Αυτό γίνεται, για να διασφαλιστεί ότι οι μεταβλητές στο σύστημα εισόδου και στο σύστημα εξόδου θα υπακούν στις ίδιες «αρχές και περιορισμούς».

Έπειτα, ακολουθεί το πρόβλημα των περιττών δεδομένων, καθώς πολλές φορές μια δεδομένη μεταβλητή μπορεί να αντικατοπτρίζεται από μία άλλη ή και από συνδυασμό άλλων μεταβλητών. Το αν μια μεταβλητή είναι περιττή ή όχι αναδεικνύεται από την ανάλυση σχέσεων των μεταβλητών (correlation analysis), η οποία βασίζεται στο πόσο εξαρτάται μια μεταβλητή από μια άλλη ανάλογα με τα δεδομένα που υπάρχουν. Οι τεχνικές που χρησιμοποιούνται στην συσχέτιση των δεδομένων είναι, **η μέθοδος του χ^2 για τα ονομαστικά δεδομένα**, ενώ για **αριθμητικά δεδομένα χρησιμοποιούνται ο συσχετισμός απόδοσης και η συνδιακύμανση** (coefficient correlation and covariance), οι οποίοι αναγνωρίζουν κατά πόσο οι τιμές μιας μεταβλητής διαφέρουν από αυτές μίας άλλης.

Σχετικά με τα ονομαστικά δεδομένα και την μέθοδο χ^2 ή στατιστική μέθοδος χ^2 του Pearson, όπως αλλιώς ονομάζεται. Ας υποθέσουμε, ότι έχουμε μια μεταβλητή Α με επιμέρους τιμές (α_1,α_2,α_3,…,α_n) και μια μεταβλητή Β (b_1,b_2,b_3,…,b_n) και με αυτές σχηματίζουμε έναν πίνακα, όπου τα στοιχεία της μεταβλητής Α θα είναι οι στήλες και της μεταβλητής Β θα είναι οι γραμμές (A_i, B_j). Το κάθε κελί αυτού του πίνακα εκφράζει κάθε πιθανό συσχετισμό μεταξύ των (A_i, B_j). Το χ^2 υπολογίζεται από την σχέση: $\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ (Σχέση 3.1), όπου o_{ij} είναι ο δείκτης

συχνότητας στον πίνακα για το συγκεκριμένο στοιχείο και e_{ij} είναι ο δείκτης αναμενόμενης συχνότητας, οποίος προκύπτει από την σχέση $e_{ij} = \frac{count(a_i) \times count(b_j)}{n}$ (Σχέση 3.2), όπου n είναι το συνολικό πλήθος των στοιχείων α_i είναι ο αριθμός των γραμμών του πίνακα που έχουν την συγκεκριμένη τιμή και b_j ο αριθμός των γραμμών που έχουν αυτή την τιμή. Το άθροισμα υπολογίζεται με βάση όλο το μέγεθος του πίνακα ($\alpha_n \times b_n$) Αυτό που καθορίζει την τιμή του χ^2 είναι τα στοιχεία του πίνακα των οποίων η πραγματική συχνότητα, διαφέρει αρκετά από την αναμενόμενη. Ο έλεγχος αυτός δείχνει κατά πόσο οι δυο μεταβλητές Α και Β είναι ανεξάρτητες μεταξύ τους και βασίζεται σε έναν βαθμό σπουδαιότητας με $(\alpha_n - 1) \times (b_n - 1)$ βαθμούς ελευθερίας. Εάν, το αποτέλεσμα του παραπάνω ελέγχου απορριφθεί, τότε μπορούμε να πούμε ότι τα δύο δεδομένα συσχετίζονται.

Σε αυτό το σημείο θα παρουσιαστεί ένα παράδειγμα για την καλύτερη κατανόηση της μεθοδολογίας. Ας υποθέσουμε ότι πραγματοποιούμε μια έρευνα σε πλήθος 1500 ατόμων και των δύο φύλων. Κάθε άτομο ρωτήθηκε αν προτιμά σπιτικό φαγητό (home_food) ή φαγητό από έξω (delivery). Επομένως, έχουμε δύο μεταβλητές (attributes), το γένος (gender) και το προτιμώμενο φαγητό (preferred_food). Η συχνότητα που εμφανίστηκε για κάθε πιθανό συνδυασμό φαίνεται στον πίνακα 3.1 και στις παρενθέσεις είναι οι εκτιμώμενες συχνότητες οι οποίες υπολογίζονται από την σχέση 3.2. Η σχέση 3.2 γίνεται: $e_{11} = \frac{count(male) \times count(home\_food)}{n} = \frac{300 \times 450}{1500} = 90$

preferred_food	gender		
	male	female	Total
home_food	250(90)	200(360)	450
delivery	50(210)	1000(840)	1050
Total	300	1200	1500

Πίνακας 3.1 Προτιμήσεις φαγητού

Χρησιμοποιώντας την σχέση 3.1 για το χ^2 έχουμε:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284{,}44 + 121{,}90 + 71{,}11 + 30{,}48 = 507{,}93$$

Στο συγκεκριμένο παράδειγμα που είναι 2×2, οι βαθμοί ελευθερίας είναι (2-1)×(2-1)=1. Για 1 βαθμό ελευθερίας, συμβουλευόμενοι κάποιο από τα εγχειρίδια στατιστικής μπορούμε να δούμε ότι η τιμή που πρέπει να έχει το χ^2, ώστε να απορριφθεί η υπόθεση διαφοράς στο 0,001 επίπεδο σημαντικότητας, είναι 10,828 κι εφόσον η τιμή που βρήκαμε είναι μεγαλύτερη από αυτή, μπορούμε να πούμε ότι οι δύο μεταβλητές συσχετίζονται άμεσα.

Για τα αριθμητικά δεδομένα η συσχέτιση απόδοσης (correlation coefficient), είναι ο ένας εκ των δύο μεθόδων που χρησιμοποιούνται για την αξιολόγηση των σχέσεων δύο μεταβλητών Α και Β και δίνεται από την παρακάτω σχέση:

$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\,\bar{B}}{n\sigma_A\sigma_B}$ (Σχέση 3.3) όπου, n είναι το σύνολο των γραμμών, a_i και b_i είναι η κάθε τιμή κάθε γραμμής για τα Α και Β, σ_A και σ_B είναι οι συντελεστές απόκλισης των Α και Β και το $\Sigma(a_i,b_i)$ είναι το άθροισμα του γινομένου των Α και Β. Σημειώνεται ότι $-1 \leq r_{A,B} \leq +1$.

Αν το $r_{A,B}$ είναι μεγαλύτερο του 0, τότε τα Α και Β είναι θετικά συσχετιζόμενα μεταξύ τους, πράγμα που σημαίνει ότι όταν η τιμή του ενός αυξάνεται θα αυξάνεται του άλλου, ενώ αν το $r_{A,B}$ είναι μικρότερο του 0 τότε τα Α και Β έχουν αρνητική συσχέτιση κι επομένως όταν η μια μεταβλητή αυξάνεται η άλλη θα μειώνεται. Μάλιστα, όσο πιο μεγάλη είναι η τιμή του $r_{A,B}$ τόσο πιο δυνατό συσχετισμό έχουν οι δυο μεταβλητές. Όταν, το $r_{A,B}$ είναι ίσο με το 0 τότε οι δυο μεταβλητές είναι ανεξάρτητες η μια από την άλλη.

Παρόλα αυτά, είναι σύνηθες να μπερδεύεται η συσχέτιση με την αιτιότητα. Δηλαδή, το γεγονός ότι δυο μεταβλητές μπορεί να συσχετίζονται μεταξύ τους δεν σημαίνει ότι η μια είναι αποτέλεσμα της άλλης.

Η δεύτερη μέθοδος για τα αριθμητικά δεδομένα είναι αυτή της συνδιακύμανσης (covariance), η οποία μοιάζει πολύ με την προηγούμενη, αφού και οι δύο μέθοδοι μας δείχνουν κατά πόσο δύο μεταβλητές αλλάζουν ταυτόχρονα. Για τις δύο παραπάνω μεταβλητές Α και Β γνωρίζουμε ότι η μέση ή αναμενόμενη τιμή είναι : $E(A) = \frac{\sum_{i=1}^{n} a_i}{n}$ για την μεταβλητή Α και αντίστοιχα για την Β.

Η συνδιακύμανση μεταξύ των Α και Β υπολογίζεται από την σχέση:

$$Cov(A,B) = E\big((A - \bar{A})(B - \bar{B})\big) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$ (Σχέση 3.4)

Η ίδια σχέση γράφεται και με την μορφή: $Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$ (Σχέση 3.5), η οποία χρησιμεύει για την διευκόλυνση των υπολογισμών.

Στην συγκεκριμένη μέθοδο οι μεταβλητές Α και Β τείνουν να μεταβάλλουν η μια την άλλη, εάν η συγκεκριμένη τιμή της Α (Α) είναι μεγαλύτερη από την εκτιμώμενη τιμή της Α ($\bar{A}$), τότε το ίδιο θα συμβαίνει και την Β. Σε αυτή την περίπτωση οι δυο μεταβλητές συσχετίζονται θετικά. Αντιθέτως, όταν μια από τις δυο μεταβλητές έχει μεγαλύτερη τιμή από την αναμενόμενη και η δεύτερη έχει μικρότερη, τότε η συνδιακύμανση μεταξύ των δυο μεταβλητών είναι αρνητική. Τέλος, όταν η διακύμανση είναι μηδενική, τότε οι δυο μεταβλητές είναι ανεξάρτητες, αλλά μόνο κάτω από κάποιες συγκεκριμένες παραδοχές μπορεί να συμβεί αυτό.

Παρακάτω παρατίθεται παράδειγμα που δείχνει πώς λειτουργεί η συνδιακύμανση.

Ας υποθέσουμε ότι έχουμε τις τιμές δυο προϊόντων, σε πέντε διαφορετικές χρονικές στιγμές. Οι τιμές θα επηρεαστούν το ίδιο από την συμπεριφορά των καταναλωτών ή όχι;

Χρονική Στιγμή	Product1	Product2
T1	6	20
T2	5	10
T3	4	14
T4	3	5
T5	2	5

Πίνακας 3.2 Τιμές Προϊόντων

Αρχικά θα υπολογίσουμε τις εκτιμώμενες τιμές για τα δύο προϊόντα:

$$E(Product1) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = 4€$$

$$E(Product2) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = 10.80€$$

Έπειτα, υπολογίζουμε την συνδιακύμανση μεταξύ των δυο προϊόντων:

$$Cov(Product1, Product2) = \frac{6 \times 10 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10{,}80$$
$$= 50{,}2 - 43{,}2 = 7$$

Επομένως, εφόσον η συνδιακύμανση είναι θετική , μπορούμε να εκτιμήσουμε ότι οι δύο τιμές θα ανέβουν ή θα πέσουν μαζί.

Στην ενσωμάτωση δεδομένων, εκτός από τα περιττά δεδομένα στις μεταβλητές των βάσεων, πρέπει να ελέγχεται και αν υπάρχουν διπλές γραμμές στις βάσεις δεδομένων. Δηλαδή γίνεται έλεγχος, για τον αν υπάρχουν παραπάνω από μια γραμμές που να έχουν την ίδια τιμή. Αυτές, μπορεί να προκύψουν όταν δεν γίνεται σωστή καταχώρηση ή ενημέρωση των δεδομένων καθώς και όταν χρησιμοποιούνται ακανόνιστοι πίνακες. Ακόμη, όταν συνδυάζονται διάφορες πηγές τυγχάνει πολλές τιμές των δεδομένων, ενώ αφορούν την ίδια μεταβλητή σε όλες τις συνδυαζόμενες πηγές να χρησιμοποιούν διαφορετικό τρόπο απεικόνισης, όπως είναι η ταχύτητα που εκφράζεται σε μίλια ανά ώρα και σε χιλιόμετρα ανά ώρα. Το ίδιο ισχύει και για τις μεταβλητές, καθώς μια μεταβλητή μπορεί να αντιπροσωπεύει το ίδιο πράγμα με μια άλλη, αλλά σε διαφορετική κλίμακα. Για παράδειγμα, σε μια βάση δεδομένων για μια αλυσίδα καταστημάτων, υπάρχει η μεταβλητή για τις συνολικές πωλήσεις της χρονιάς για όλα τα καταστήματα, ενώ σε μια άλλη βάση δεδομένων για την ίδια αλυσίδα, υπάρχει μεταβλητή με το ίδιο όνομα αλλά να εκφράζει τις συνολικές πωλήσεις της χρονιάς για ένα συγκεκριμένο κατάστημα και όχι για ολόκληρη την αλυσίδα καταστημάτων.

3.3.3 Μείωση Δεδομένων (Data Reduction)

Δεν είναι λίγες οι φορές που ο όγκος των δεδομένων θα είναι τόσο μεγάλος, που θα καθιστά χρονοβόρα και δύσκολη την ανάλυση τους. Για αυτό, χρησιμοποιούνται τεχνικές δεδομένων ώστε να μειώσουν τον όγκο των δεδομένων σε έναν πολύ μικρότερο, αλλά ταυτόχρονα να διατηρείται η ακεραιότητα των αρχικών δεδομένων ώστε να μην υπάρχει διαφορά στο τελικό αποτέλεσμα της ανάλυσης ακόμα κι αν χρησιμοποιούνται τα μειωμένα δεδομένα. Η μείωση των δεδομένων (data reduction), περιλαμβάνει τεχνικές που οδηγούν στην μείωση των διαστάσεων (dimensionality reduction), την αριθμητική μείωση (numerosity reduction) και την συμπίεση των δεδομένων (data compression).

Η μείωση των διαστάσεων (dimensionality reduction), στοχεύει στην μείωση τυχαίων μεταβλητών, οι οποίες σε πρώτη φάση θεωρούνται σημαντικές για την ανάλυση. Κάποιες από τις τεχνικές που ακολουθούνται ώστε να μειωθούν οι διαστάσεις των δεδομένων είναι, η τεχνική μετασχηματισμού wavelet, η ανάλυση κυρίαρχων συνιστωσών (principal component analysis), η επιλογή μεταβλητών μέσα από το σετ δεδομένων (attribute subset selection), η γραμμική παλινδρόμηση και η χρήση λογαριθμικών μοντέλων (regression and log-linear models), η ομαδοποίηση των δεδομένων (clustering) και την χρήση ιστογραμμάτων.

Ο μετασχηματισμός wavelet (discrete wavelet transform-DWT), είναι μια τεχνική επεξεργασίας γραμμικών σημάτων που μετατρέπει μια «σειρά» δεδομένων, σε μια νέα «σειρά» δεδομένων που διαφέρει αριθμητικά από την αρχική. Παρόλο που τα νέα «wavelet» δεδομένα μπορούν να ελαχιστοποιηθούν σε μεγάλο βαθμό, επειδή κρατούν τα πιο έντονα συσχετιζόμενα στοιχεία, η ακρίβεια τους πλησιάζει κατά πολύ αυτή των αρχικών. Με αυτόν τον τρόπο οι υπολογιστικές διαδικασίες γίνονται πολύ γρήγορα στον χώρο wavelet και αν έχουμε στους συσχετισμούς wavelet εφαρμόζοντας την αντίστροφη τεχνική μπορούμε να αποκτήσουμε μια εικόνα για τα αρχικά δεδομένα. Τα βήματα που ακολουθούνται για να εφαρμοστεί ο μετασχηματισμός DWT είναι:

1. Το μήκος της «σειράς» των δεδομένων πρέπει να είναι ακέραιος αριθμός της δύναμης του 2. Αν δεν είναι συμπληρώνονται μηδενικά στοιχεία μέχρι να επιτευχθεί.
2. Εφαρμόζονται δύο συναρτήσεις σε κάθε μετασχηματισμό. Η πρώτη εξομαλύνει τα δεδομένα, όπως γίνεται στην συνάρτηση αθροίσματος και η δεύτερη συνάρτηση βρίσκει ποιες κατηγορίες των δεδομένων είναι πιο σημαντικές.
3. Οι δύο συναρτήσεις εφαρμόζονται σε «ζευγάρια» στοιχείων της σειράς δεδομένων κι έτσι δημιουργούνται δυο νέα dataset του μισού μεγέθους, τα οποία είναι πιο καθαρά.
4. Οι δυο συναρτήσεις εφαρμόζονται έως ότου τα dataset που προκύπτουν να είναι μήκους 2.
5. Επιλεγμένες τιμές από τα τελικά dataset προηγουμένως καθορίζονται ως οι συσχετισμοί wavelet των μετασχηματισμένων δεδομένων.

Η ανάλυση κυρίαρχων συνιστωσών (principal component analysis) μπορεί να χρησιμοποιηθεί σε δυσδιάστατα και πολυδιάστατα δεδομένα καθώς και σε αλλοιωμένα ή διασκορπισμένα δεδομένα. Ας υποθέσουμε, ότι τα δεδομένα που θέλουμε να μειώσουμε μπορούν να αναλυθούν σε έναν

αριθμό ‘ν’ διαστάσεων. Η ανάλυση κυρίαρχων συνιστωσών ψάχνει ‘χ’ τέτοιους ορθογωνιακούς τομείς ($\chi \leq \nu$) που μπορούν να παρουσιάσουν τα αρχικά δεδομένα όσο το δυνατόν καλύτερα. Πιο αναλυτικά τα βήματα της διαδικασίας είναι:

1. Τα δεδομένα εισόδου κανονικοποιούνται, έτσι ώστε όλα να έχουν το ίδιο μέγεθος και όλες οι κατηγορίες να αντιμετωπίζονται ισότιμα.
2. Υπολογίζονται οι νέοι «τομείς» που θα αποτελέσουν την βάση για τα νέα κανονικοποιημένα δεδομένα.
3. Γίνεται η ταξινόμηση των κύριων συνιστωσών με φθίνουσα σειρά με βάση την σημαντικότητα τους, δίνοντας έτσι πολύτιμες πληροφορίες σχετικά με την διακύμανση που έχουν μεταξύ τους.
4. Τα δεδομένα που είναι λιγότερο σημαντικά διαγράφονται μειώνοντας ακόμη περισσότερο την το τελικό μέγεθος των δεδομένων.

Ένας ακόμη τρόπος για να μειώνουμε τον όγκο μεγάλων δεδομένων, είναι να μην χρησιμοποιούμε τις μεταβλητές που δεν έχουν σχέση με την εφαρμογή που θέλουμε να υλοποιήσουμε. Αυτό γίνεται με την διαδικασία επιλογής μεταβλητών μέσα από το σετ (attribute subset selection). Τελικός στόχος είναι να αφαιρεθούν όσες μεταβλητές δεν χρειάζονται και να δημιουργηθεί ένα νέο σετ δεδομένων, με όσο το δυνατόν λιγότερες μεταβλητές αλλά χωρίς να αλλοιώνεται η αρχική πληροφορία. Οι μεταβλητές αξιολογούνται με βάση, το πόσο σημαντικές είναι στατιστικά και διαλέγονται χρησιμοποιώντας τοπικά μέγιστα και ελπίζοντας ότι έτσι θα βρεθεί η βέλτιστη λύση.

Οι ευρηματικές τεχνικές που χρησιμοποιούνται εδώ είναι:

1. Βηματική επιλογή (stepwise forward selection): Το νέο σετ δεδομένων είναι κενό, ενώ σε κάθε επανάληψη προστίθενται οι καλύτερες μεταβλητές.
2. Αντίστροφη βηματική διαγραφή (stepwise backward elimination): Το νέο σετ είναι ίδιο με το αρχικό και σε κάθε επανάληψη αφαιρούνται οι πιο ασήμαντες μεταβλητές.
3. Δέντρα απόφασης (decision tree): Οι μεταβλητές των δεδομένων εισόδου, που δεν υπάρχουν στα δέντρα απόφασης χαρακτηρίζονται ως άσχετες με την εφαρμογή.

Επιπλέον, η μείωση των παραμέτρων που χρησιμοποιούνται γίνεται με γραμμική παλινδρόμηση ή με λογαριθμικά μοντέλα. Στην απλή γραμμική παλινδρόμηση (linear regression), μια μεταβλητή παρουσιάζεται ως γραμμική συνάρτηση μιας άλλης μεταβλητής σε συνδυασμό με κάποιες

μεταβλητές, οι οποίες είναι και οι συντελεστές απόδοσης. Στην πολλαπλή γραμμική παλινδρόμηση (multiple linear regression) χρησιμοποιούνται δύο ή και παραπάνω μεταβλητές για να εκφράσουν μια άλλη.

Τα λογαριθμικά μοντέλα (log-linear models) από την άλλη, χρησιμοποιούνται για την απεικόνιση μιας μεταβλητής στον χώρο. Κατά την διαδικασία αυτή κάθε καταχώρηση θεωρείται ως ένα σημείο ενός χώρου 'ν' διαστάσεων. Τα μοντέλα αυτά λοιπόν, εκτιμούν τις πιθανότητες που έχει κάθε καταχώρηση στον χώρο αυτό με βάση επιλεγμένες μεταβλητές. Επομένως, τα λογαριθμικά μοντέλα εξυπηρετούν στην μείωση των δεδομένων αφού, καταλαμβάνουν λιγότερο χώρο από τα αρχικά δεδομένα.

Μία ακόμα πολύ καλή προσέγγιση για την μείωση των δεδομένων είναι η χρήση ιστογραμμάτων. Μέσα από τα ιστογράμματα έχουμε μια πολύ απεικόνιση για την συχνότητα εμφάνισης δεδομένων αλλά και για την ομαδοποίηση τους, καθώς στα ιστογράμματα το πλάτος τους μπορεί να εκφράζει ένα συγκεκριμένο εύρος τιμών που υπάρχουν στο dataset.

Τέλος, η ομαδοποίηση των δεδομένων (clustering) όταν χρησιμοποιείται στην μείωση των δεδομένων είναι πολύ χρήσιμη, διότι χωρίζουν τα δεδομένα σε ομάδες ανάλογα με το πόσο «μοιάζουν' μεταξύ τους κι έτσι γίνεται πιο ξεκάθαρο ποια δεδομένα χρειάζονται για την εφαρμογή που θέλουμε να πραγματοποιήσουμε αγνοώντας τα δεδομένα που δεν μας είναι χρήσιμα.

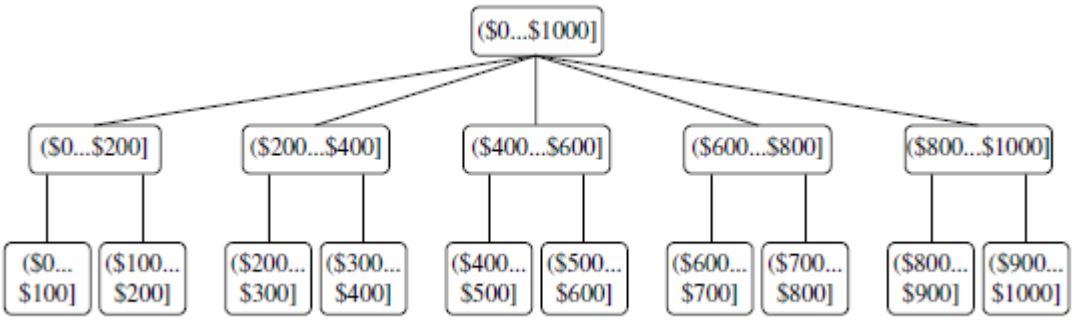
3.3.4 Μετατροπή και διακριτοποίηση δεδομένων (Data transformation and discretization)

Σε αυτό το βήμα της προεπεξεργασίας των δεδομένων, πραγματοποιείται η μετατροπή ή ενοποίηση των δεδομένων ώστε η διαδικασία της εξόρυξης να είναι πιο αποδοτική και τα τελικά δεδομένα πιο ευκολονόητα.

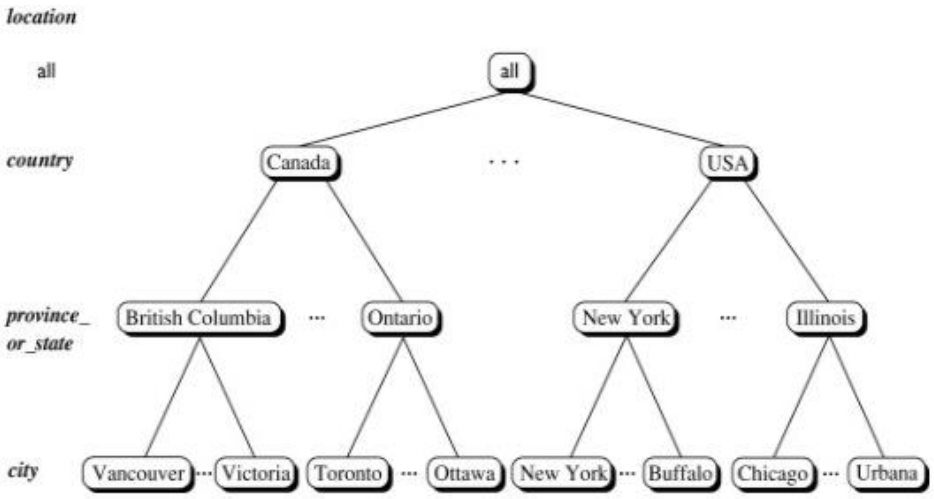
Οι τρόποι με τους οποίους επιτυγχάνεται η μετατροπή των δεδομένων, περιλαμβάνουν τα παρακάτω:

1. Ομαλοποίηση (smoothing) των δεδομένων ώστε, τα δεδομένα να απαλλαγούν από τον θόρυβο. Αυτό επιτυγχάνεται με την τεχνικές όπως το regression ή το clustering.

2. Δημιουργία μεταβλητών (feature construction), όπου νέες μεταβλητές προστίθενται με βάση τις προηγούμενες για να διευκολύνουν την διαδικασία.
3. Συνάθροιση (aggregation), των δεδομένων με σκοπό την πιο συγκεντρωτική απεικόνιση αυτών.
4. Κανονικοποίηση (normalization), με την οποία όλα τα δεδομένα εκφράζονται στην ίδια κλίμακα κι έχουν το ίδιο «βάρος».
5. Διάκριση (discretization), όπου απλά αριθμητικά δεδομένα (π.χ. τιμή) αποκτούν ετικέτες αριθμητικού εύρους (π.χ. 0-5, 6-10) είτε ονομαστικές ετικέτες (π.χ. ακριβό, φθηνό). Αυτές οι ετικέτες εξυπηρετούν στην καλύτερη οργάνωση των δεδομένων, η οποία οδηγεί σε ευρύτερες έννοιες δημιουργώντας παράλληλα μια μορφή «εννοιολογικής» ιεραρχίας (concept hierarchy), για τα δεδομένα μιας μεταβλητής. (Εικόνα 3.3)
6. Εννοιολογική ιεραρχία για ονομαστικά δεδομένα (concept hierarchy for nominal data), με την οποία ονομαστικές μεταβλητές (π.χ. πόλη) μπορούν να γενικευθούν σε ευρύτερες έννοιες (π.χ. περιφέρεια ή χώρα). (Εικόνα 3.4)



Εικόνα 3.3: Εννοιολογική ιεραρχία με αριθμητικά δεδομένα [2]



Εικόνα 3.4: Εννοιολογική ιεραρχία με ονομαστικά δεδομένα [2]

Ανάλογα με τον τύπο των δεδομένων που έχουμε εφαρμόζονται και οι ανάλογες τεχνικές, όταν θέλουμε να βρούμε την εννοιολογική ιεραρχία κατά την διαδικασία της διάκρισης. Στα αριθμητικά δεδομένα εφαρμόζονται τεχνικές που χρησιμοποιούνται και σε άλλα στάδια του preprocessing, όπως το binning και το clustering στο data cleaning. Όμως στα ονομαστικά δεδομένα η εννοιολογική ιεραρχία μπορεί να δημιουργηθεί με τους εξής τρόπους:

- Προσδιορίζοντας τις μεταβλητές που θα χρησιμοποιηθούν, κατά την διάρκεια της σχεδίασης, από τους ειδικούς.
- Προσδιορίζοντας ένα μέρος της εννοιολογικής ιεραρχίας, με την ομαδοποίηση των δεδομένων σε σαφής ομάδες.
- Ανάλογα με τον αριθμό καταχωρήσεων, που έχει η κάθε μεταβλητή. Η μεταβλητή που θα είναι στην κορυφή της ιεραρχίας συνήθως είναι αυτή που έχει τον μικρότερο αριθμό μεταβλητών, αλλά πάντα ελέγχεται από τον χρήστη γιατί αυτό εξαρτάται από την εφαρμογή.
- Χρησιμοποιώντας μόνο ένα μέρος από τις μεταβλητές.

Τέλος, η διαδικασία της διάκρισης μπορεί να χρησιμοποίει πληροφορίες σχετικά με το πρόβλημα, τότε λέμε ότι η διαδικασία είναι επιβλεπόμενη (supervised), ενώ σε κάθε άλλη περίπτωση δεν είναι επιβλεπόμενη (unsupervised). Επίσης, αν η διαδικασία, όσο επαναλαμβάνεται, αρχίζει

βρίσκοντας σημεία τέτοια ώστε να μπορεί να χωρίσει τα δεδομένα μίας μεταβλητής, σε μικρότερα σύνολα, τότε ονομάζεται διάκριση διαχωρισμού (splitting). Από την άλλη πλευρά, όταν η διαδικασία της διάκρισης, «συγχωνεύει» δεδομένα μιας μεταβλητής με άλλα γειτονικά, για να σχηματίσει επιμέρους σύνολα, τότε ονομάζεται διάκριση συγχώνευσης (merging).

Κεφάλαιο 4: Δεδομένα και Μηχανική Μάθηση

4.1 Τι είναι η μηχανική μάθηση;

Μηχανική μάθηση (machine learning) είναι το επιστημονικό πεδίο του προγραμματισμού των ηλεκτρονικών υπολογιστών, με το οποίο μαθαίνουν με βάση τα δεδομένα.

Μια πιο σφαιρική προσέγγιση για την μηχανική μάθηση υποστηρίζει ότι:

Μηχανική μάθηση (machine learning) είναι το αντικείμενο μελέτης, που επιτρέπει στους υπολογιστές να «μαθαίνουν» χωρίς να είναι εξ' ολοκλήρου προγραμματισμένοι.

– Arthur Samuel,1959

Μια πιο «μηχανική» προσέγγιση προσδιορίζει την μηχανική μάθηση ως εξής:

Ένα πρόγραμμα υπολογιστή μαθαίνει εμπειρικά έχοντας ως γνώμονα ένα συγκεκριμένο πρόβλημα στο οποίο καλείται να βρει την βέλτιστη λύση και ένα μέτρο απόδοσης, αν η απόδοση του στο συγκεκριμένο πρόβλημα, βελτιώνεται με βάση την εμπειρία που έχει.

– Tom Mitchell, 1997

4.2 Γιατί χρησιμοποιήσουμε την μηχανική μάθηση;

Όταν έχουμε να αντιμετωπίσουμε ένα πιο σύνθετο πρόβλημα, είναι πολύ δύσκολο και χρονοβόρο να ακολουθήσουμε μια πιο παραδοσιακή προσέγγιση γράφοντας έναν αλγόριθμο κι αυτό διότι, προκειμένου να καλύψουμε όλες τις περιπτώσεις θα πρέπει, να γράφουμε ένα νέο κομμάτι για τον αλγόριθμο που να αντιστοιχεί σε κάθε περίπτωση και να ελέγχουμε το αποτέλεσμα κάθε μέρους του αλγορίθμου, ώστε στο τέλος να είμαστε σίγουροι ότι το τελικό αποτέλεσμα του αλγόριθμου μας καλύπτει. Επίσης, τα προβλήματα, για τα οποία δεν υπάρχουν αλγόριθμοι επίλυσης, είναι ακόμα πιο δύσκολο να λυθούν με την χρήση αλγορίθμων καθώς αυτοί θα πρέπει να δημιουργηθούν από την αρχή.

Η μηχανική μάθηση με τα μέσα που προσφέρει όχι μόνο επιλύει τα παραπάνω καίρια αλγοριθμικά ζητήματα, αλλά εξυπηρετεί τους χρήστες γενικότερα διότι:

- Η μηχανική μάθηση κάνει τον κώδικα πιο απλό και φέρνει καλύτερα αποτελέσματα συγκριτικά με την παραδοσιακή προσέγγιση.
- Σε προβλήματα όπου η λύση δεν μπορεί να βρεθεί μέσω αλγορίθμων, οι τεχνικές μηχανικής μάθησης ίσως τα καταφέρουν.
- Τα συστήματα μηχανικής μάθησης, έχουν την δυνατότητα να προσαρμόζονται σε νέα δεδομένα.
- Η μηχανική μάθηση βοηθά τον χρήστη να κατανοήσει καλύτερα τα προβλήματα και τα δεδομένα που έχει, δίνοντας του επιπλέον πληροφορίες.

4.3 Τύποι συστημάτων μηχανικής μάθησης

Τα συστήματα της μηχανικής μάθησης, επειδή είναι πολυάριθμα έχουν χωριστεί σε ευρύτερες κατηγορίες με βάση συγκεκριμένα κριτήρια. Τα κριτήρια αυτά αναφέρονται στο αν τα συστήματα εκπαιδεύονται υπό την επίβλεψη του προγραμματιστή (supervised/unsupervised learning), αν μπορούν να «μάθουν» σταδιακά ή χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα (online/batch learning) και αν λειτουργούν συγκρίνοντας νέα δεδομένα με αυτά που έχουν ήδη ή βρίσκοντας συσχετισμούς στα δεδομένα εκπαίδευσης δημιουργώντας ένα μοντέλο πρόβλεψης (instance-based/model-based learning).

4.*3.1 Supervised/Unsupervised Learning*

Αρχικά, στην επιβλεπόμενη μάθηση (supervised learning), το σετ δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου, περιέχει ήδη δεδομένα με τις επιθυμητές λύσεις (labels), τις χρησιμοποιεί ως παράδειγμα και ανάλογα κατηγοριοποιεί ή δίνει κάποια πρόβλεψη για τα νέα δεδομένα.

Στην μάθηση χωρίς επίβλεψη, όπως είναι αναμενόμενο τα δεδομένα εκπαίδευσης δεν έχουν, τις επιθυμητές λύσεις κι έτσι το σύστημα προσπαθεί να μάθει μόνο του.

Όσον αφορά, την ημι-επιβλεπόμενη μάθηση (semisupervised learning), είναι προφανές ότι πρόκειται για έναν συνδυασμό της μάθησης με επίβλεψη και της μάθησης χωρίς επίβλεψη. Δηλαδή, μερικοί αλγόριθμοι μπορούν μόνο με ένα μικρό μέρος των δεδομένων, να έχει κατηγοριοποιηθεί ενώ το μεγαλύτερο όχι, να αποδώσουν ένα καλό αποτέλεσμα.

Στην ενισχυμένη μάθηση (reinforcement learning), το σύστημα παρατηρεί το περιβάλλον των δεδομένων, εκτελεί κάποιες ενέργειες και παίρνει κάποια ανταμοιβή ή ποινή ανάλογα με το αποτέλεσμα. Το σύστημα με αυτόν τον τρόπο, μαθαίνει από μόνο του ποια είναι η καλύτερη στρατηγική και ενέργειες για να παίρνει περισσότερες αμοιβές όσο προσπαθεί.

4.*3.2 Online/Batch Learning*

Ένα ακόμη κριτήριο κατηγοριοποίησης των συστημάτων της μηχανικής μάθησης είναι, αν μπορούν να εκπαιδευτούν μαθαίνοντας σταδιακά από ένα πλήθος νέων δεδομένων ή αν χρησιμοποιούν όλα τα δεδομένα που έχουν διαθέσιμα για την εκπαίδευση τους.

Στο batch learning, το μοντέλο εκπαιδεύεται χρησιμοποιώντας κατευθείαν όλα τα διαθέσιμα δεδομένα, το οποίο είναι αρκετά χρονοβόρο και συνήθως γίνεται εκτός δικτύου (offline). Το μοντέλο αφού εκπαιδευτεί στα δεδομένα αυτά, τίθεται σε λειτουργία χωρίς να μαθαίνει πλέον και εφαρμόζοντας ότι έχει μάθει. Για αυτό το batch learning, είναι γνωστό και ως offline learning. Αν το μοντέλο που ακολουθεί την συγκεκριμένη μέθοδο, θέλουμε να εκπαιδευτεί σε νέα δεδομένα πρέπει να ξανά εκπαιδευτεί σε νέο σετ δεδομένων, που περιέχει τα παλιά αλλά και τα νέα δεδομένα, ξανά από την αρχή. Στην μηχανική μάθηση αυτή η διαδικασία μπορεί να αυτοματοποιηθεί αρκετά εύκολα, οπότε και το batch learning χαρακτηρίζεται από μια προσαρμοστικότητα. Παρόλα αυτά, το να χρησιμοποιεί κανείς όλα τα δεδομένα που διαθέτει από την αρχή, είναι αρκετά χρονοβόρο και χρήσιμο μόνο όταν τα δεδομένα δεν μεταβάλλονται συχνά. Επίσης, όταν χρησιμοποιούμε ολόκληρα σετ δεδομένων για την εκπαίδευση ενός μοντέλου, οι υπολογιστικές απαιτήσεις αυξάνονται, ειδικότερα αν το μοντέλο εκπαιδεύεται συχνά από την αρχή, το κόστος γίνεται πολύ μεγαλύτερο. Ακόμα αν τα δεδομένα είναι πάρα πολλά μπορεί να είναι αδύνατον να ακολουθήσουμε την μέθοδο του batch learning, καθώς αν το σύστημα που έχουμε είναι περιορισμένων δυνατοτήτων και θέλουμε να μπορεί να μαθαίνει αυτόνομα, το να φορτώνουμε μεγάλους όγκους δεδομένων και να δαπανούμε πόρους και χρόνο για την εκπαίδευση του, προφανώς είναι ασύμφορο.

Από την άλλη, με το online learning, το μοντέλο εκπαιδεύεται με εκθετική πρόοδο καθώς τα δεδομένα εκπαίδευσης τροφοδοτούνται σε αυτό έχοντας μια συνέχεια, σε μικρά σύνολα (mini batches) ή ξεχωριστά. Κάθε μέρος της συγκεκριμένης διαδικασίας έχει χαμηλό κόστος και γίνεται

πολύ γρήγορα, καθιστώντας τα μοντέλα της συγκεκριμένη μεθόδου ιδανικά για εφαρμογές όπου υπάρχει συνεχής ροή δεδομένων και τα συστήματα πρέπει να προσαρμόζονται στιγμιαία ή να είναι αυτόνομα. Επιπλέον, το online learning είναι ιδανικό όταν δεν υπάρχουν αρκετοί πόροι, αφού μόλις το μοντέλο μάθει από τα νέα δεδομένα, τα καταστρέφει κάνοντας τα μοντέλα του online learning χρήσιμα σε περιπτώσεις όπου το dataset είναι τόσο μεγάλο που δεν χωράει στην μνήμη. Το σύστημα, παίρνει μέρος των δεδομένων, εκπαιδεύεται σε αυτό και έπειτα συνεχίζει μέχρι να καλυφθεί όλο το dataset.

Στα συστήματα online learning, υπάρχει η πολύ σημαντική παράμετρος του ρυθμού εκμάθησης, που εκφράζει το πόσο γρήγορα μπορούν τα συστήματα να προσαρμόζονται σε νέα δεδομένα. Αν αυξήσουμε κατά πολύ αυτήν την παράμετρο, τότε το σύστημα θα προσαρμόζεται στα νέα δεδομένα με μεγάλη ταχύτητα αλλά θα «ξεχνάει» εξίσου γρήγορα τα παλαιότερα δεδομένα. Αντίστοιχα, αν ο ρυθμός εκμάθησης είναι πολύ χαμηλός το μοντέλο θα μαθαίνει πιο αργά αλλά δεν θα είναι τόσο ευαίσθητο στο θόρυβο των δεδομένων και σε παράταιρες τιμές δεδομένων (outliers).

Όπως, έχει προαναφερθεί ο όγκος και η ποιότητα των δεδομένων έχουν καθοριστικό ρόλο στην μηχανική μάθηση και ακόμα περισσότερο στα μοντέλα online learning διότι, εφόσον το μοντέλο προσαρμόζεται στιγμιαία οποιαδήποτε ανωμαλία στα δεδομένα εισόδου του μοντέλου θα φανεί στην έξοδο του. Για αυτό πρέπει τα δεδομένα να παρακολουθούνται και να επιτηρείται το σύστημα, έτσι ώστε αν πέσει η απόδοση να γίνουν οι απαιτούμενες ενέργειες.

4.3.3 Instance-Based/Versus Model-Based Learning

Μια ακόμη κατηγορία των μοντέλων της μηχανικής μάθησης, είναι αυτή της γενίκευσης. Το μεγαλύτερο μέρος της μηχανικής μάθησης αφορά τις προβλέψεις. Αυτό σημαίνει ότι έχοντας ως παράδειγμα κάποια δεδομένα, το σύστημα μπορεί να κάνει μια καλή πρόβλεψη για πιο γενικά δεδομένα που δεν έχει ξαναδεί. Το να μπορεί ένα σύστημα να ανταποκρίνεται καλά στα δεδομένα εκπαίδευσης αν και καλό, είναι ανεπαρκές. Σκοπός είναι να ανταποκρίνεται καλά σε νέα δεδομένα.

Ο πιο συνηθισμένος τρόπος μάθησης είναι αυτός της αποστήθισης. Με την γνώση αυτή μπορεί ένα σύστημα να κατηγοριοποιεί τα νέα δεδομένα με βάση τις ομοιότητες που έχουν με τα

δεδομένα εκπαίδευσης. Για να γίνει βέβαια αυτό απαιτείται ένα μέτρο ομοιότητας (similarity measure), ώστε το σύστημα να μπορεί να βρίσκει ομοιότητες μεταξύ των δεδομένων. Αυτή η μέθοδος μάθησης ονομάζεται instance-based learning (μάθηση με βάση τις μεταβλητές), διότι το μοντέλο αφού «αποστηθίσει» τα δεδομένα εκπαίδευσης, χρησιμοποιώντας το μέτρο ομοιότητας συγκρίνει τα νέα δεδομένα με αυτά που έχει αποστηθίσει.

Πέρα από την αποστήθιση των δεδομένων, ένας ακόμη τρόπος για να διευρύνουμε τους ορίζοντες μας με βάση ένα σετ δεδομένων είναι να πάρουμε τα δεδομένα αυτά, να φτιάξουμε ένα μοντέλο και έπειτα να χρησιμοποιήσουμε αυτό το μοντέλο για να κάνουμε προβλέψεις. Αυτός ο τρόπος μάθησης ονομάζεται model-based learning (μάθηση με βάση ένα μοντέλο). Αρχικά, πρέπει να επιλέξουμε ποιο μοντέλο θα χρησιμοποιήσουμε για τα δεδομένα που έχουμε. Αφού γίνει αυτό, πρέπει να βρεθούν οι παράμετροι με τις οποίες το μοντέλο μας θα έχει την καλύτερη απόδοση και αυτό γίνεται είτε με κάποια συνάρτηση απόδοσης που μετρά πόσο καλό είναι το μοντέλο, είτε με κάποια συνάρτηση κόστους που μετρά πόσο κακό είναι το μοντέλο. Με αυτόν τον τρόπο φτάνουμε στο τελικό στάδιο όπου μπορούμε να χρησιμοποιήσουμε το τελικό μας μοντέλο, για να κάνουμε νέες προβλέψεις.

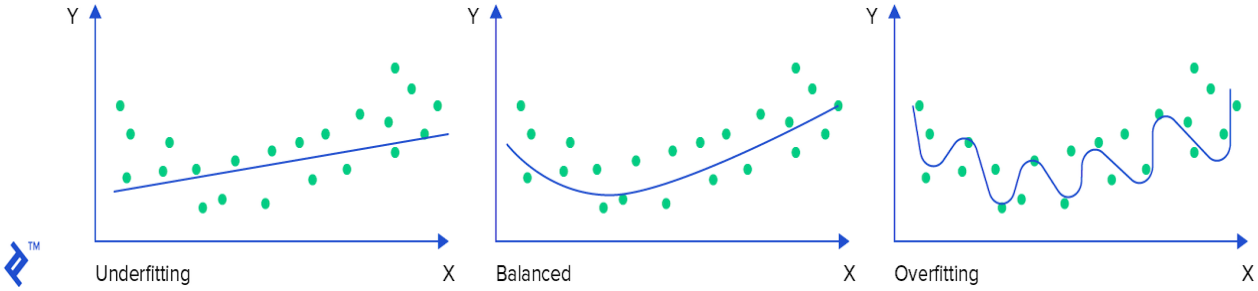
4.4 Οι προκλήσεις της Μηχανικής Μάθησης

Η μηχανική μάθηση έχει φτάσει σε τέτοιο επίπεδο που πολλές φορές, τα συστήματα μηχανικής μάθησης ξεπερνούν τον άνθρωπο. Αυτό όμως δεν ισχύει, ακόμα, και για τον τρόπο με τον οποίο μαθαίνουν τα συστήματα ή το πόσο γρήγορα μαθαίνουν. Για παράδειγμα, αν δείξουμε μερικές φορές σε ένα μικρό παιδί, πως είναι μια μπάλα το παιδί θα μάθει να αναγνωρίζει μια μπάλα σε διάφορα χρώματα και σχέδια. Για ένα, σύστημα μηχανικής μάθησης δεν είναι τόση εύκολη διαδικασία, καθώς χρειάζονται πάρα πολλά δεδομένα και χιλιάδες παραδείγματα προκειμένου να είναι σε θέση να πραγματοποιήσει κάτι τέτοιο. Επίσης, όπως έχει αναφερθεί και σε προηγούμενα κεφάλαια, προκείμενου να έχουν καλή απόδοση τα μοντέλα, είναι απαραίτητο τα δεδομένα να ποιοτικά, δηλαδή να μην έχουν θόρυβο, να μην τους λείπουν καταχωρήσεις και οι μεταβλητές που χρησιμοποιούνται να είναι στενά συνδεόμενες με την εφαρμογή που θα υλοποιηθεί.

Επιπλέον, άλλο ένα πρόβλημα που αντιμετωπίζεται στην μηχανική μάθηση είναι αυτό της εξαπάτησης του χρήστη από το μοντέλο με βάση τα δεδομένα εκπαίδευσης, δηλαδή το μοντέλο

να αποδίδει πολύ καλά στα δεδομένα εκπαίδευσης αλλά να μην μπορεί να αποδώσει καλά σε δεδομένα, που αντιμετωπίζει για πρώτη φορά. Αυτό το φαινόμενο ονομάζεται overfitting και συμβαίνει όταν το μοντέλο είναι αρκετά πιο σύνθετο συγκριτικά με τα δεδομένα ή όταν τα δεδομένα περιέχουν πολύ θόρυβο. Διαλέγοντας ένα μοντέλο, με λιγότερες παραμέτρους ή μειώνοντας το πλήθος των μεταβλητών που χρησιμοποιούνται, συλλέγοντας περισσότερα δεδομένα ή ακόμη, «καθαρίζοντας» τα δεδομένα εκπαίδευσης, μπορούμε να αποφύγουμε το overfitting και να αυξήσουμε την απόδοση του μοντέλου. Αυτή η διαδικασία περιορισμού του μοντέλου, για την αποφυγή του overfitting, ονομάζεται ρύθμιση ή κανονικοποίηση (regularization) του μοντέλου. Το πόσο ρυθμίζεται ένα μοντέλο ελέγχεται από μια ή περισσότερες υπέρ-παραμέτρους (hyper-parameters), η οποίες είναι παράμετροι του αλγόριθμου μάθησης και η ρύθμιση τους καθορίζουν την λειτουργία του τελικού συστήματος.

Όπως ένα σύνθετο μοντέλο μπορεί να «κλέψει» μαθαίνοντας πολύ καλά τα δεδομένα εκπαίδευσης, επίσης μπορεί να είναι και τόσο απλό που να χρησιμοποιεί τα χαμηλότερα δεδομένα εκπαίδευσης. Αυτό, ονομάζεται underfitting και επιλύεται χρησιμοποιώντας πιο ισχυρά μοντέλα, δίνοντας καλύτερες μεταβλητές στο μοντέλο και μειώνοντας τους περιορισμούς του μοντέλου.



Εικόνα 4.1: Παράδειγμα Γραμμικής Παλινδρόμησης [12]

Κεφάλαιο 5: Μοντέλα Επιβλεπόμενης Μάθησης

5.1 Γραμμική Παλινδρόμηση (Linear Regression)

Όπως προδίδει και το όνομα του, το συγκεκριμένο μοντέλο στηρίζει την λειτουργία του σε μια γραμμική συνάρτηση. Το μοντέλο αυτό, κάνει τις προβλέψεις του με το να υπολογίζει ένα άθροισμα βαρών των μεταβλητών στην είσοδο σε συνδυασμό με μια σταθερά, η οποία ονομάζεται bias term και αναφέρεται στον θόρυβο.

Η εξίσωση του μοντέλου πρόβλεψης της Γραμμικής Παλινδρόμησης:

$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$ (Σχέση 5.1)

Όπου: $\hat{y}$ είναι η τιμή για την οποία γίνεται η πρόβλεψη, a_0 είναι το bias term, x_n είναι η νιοστή μεταβλητή και a_n η νιοστή μεταβλητή βάρους. Η παραπάνω εξίσωση γράφεται και σε διανυσματική μορφή $\hat{y} = \alpha \cdot x$ (Εξίσωση 5.2), όπου α το διάνυσμα των σταθερών βάρους και της σταθεράς σφάλματος a_0 και *x* το διάνυσμα των μεταβλητών.

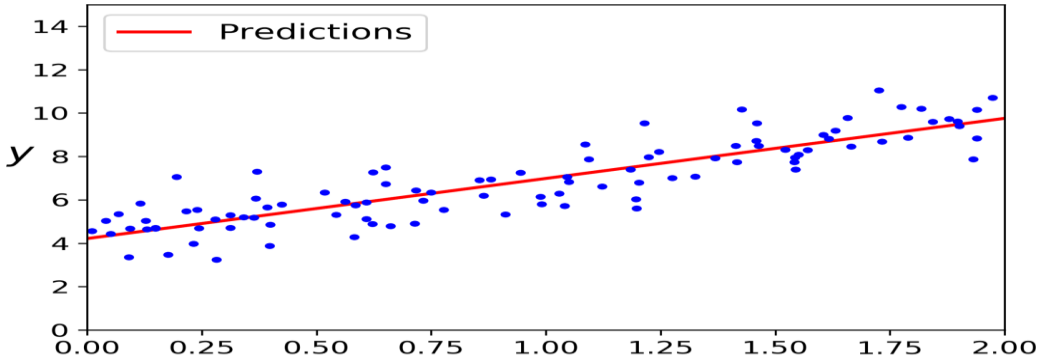
Στην μηχανική μάθηση συνηθίζεται η χρήση διανυσματικών απεικονίσεων, σε μορφή πινάκων δύο διαστάσεων με μια στήλη. Με αυτή την παραδοχή η προβλεπόμενη τιμή υπολογίζεται από την σχέση $\hat{y} = \alpha^T \cdot x$ (Εξίσωση 5.3) , όπου α^T ο ανάστροφος πίνακας του α, ενώ η τιμή πρόβλεψης παραμένει η ίδια απλά σε μορφή ενός πίνακα με μόνο ένα κελί.

Αφού το μοντέλο είναι έτοιμο, πρέπει να εκπαιδευτεί. Για να γίνει αυτό, πρέπει να βρεθούν οι μεταβλητές ώστε το μοντέλο να ταιριάζει όσο το δυνατόν περισσότερο με στα δεδομένα εκπαίδευσης. Ο πιο κοινός τρόπος για την μέτρηση της απόδοσης ενός μοντέλου είναι η εύρεση της Ρίζας του Μέσου Τετραγωνικού Σφάλματος (Root Mean Square Error - RMSE). Επομένως, πρέπει να βρεθούν οι μεταβλητές *α*, που ελαχιστοποιούν την RMSE. Για να βρεθούν οι τιμές του α, που ελαχιστοποιούν το σφάλμα, μπορεί να χρησιμοποιηθεί μια εξίσωση που δίνει κατευθείαν το αποτέλεσμα.

Αυτή είναι η ονομαζόμενη, Κανονική Εξίσωση (Normal Equation) : $\hat{\alpha} = x^T \cdot y$ (Εξίσωση 5.4)

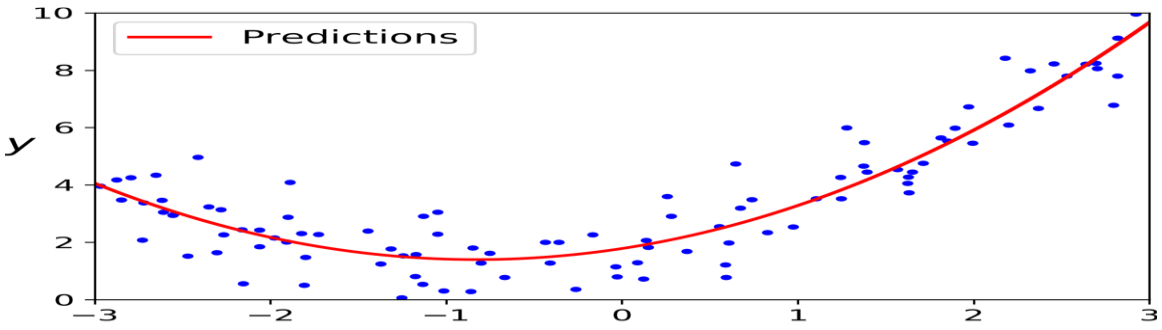
Το θετικό της γραμμικής παλινδρόμησης είναι ότι, είναι ένα μοντέλο αρκετά απλό. Ακόμα και αν αυξήσουμε τις μεταβλητές του μοντέλου, αυξάνοντας έτσι και τον βαθμό πολυπλοκότητας του, οι

προβλέψεις θα γίνουν αρκετά γρήγορα άπαξ και το μοντέλο έχει ήδη εκπαιδευτεί και ο παραπάνω χρόνος που θα χρειαστεί για τις νέες μεταβλητές είναι ανάλογος με τον αριθμό τους.



Εικόνα 5.1: Προβλέψεις με την Γραμμική Παλινδρόμηση [3]

Μια παραλλαγή της γραμμικής παλινδρόμησης, που χρησιμεύει όταν τα δεδομένα είναι πιο πολύπλοκα από μια απλή γραμμή, είναι η Πολυωνυμική Παλινδρόμηση (Polynomial Regression). Με αυτή την μέθοδο μπορεί να εφαρμοστεί ένα γραμμικό μοντέλο σε μη γραμμικά δεδομένα, συμπληρώνοντας στην εξίσωση του μοντέλου νέες μεταβλητές, οι οποίες είναι δυνάμεις των μεταβλητών (π.χ.x_1^2) που υπήρχαν ήδη στο μοντέλο. Επίσης, όταν υπάρχουν πολλές μεταβλητές η Πολυωνυμική Παλινδρόμηση μπορεί να βρίσκει και σχέσεις μεταξύ των μεταβλητών, κάτι το οποίο δεν μπορεί να κάνει η απλή Γραμμική Παλινδρόμηση. Αυτό συμβαίνει διότι, στην Πολυωνυμική Παλινδρόμηση προσθέτει και όλους τους συνδυασμούς των μεταβλητών, μέχρι τον βαθμό του πολυωνύμου που έχει προσδιοριστεί.



Εικόνα 5.2: Πολυωνυμική Παλινδρόμηση (Polynomial Regression) [3]

5.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Μερικοί αλγόριθμοι παλινδρόμησης χρησιμεύουν περισσότερο σε προβλήματα κατηγοριοποίησης. Η Λογιστική Παλινδρόμηση (Logistic Regression) χρησιμοποιείται ευρέως, για να υπολογιστεί η πιθανότητα που έχει μια νέα τιμή να ανήκει σε μια συγκεκριμένη κατηγορία (π.χ. ένα email να είναι ανεπιθύμητο). Αν η πιθανότητα που προβλέπει το μοντέλο, για το αν η τιμή ανήκει σε μια κατηγορία, είναι μεγαλύτερη από το 50%, τότε το μοντέλο συμπεραίνει ότι η τιμή ανήκει στην συγκεκριμένη κατηγορία. Μιλάμε ουσιαστικά για ένα μοντέλο δυαδικής κατηγοριοποίησης, αφού έχει μόνο δύο πιθανά αποτελέσματα, την θετική κλάση (positive class) που εκφράζεται με το δυαδικό ‘1’ και την αρνητική κλάση (negative class) που εκφράζεται με το δυαδικό ‘0’.

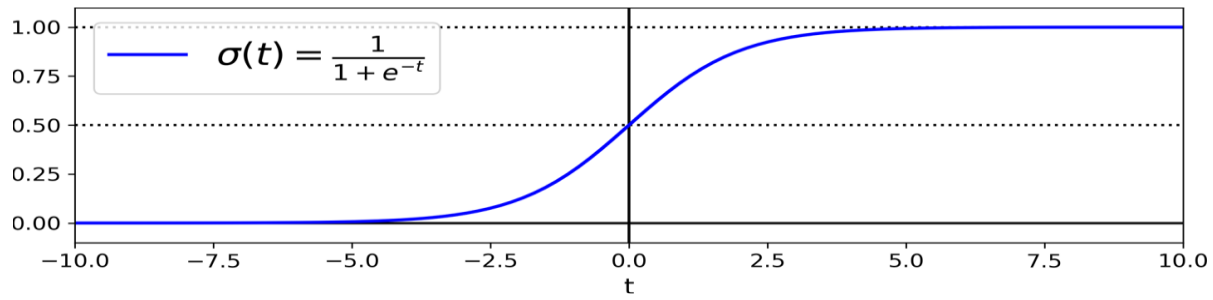
Η Λογιστική Παλινδρόμηση, λειτουργεί ακριβώς όπως και η Γραμμική Παλινδρόμηση, με μόνη διαφορά να είναι η έξοδος. Η Λογιστική Παλινδρόμηση αντί να δίνει κατευθείαν το αποτέλεσμα όπως η Γραμμική, δίνει στην έξοδο την λογαριθμική μορφή του αποτελέσματος.

Το μοντέλο της Λογιστικής Παλινδρόμησης σε διανυσματική μορφή:

$\hat{p} = \sigma(x^T \cdot \alpha)$ (Εξίσωση 5.5), όπου σ είναι η σιγμοειδής συνάρτηση που δίνει ως αποτέλεσμα αριθμούς ανάμεσα στο 0 και το 1.

Παρακάτω δίνεται και ο μαθηματικός τύπος της λογιστικής συνάρτησης (logistic function).

$\sigma(t) = \frac{1}{1+\exp(-t)}$ (Εξίσωση 5.6)



Εικόνα 5.3: Γραφική απεικόνιση της λογιστικής συνάρτησης [3]

Μόλις το μοντέλο υπολογίσει την πιθανότητα μιας καταχώρησης να ανήκει στην θετική κλάση, τότε μπορεί να κάνει εύκολα την πρόβλεψη, με βάση τον παραπάνω συλλογισμό.

Όσο για την εκπαίδευση του συγκεκριμένου μοντέλου, τελικός σκοπός είναι να βρεθούν οι τιμές του διανύσματος α έτσι ώστε, το μοντέλο να δίνει υψηλές πιθανότητες για θετικές τιμές και χαμηλές πιθανότητες για αρνητικές τιμές. Αντίστοιχα με την RMSE που είναι η συνάρτηση κόστους στην Γραμμική Παλινδρόμηση, εδώ η συνάρτηση κόστους για μια μεταβλητή έχει ως εξής:

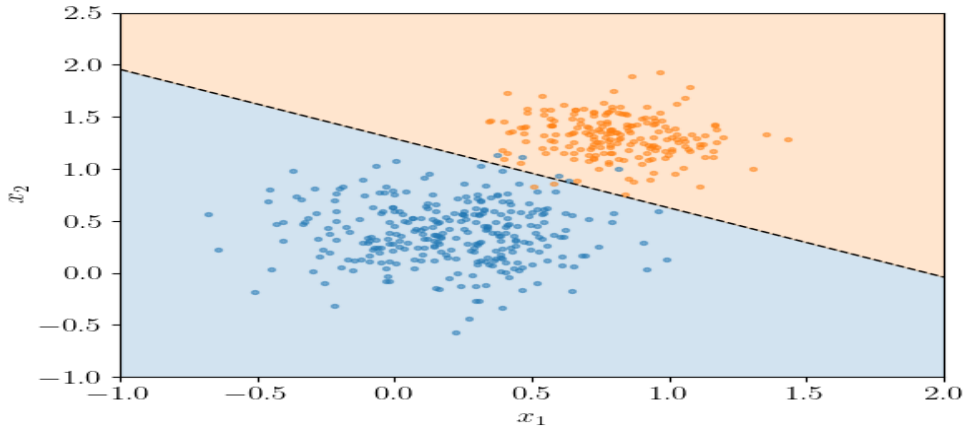
$$c(a) = \begin{cases} -\log(\hat{p}), \alpha\nu\ y = 1 \\ -\log(1-\hat{p}), \alpha\nu\ y = 0 \end{cases}$$ (Εξίσωση 5.7)

Αυτή η συνάρτηση κόστους είναι ιδανική, διότι ο αρνητικός λογάριθμος –log(t) παίρνει μεγάλες τιμές όσο το t πλησιάζει στο 0, επομένως το κόστος θα είναι μεγάλο αν το μοντέλο δώσει μια πιθανότητα κοντά στο 0 για μια θετική τιμή, όπως επίσης και αν το μοντέλο δώσει μια πιθανότητα κοντά στο 1 για μια αρνητική τιμή. Από την άλλη, ο –log(t) τείνει να μηδενιστεί όταν το t πλησιάζει στο 1, κάνοντας το κόστος μηδενικό αν η εκτιμώμενη πιθανότητα είναι κοντά στο 0 για μια αρνητική τιμή, ενώ θα είναι κοντά στο 1 για μια θετική τιμή. Για ολόκληρο το σετ δεδομένων εκπαίδευσης, η συνάρτηση κόστους είναι το μέσο κόστος από όλα τα δεδομένα εκπαίδευσης.

Συνάρτηση κόστους Λογιστικής Παλινδρόμησης (Logistic Regression cost function):

$$J(a) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(\hat{p}^{(i)}) + (1-y^{(i)})\log(1-(\hat{p}^{(i)})$$ (Εξίσωση 5.8)

Για να είμαστε σίγουροι ότι το μοντέλο κατηγοριοποίησης της Λογιστικής Παλινδρόμησης, θα κάνει σωστή πρόβλεψη πιθανοτήτων για το σε ποια κατηγορία ανήκει μια νέα τιμή θα πρέπει να προσδιοριστεί και ένα όριο. Μόλις εκπαιδευτεί, βασιζόμενο σε δύο μεταβλητές που έχουν καθοριστεί από τον χρήστη, μπορεί να προβλέψει την κατηγορία μιας νέα τιμής.

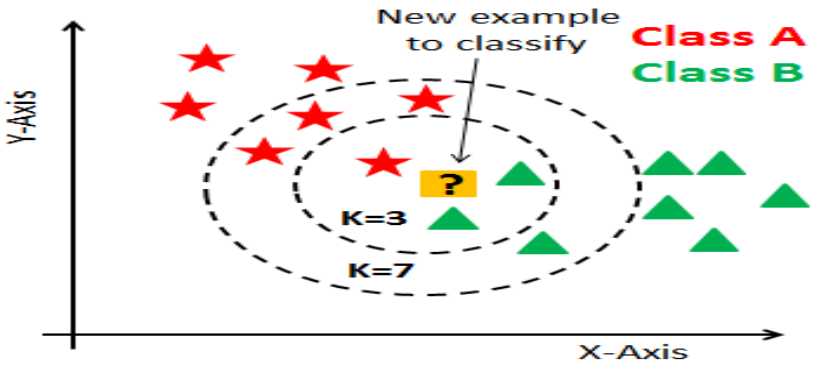


Εικόνα 5.4: Όριο Απόφασης Λογιστικής Παλινδρόμησης [13]

Πηγαίνοντας ένα βήμα παρακάτω, το μοντέλο της Λογιστικής Παλινδρόμησης μπορεί να πάρει μια πιο γενικευμένη μορφή, ώστε να μπορεί να υπολογίζει τις πιθανότητες για παραπάνω από δύο κλάσεις ταυτόχρονα, χωρίς να είναι απαραίτητη η εκπαίδευση και ο συνδυασμός διαφόρων δυαδικών μοντέλων κατηγοριοποίησης. Αυτή η πιο γενική μορφή ονομάζεται Παλινδρόμηση Softmax (Softmax Regression) και αυτό που κάνει είναι για μια νέα τιμή, υπολογίζει μια απόδοση για κάθε κατηγορία που υπάρχει και έπειτα υπολογίζει τις πιθανότητες για κάθε κατηγορία, με μια ειδική συνάρτηση.

5.3 k-Πλησιέστεροι Γείτονες (k-Nearest Neighbors)

Ο αλγόριθμος κ-πλησιέστερων γειτόνων (k-Nearest Neighbors ή kNN) είναι ένας αλγόριθμος ψήφων. Πολύ απλά χρησιμοποιεί *k* αριθμό γειτονικών δειγμάτων μέσα σε μια δεδομένη απόσταση και με βάση το πλήθος των δειγμάτων που υπερτερεί, κατατάσσεται στην ανάλογη κατηγορία. Ο αλγόριθμος αυτός, είναι πολύ απλός και τα όρια απόφασης είναι εύκολα να προσαρμοστούν ακόμα και σε πιο περίπλοκες μορφές. Λόγω της απλότητας του και της υψηλής του απόδοσης σε πολλές εφαρμογές, ο συγκεκριμένος αλγόριθμος χρησιμοποιείται σε μεγάλο εύρος, αν και αποφεύγεται σε εφαρμογές όπου υπάρχει μεγάλος όγκος δεομένων εξαιτίας του μεγάλου υπολογιστικού κόστους.



Εικόνα 5.5: Απεικόνιση kNN [14]

Στο παραπάνω πρόβλημα κατηγοριοποίησης (Εικόνα 5.5) υπάρχουν δύο κατηγορίες και ένα νέο σημείο ενδιαφέροντος και προσπαθούμε να προσδιορίσουμε σε ποια κατηγορία ανήκει. Αν πάρουμε k=3, τότε παρατηρούμε ότι το νέο σημείο θα καταταχθεί στην κατηγορία B, ενώ αν πάρουμε k=7, τότε η κατηγορία θα είναι η A. Είναι πολύ σημαντική η σωστή επιλογή του k, διότι, μπορεί ο αλγόριθμος να βρεθεί σε δίλημμα αν τα δεδομένα των δυο κατηγοριών είναι ισάριθμα και δεν θα μπορεί να γίνει η κατηγοριοποίηση. Για να αποφευχθεί η ισοπαλία είναι καλό το k να είναι μονός αριθμός.

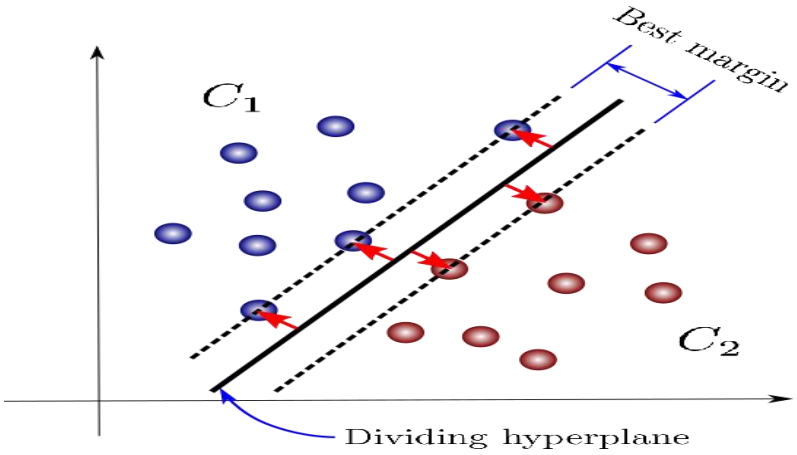
Πολύ χαμηλή τιμή του k, μπορεί να οδηγήσει σε υπερμοντελοποιήση (overfitting) και ο αλγόριθμος να είναι πιο ευάλωτος σε δεδομένα με θόρυβο, ενώ πολύ μεγάλη τιμή του k μπορεί να οδηγήσει σε μεγαλύτερο σφάλμα και χαμηλότερη ακρίβεια, διότι μπορεί να περιλαμβάνει δείγματα που να μην είναι «γείτονες» του νέου δείγματος. Επίσης, οι αποστάσεις που υπολογίζονται ανάμεσα στα δείγματα είναι ένας καθοριστικός παράγοντας και για αυτό προτείνεται πριν την εκτέλεση του αλγόριθμου, μια κανονικοποίηση ή αλλαγή της κλίμακας των δεδομένων ώστε οι αποστάσεις να είναι συγκρίσιμες.

5.4 Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines)

Οι μηχανές υποστήριξης διανυσμάτων, είναι ένα σύνολο δυνατών εργαλείων που χρησιμεύουν στην κατηγοριοποίηση, την παλινδρόμηση, την εξόρυξη δεδομένων, την βελτιστοποίησης, την τεχνητή νοημοσύνη και για αυτό είναι από τα πιο γνωστά μοντέλα στην Μηχανική Μάθηση.

Το κυριότερο πλεονέκτημα αυτών των μοντέλων σε σχέση με τα απλά γραμμικά, είναι ότι οι SVM διαχωρίζουν τις κλάσεις με τον καλύτερο δυνατό τρόπο, σχετικά με την μελλοντική ικανότητα γενίκευσης. Το πόσο καλός θα είναι ο μελλοντικός διαχωρισμός, εξαρτάται από την απόσταση των δεδομένων από την διαχωριστική επιφάνεια, καθώς όσο πιο μικρή είναι η απόσταση τόσο πιθανότερο είναι τα στοιχεία των δεδομένων να περάσουν στην λάθος πλευρά, ενώ με μεγαλύτερη απόσταση, υπάρχει μεγαλύτερη ασφάλεια. Επομένως, οι SVM αναζητούν την επιφάνεια με το μεγαλύτερο δυνατό περιθώριο από το πλησιέστερο στοιχείο των δεδομένων.

Στην περίπτωση κατηγοριοποίησης έχοντας δύο κατηγορίες, οι γραμμικές SVM, αφού χωρίσουν τα δεδομένα βάση του ορίου απόφασης, δημιουργούν δυο διανύσματα υποστήριξης (support vectors), ώστε να τα δεδομένα να είναι όσο πιο μακριά γίνεται από την γραμμή του ορίου απόφασης. Με αυτόν τον τρόπο δημιουργούνται δύο ακόμη υπερεπίπεδα (hyperplanes) με σκοπό την μέγιστη απόσταση μεταξύ τους και ανάμεσα τους δεν πρέπει να υπάρχουν δεδομένα.

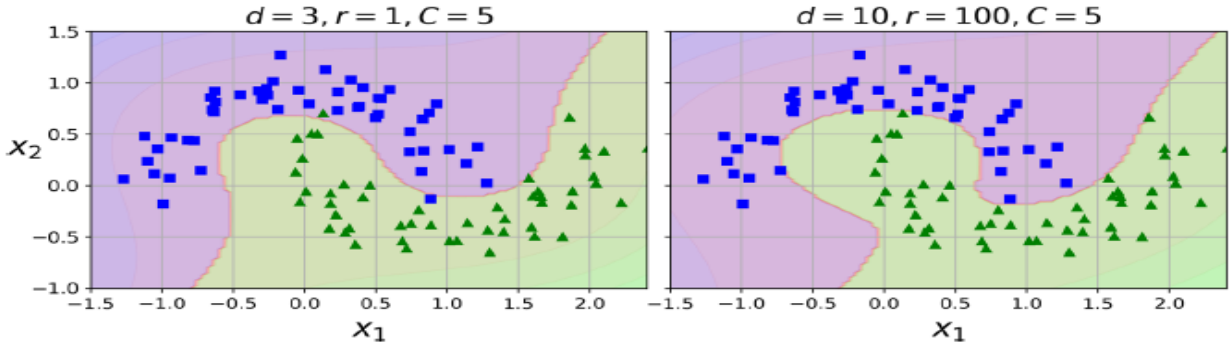


Εικόνα 5.6: Απεικόνιση SVM [15]

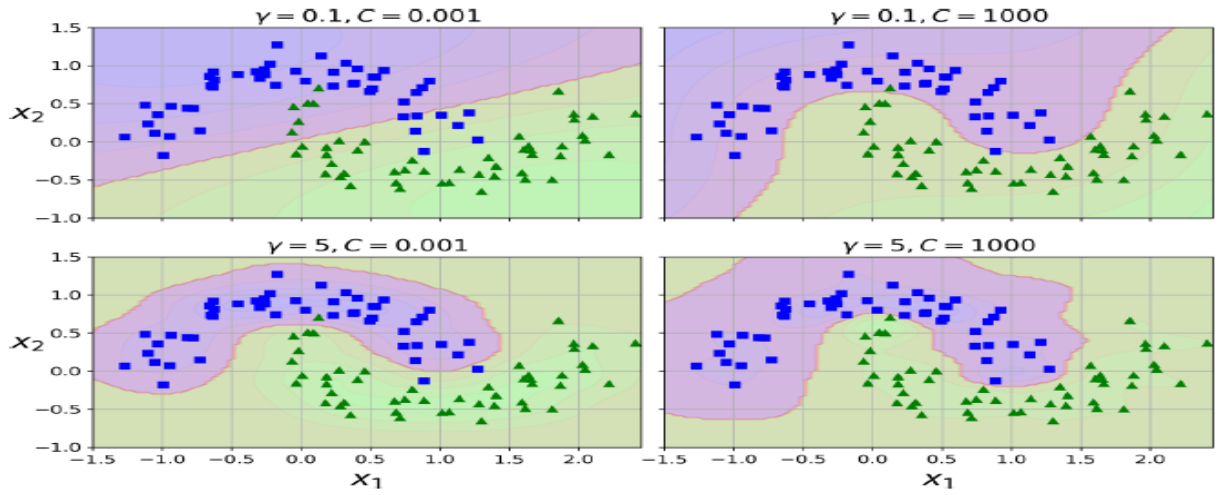
Όμως, είναι σύνηθες να εντοπίζεται παραβίαση των ορίων ή ύπαρξη διαφορετικών δεδομένων στις κατηγορίες. Για την αποφυγή αυτών, είναι ανάγκη τα δεδομένα να είναι στην ίδια κλίμακα και οι υπερπαράμετροι του μοντέλου να ρυθμίζονται σωστά.

Παρόλο που οι αλγόριθμοι SVM, αποδίδουν καλά σε γραμμικά προβλήματα, στην πραγματικότητα τα περισσότερα προβλήματα είναι μη γραμμικά. Αυτό λύνεται χρησιμοποιώντας συναρτήσεις πυρήνα (kernel functions) και η τεχνική αυτή ονομάζεται «κόλπο συνάρτησης του πυρήνα» (kernel function trick). Οι συναρτήσεις αυτές μπορεί να είναι πολυωνυμικές, σιγμοειδείς, γραμμικές, μη γραμμικές ή η συνάρτηση βάσης-ακτίνας του Gauss για τα νευρωνικά δίκτυα

(Gaussian Radial Basis Function-RBF). Οι συναρτήσεις αυτές, μετατρέπουν τα μη γραμμικά δεδομένα σε μορφή, τέτοια ώστε να αναπαρασταθεί σε έναν χώρο πολλών διαστάσεων κάνοντας το πρόβλημα κατηγοριοποίησης γραμμικό.

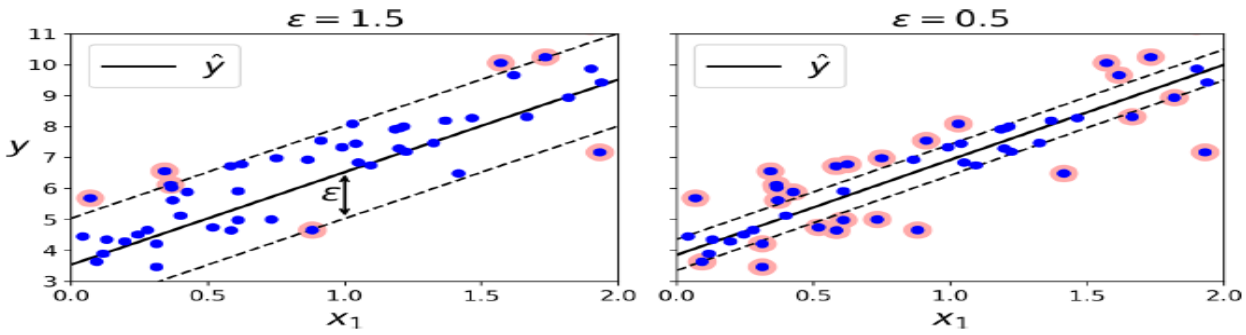


Εικόνα 5.7: SVM με πολυωνυμική συνάρτηση πυρήνα



Εικόνα 5.7: SVM με συνάρτηση πυρήνα RBF [3]

Όπως προαναφέρθηκε, οι αλγόριθμοι SVM είναι ευέλικτοι. Όπως υποστηρίζουν γραμμική και μη γραμμική κατηγοριοποίηση, το ίδιο ισχύει και για την παλινδρόμηση. Για να χρησιμοποιηθούν οι SVM στην παλινδρόμηση, πρέπει να γίνει το αντίθετο από ότι στην κατηγοριοποίηση. Δηλαδή, αντί να αποφεύγεται η είσοδος των δεδομένων στα υπερεπίπεδα χωρίς να γίνεται παραβίαση του ορίου απόφασης, η παλινδρόμηση SVM (Support Vector Regression-SVR), επιδιώκει να χωρέσει όσο πιο πολλά δεδομένα μπορεί, με όσο το δυνατόν λιγότερες παραβιάσεις γίνεται.



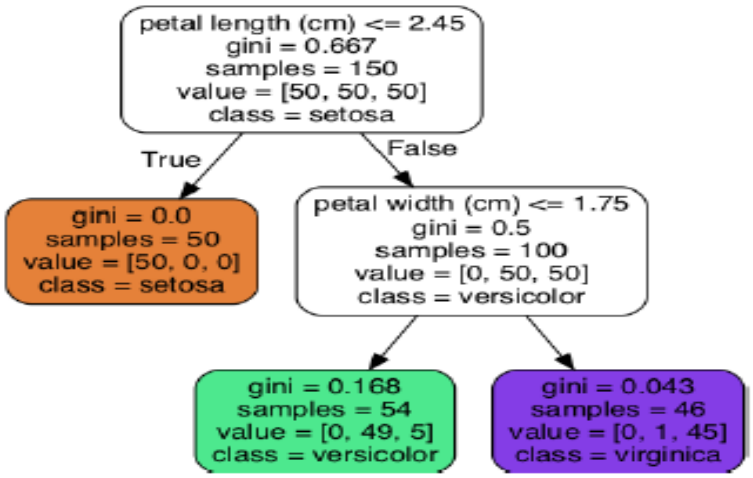
Εικόνα 5.8: Support Vector Regression [3]

5.5 Δέντρο Απόφασης (Decision Tree)

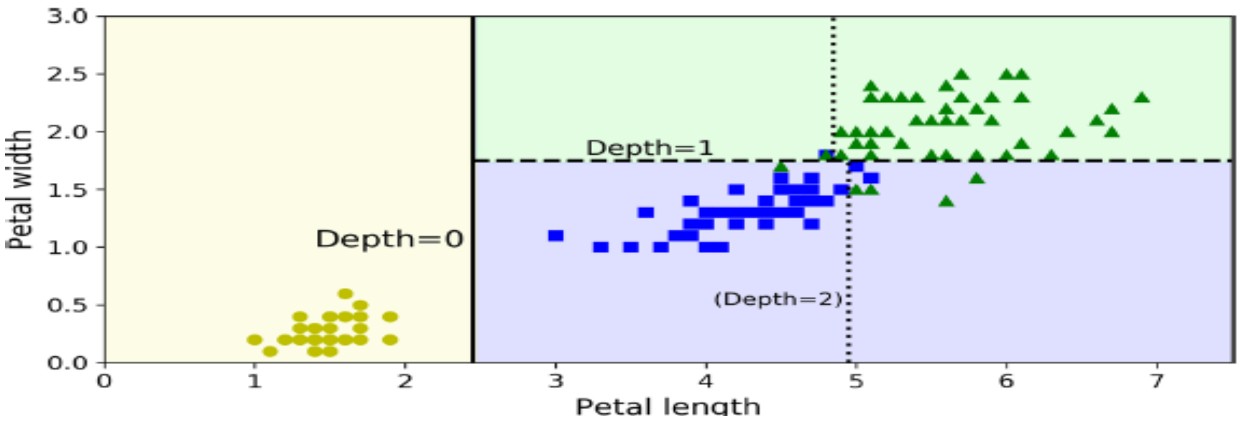
Τα δέντρα απόφασης (decision trees) είναι μοντέλα που βασίζονται σε μια σειρά κανόνων. Αυτοί οι κανόνες έχουν την λογική συνθήκη AN-TOTE (IF-THEN), κάνοντας τα πολύ εύκολα στην κατανόηση και στην εφαρμογή.

Τα δέντρα απόφασης, μοιάζουν με ένα διάγραμμα ροής, όπου ο κόμβος στην κορυφή αποτελεί την ρίζα του δέντρου και περιέχει όλα τα δεδομένα με έναν αρχικό έλεγχο, έπειτα κάθε κόμβος αποτελεί έναν έλεγχο για μια μεταβλητή, κάθε «κλαδί» την απάντηση σε αυτό τον έλεγχο, η οποία είναι το λογικό «ΝΑΙ» ή «ΟΧΙ». Όσο μεγαλώνει το δέντρο τα δεδομένα κατηγοριοποιούνται σε αυτά τα κλαδιά και όταν όλα τα δεδομένα σε ένα κλαδί ανήκουν στην ίδια κατηγορία, καταλήγουν σε έναν κόμβο με το όνομα της κατηγορίας και τότε σταματάει η διαδικασία. Όταν, υπάρχουν πολλές κατηγορίες, δημιουργούνται επιπλέον κόμβοι με βάση μια άλλη μεταβλητή, πραγματοποιώντας κάθε πιθανό συνδυασμό και η διαδικασία επαναλαμβάνεται μέχρι να κατηγοριοποιηθούν όλα τα δεδομένα. Επιπλέον, με αυτή ακριβώς την λογική τα δέντρα απόφασης πραγματοποιούν προβλέψεις.

Τα δέντρα απόφασης χωρίζονται μπορούν να διαχωρίζουν τα δεδομένα με δύο τρόπους, τον Gini Impurity και τον Entropy. Από προεπιλογή, χρησιμοποιείται ο Gini και η πραγματικότητα είναι ότι δεν έχουν ιδιαίτερη διαφορά στην απόδοση. Αν και τα δέντρα Gini είναι λίγο πιο γρήγορα στο να υπολογιστούν, τείνουν να απομονώνουν την πιο συχνή κατηγορία σε ένα κλαδί του δέντρου ενώ, τα δέντρα Entropy είναι πιο ισορροπημένα.



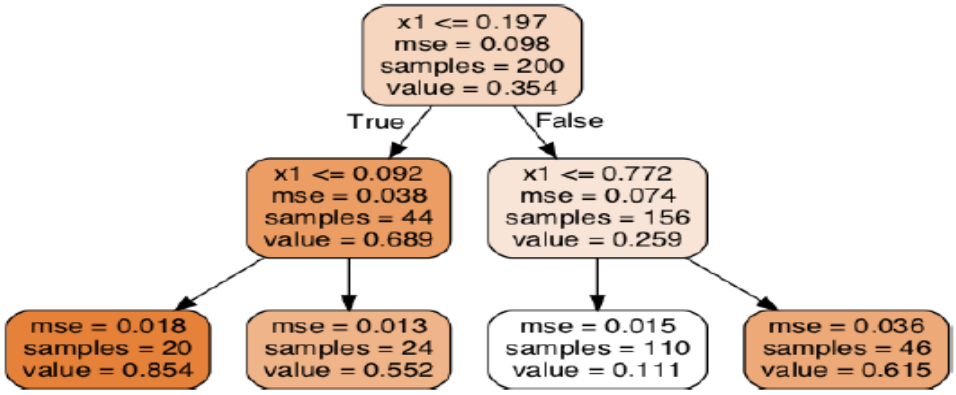
Εικόνα 5.9: Δέντρο Απόφασης στο Iris dataset [3]



Εικόνα 5.10: Πρόβλεψη Δέντρου Απόφασης-Όρια απόφασης [3]

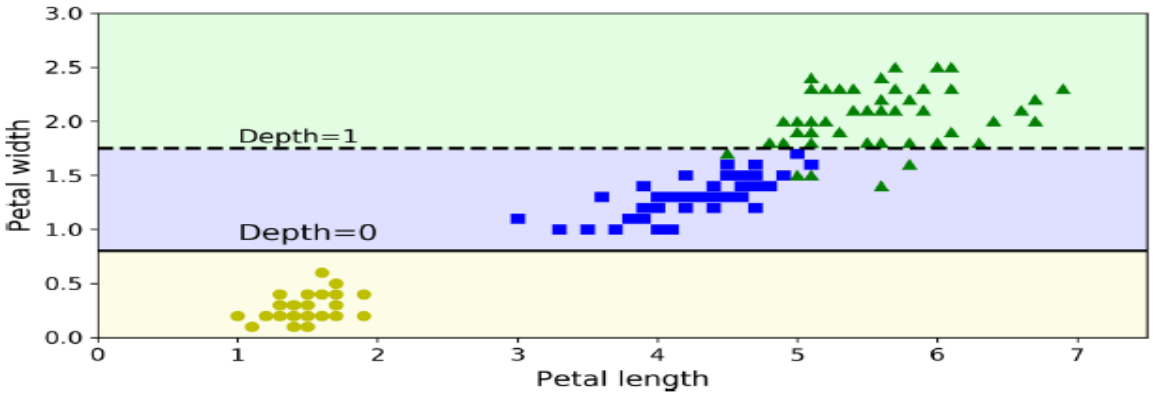
Εξαιτίας της δύναμης τους, τα δέντρα απόφασης αν δεν ελέγχονται και δεν έχουν περιορισμούς, θα προσαρμοστούν στα δεδομένα εκπαίδευσης με αποτέλεσμα να το υπερμοντελοποιήσει (overfitting). Για να αποφευχθεί αυτό πρέπει να προσδιοριστούν οι υπερπαράμετροι του δέντρου πριν την διαδικασία εκπαίδευσης. Ο περιορισμός των δέντρων μπορεί να γίνει και μετά την διαδικασία εκπαίδευσης, κλαδεύοντας τα δέντρα με την χρήση στατιστικών ελέγχων, όπως ο έλεγχος χ^2. Οι έλεγχοι αυτοί υπολογίζουν κατά πόσο ένας κόμβος του δέντρου συμβάλει στην «καθαρότητα» του αποτελέσματος και αν δεν υπάρχει ιδιαίτερη διαφορά, τον διαγράφουν μαζί τους κόμβους που είναι από κάτω.

Τα δέντρα εκτός από κατηγοριοποίηση μπορούν να πραγματοποιήσουν και παλινδρόμηση. Η λογική και η διαδικασία παραμένουν οι ίδιες, μόνο που εδώ γίνεται πρόβλεψη για μία τιμή σε κάθε κόμβο.



Εικόνα 5.11: Παλινδρόμηση με Δέντρο Απόφασης [3]

Αν και τα δέντρα απόφασης μοιάζουν το ιδανικό μοντέλο για κάθε αναλυτή, έχουν τις αδυναμίες τους. Η πιο σημαντική είναι ότι επηρεάζονται πολύ, από την παραμικρή αλλαγή υπάρξει στα δεδομένα εκπαίδευσης. Αν για παράδειγμα αφαιρέσουμε μια καταχώρηση από το μοντέλο της Εικόνας 5.9 θα έχουμε το αποτέλεσμα της Εικόνας 5.11.



Εικόνα 5.12: Ευαισθησία στις αλλαγές των δεδομένων εκπαίδευσης [3]

Ένας ακόμη λόγος που τα δέντρα απόφασης είναι τόσο σημαντικά στην Μηχανική Μάθηση, είναι διότι σε αυτά βασίζεται ένα από τα πιο δυνατά εργαλεία της, το Τυχαίο Δάσος (Random Forest).

Το Τυχαίο Δάσος ουσιαστικά, φτιάχνει πολλά δέντρα απόφασης τυχαία παίρνοντας δείγματα δεδομένων που αργότερα αντικαθίσταται. Οι τελικές αποφάσεις βασίζονται, σε «ψήφους» ανάμεσα στα δέντρα απόφασης. Στόχος αυτού του αλγόριθμου είναι να αυξήσει την ακρίβεια και να αποφύγει όσο το δυνατόν περισσότερο το overfitting.

Μέρος Β: Υλοποίηση

Κεφάλαιο 6: Εισαγωγή στα δεδομένα και τις πλατφόρμες λογισμικού

6.1 Εισαγωγή

Πραγματοποιήθηκε η μελέτη του dataset που χρησιμοποιήθηκε στην εργασία αυτή, εντοπίστηκαν κάποια από τα προβλήματα που μειώνουν την ποιότητα των δεδομένων (Κεφάλαιο 3) και έπειτα πραγματοποιήθηκε μια πρώτη γνωριμία με τις πλατφόρμες λογισμικού που χρησιμοποιήθηκαν, με τελικό σκοπό την αντιμετώπιση του προβλήματος των «κενών» στα δεδομένα.

6.2 Μελέτη των δεδομένων

Αρχικά έγινε μια προσπάθεια για την κατανόηση των δεδομένων ώστε, κατανοηθούν καλύτερα το πρόβλημα της πρόβλεψης της ταχύτητας του πλοίου στο νερό, οι μεταβλητές που υπάρχουν στο dataset και να γίνει μια πρώτη εκτίμηση της σημαντικότητας των μεταβλητών για το τελικό αποτέλεσμα.

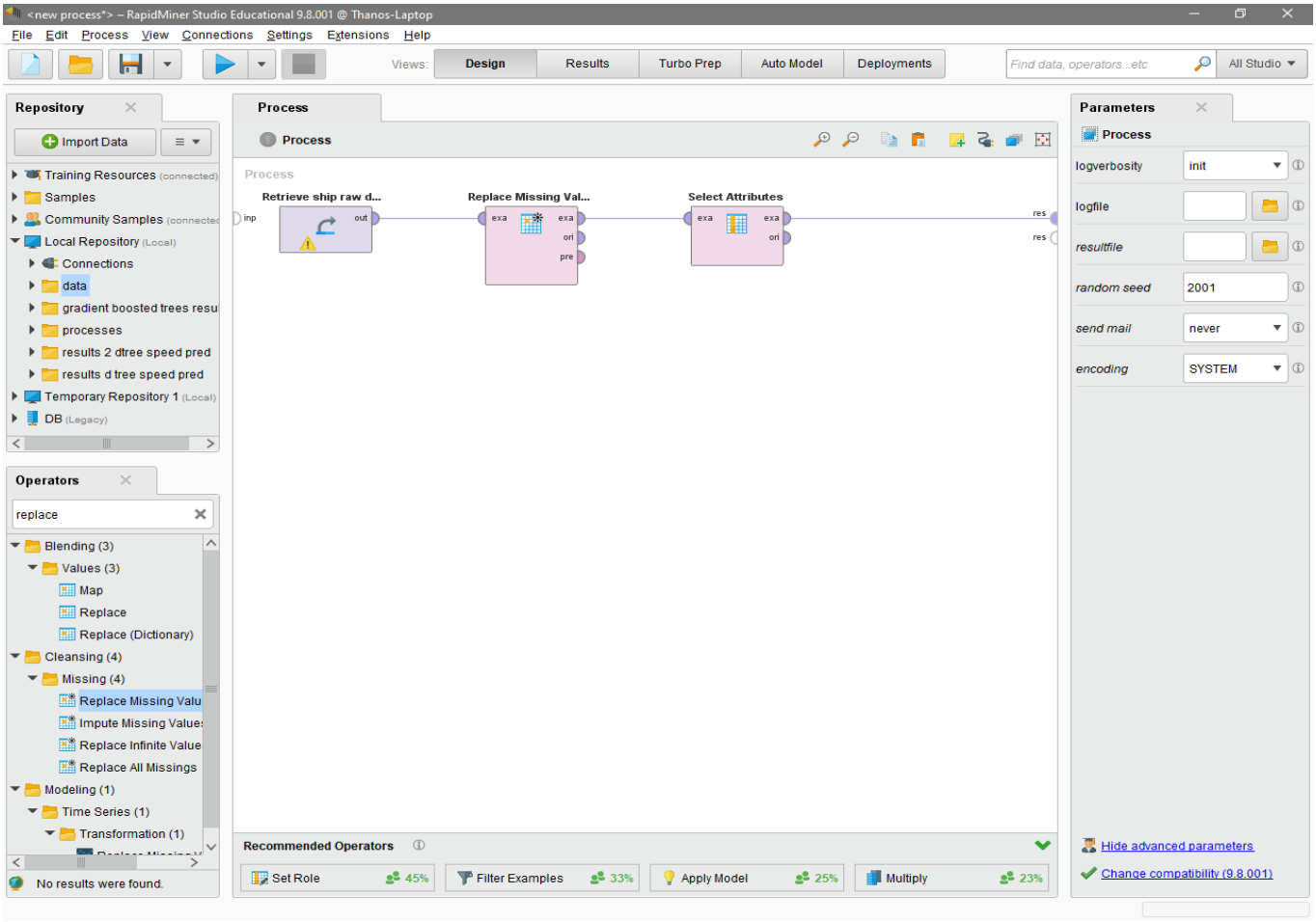
Παρατηρήθηκαν αρχικά 179 μεταβλητές, από τις οποίες κάποιες σχετίζονται σε μεγάλο βαθμό με το πρόβλημα, όπως μπορούμε να συμπεράνουμε με γνώμονα την λογική. Επίσης παρατηρείται ότι πολλές μεταβλητές έχουν εσφαλμένες ή κενές καταχωρήσεις (missing values), το οποίο είναι και το πρόβλημα που καλούμε να αντιμετωπίσουμε, στα δεδομένα αυτά. Επιπλέον, θεωρήθηκε σκόπιμο να μειωθεί ο όγκος των δεδομένων καθώς πολλές από τις μεταβλητές δεν επηρεάζουν την προβλεπόμενη μεταβλητή αλλά και για να διαχειρίζονται πιο εύκολα τα δεδομένα από τον χρήστη και από το μοντέλο, όπως και για να μειωθεί ο χρόνος που απαιτείται για την εκπαίδευση του μοντέλου.

6.3 Πλατφόρμες λογισμικού

Για την εκπαίδευση και την επεξεργασία των δεδομένων έγινε η χρήση της πλατφόρμας RapidMiner και του Jupyter Notebook. Η χρήση της δεύτερης πλατφόρμας έγινε, διότι δύο από τις τεχνικές που εφαρμόστηκαν ήταν εξαιρετικά πολύπλοκο να υλοποιηθούν στην πλατφόρμα του RapidMiner η οποία χρησιμοποιήθηκε για το υπόλοιπο μέρος της εργασίας.

6.3.1 RapidMiner

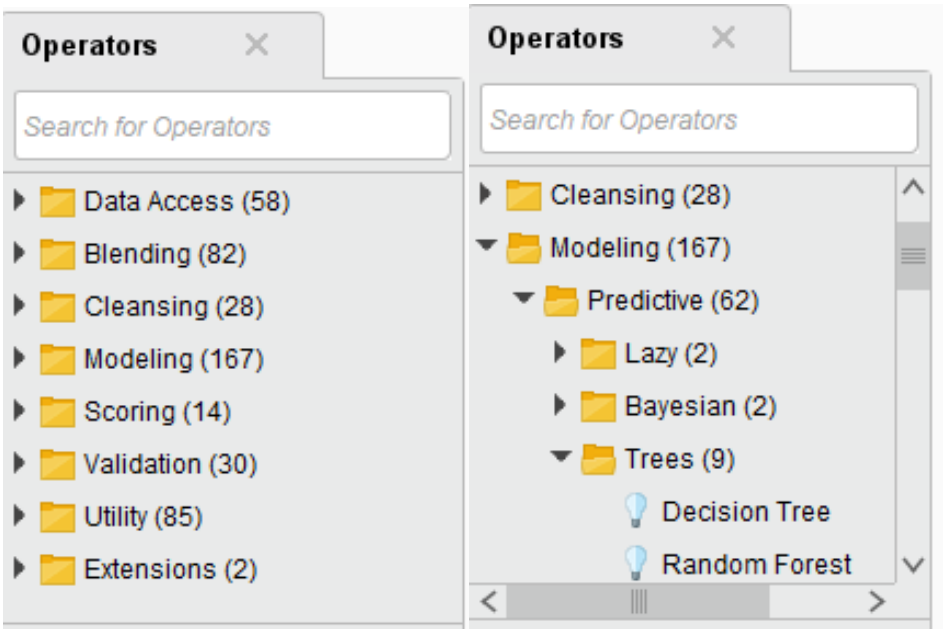
Η πλατφόρμα RapidMiner είναι ένα πρόγραμμα που χρησιμεύει στην εξόρυξη και την επεξεργασία δεδομένων και στην εκπαίδευση μοντέλων μηχανικής μάθησης και τεχνητής νοημοσύνης. Το βασικό του πλεονέκτημα είναι το γραφικό του περιβάλλον, καθώς έχει προγραμματιστεί στην γλώσσα προγραμματισμού Java και «κάτω από το καπό» χρησιμοποιεί την γλώσσα προγραμματισμού Python, αλλά με το σύστημα drag and drop που έχει ενσωματωθεί σε αυτό αλλά και με τα χρήσιμα εργαλεία που παρέχει, όπως το Turbo Prep και το Auto Model, διευκολύνει τον χρήστη σε μεγάλο βαθμό και παρέχει τα ίδια αποτελέσματα με την χρήση κάποιου code editor, με την διαφορά όμως ότι είναι πιο εύκολο να γίνει πιο εύκολα η επεξήγηση της διαδικασίας λόγω των γραφικών στοιχείων.



Εικόνα 6.1: Το περιβάλλον του RapidMiner

Ενδεικτικά, έχουν προστεθεί μερικά «blocks» που επιτελούν συγκεκριμένες λειτουργίες. Από τα αριστερά προς τα δεξιά, έχουμε το block με το οποίο εισάγουμε τα δεδομένα, έπειτα έχουμε το block με το οποίο αντικαθίστανται τα κενά στα δεδομένα (missing values) με έναν από τους τρόπους που δίνονται και στο τελευταίο block μπορεί να γίνει η επιλογή των μεταβλητών ή της μεταβλητής με βάση κάποιο κριτήριο από αυτά που υπάρχουν.

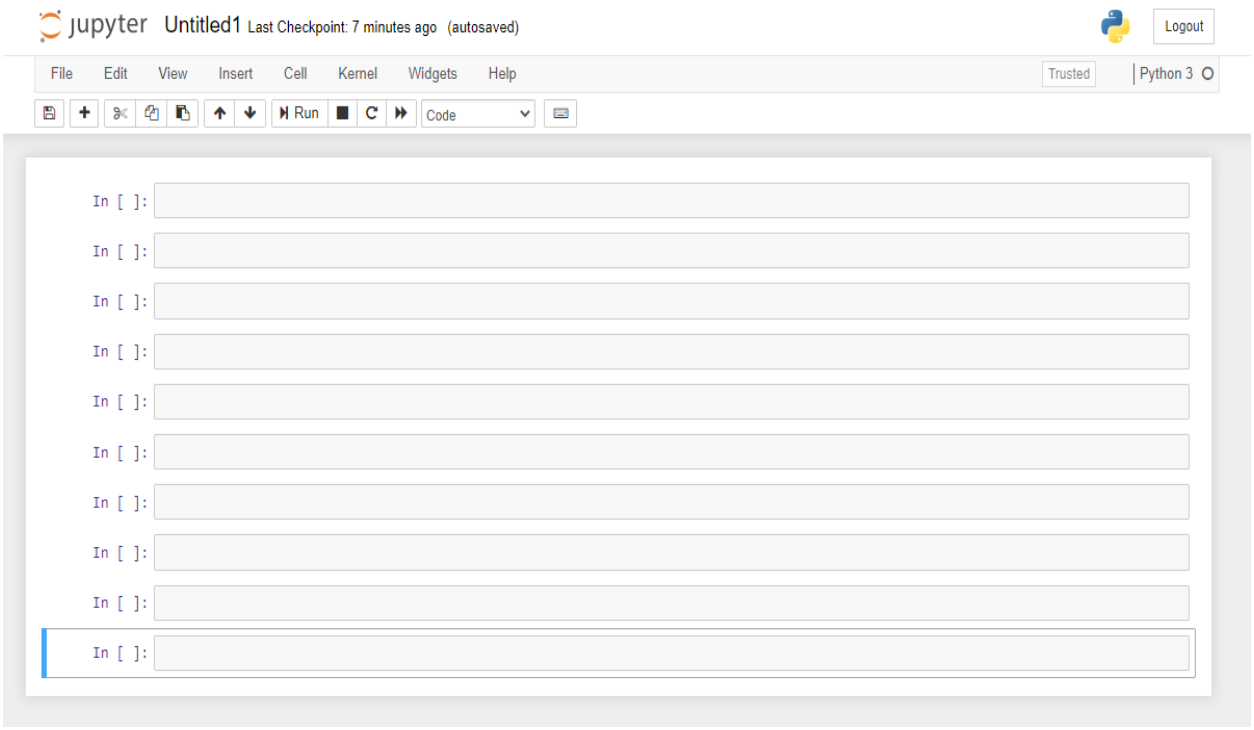
Φυσικά, εκτός από αυτά, υπάρχουν και blocks που εκτελούν λειτουργίες μοντέλων και τεχνικών επεξεργασίας, που μπορούν να ερμηνευθούν και με κώδικα προγραμματισμού. Οι κατηγορίες των λειτουργιών φαίνονται στις παρακάτω εικόνες (Εικόνα 6.2 και Εικόνα 6.3).



Εικόνες 6.2 – 6.3: Κατηγορίες blocks - Λειτουργίες RapidMiner

6.3.1 Jupyter Notebook

Το Jupyter Notebook είναι ένα ιδιαίτερα απλό περιβάλλον επεξεργασίας και εκτέλεσης κώδικα Python. Το μεγάλο θετικό σε αυτήν την πλατφόρμα είναι η απλότητα της η οποία βοηθά τον χρήστη καθώς δεν μπορεί να μπερδέψει κάποια από τις λειτουργίες που προσφέρει. Μέσα στα κελιά που δημιουργεί ο χρήστης, γράφεται ο κώδικας προς εκτέλεση και κάτω από κάθε κελί εμφανίζονται τα αποτελέσματα του κώδικα που περιέχει το κελί, αν αυτά υπάρχουν. Στην παρακάτω εικόνα (Εικόνα 6.4) απεικονίζεται το προγραμματιστικό περιβάλλον στην πλατφόρμα Jupyter Notebook.



Εικόνα 6.4: Περιβάλλον προγραμματισμού Jupyter Notebook

Κεφάλαιο 7: Επεξεργασία δεδομένων

7.1 Εισαγωγή

Μετά την μελέτη των δεδομένων, ακολούθησε η επεξεργασία τους με σκοπό την αντιμετώπιση των χαμένων τιμών που υπάρχουν σε αυτά, αλλά, συγχρόνως, και την πιο εύκολη διαχείριση τους. Σε αυτό το σημείο θα γίνει αναφορά και ανάλυση των τεχνικών επεξεργασίας που εφαρμόστηκαν στα δεδομένα με σκοπό την καλύτερη απόδοση και λειτουργία του μοντέλου.

7.2 Μείωση των δεδομένων

Το πρώτο βήμα στην επεξεργασία των δεδομένων έγινε με την δημιουργία ενός μικρότερου dataset σε σχέση με το αρχικό. Αυτό επιτεύχθηκε με την τεχνική του **Feature Ranking** (κατάταξη μεταβλητών), με την οποία εντοπίστηκαν οι μεταβλητές οι οποίες συσχετίζονται και επηρεάζουν άμεσα την μεταβλητή της ταχύτητας που θέλουμε να προβλέψουμε. Τροφοδοτώντας τα ακατέργαστα δεδομένα στο μοντέλο που επιλέχθηκε, το οποίο στην συνέχεια υπολόγισε τα «βάρη» των μεταβλητών, με κριτήριο τον βαθμό που σχετίζονται με την προβλεπόμενη μεταβλητή, το dataset έγινε μικρότερο και πιο ουσιώδες αφού έχει μεταβλητές που έχουν αντίκτυπο στην πρόβλεψη της ταχύτητας του πλοίου.

Με αυτόν τον τρόπο, έγινε μια σημαντική μείωση των δεδομένων καθώς από τις 179 μεταβλητές που υπάρχουν στο αρχικό dataset, χρησιμοποιήθηκαν 75 για την εκπαίδευση του μοντέλου με ποσοστό βάρους άνω του 19%. Η επιλογή των συγκεκριμένων ορίων, έγινε μετά από δοκιμές για περαιτέρω μείωση του μεγέθους του dataset, οι οποίες κατέληξαν είτε σε μείωση της απόδοσης είτε σε overfitting του μοντέλου.

Μερικές από τις μεταβλητές που χρησιμοποιήθηκαν μαζί με τα βάρη τους, παρουσιάζονται στον Πίνακα 7.1.

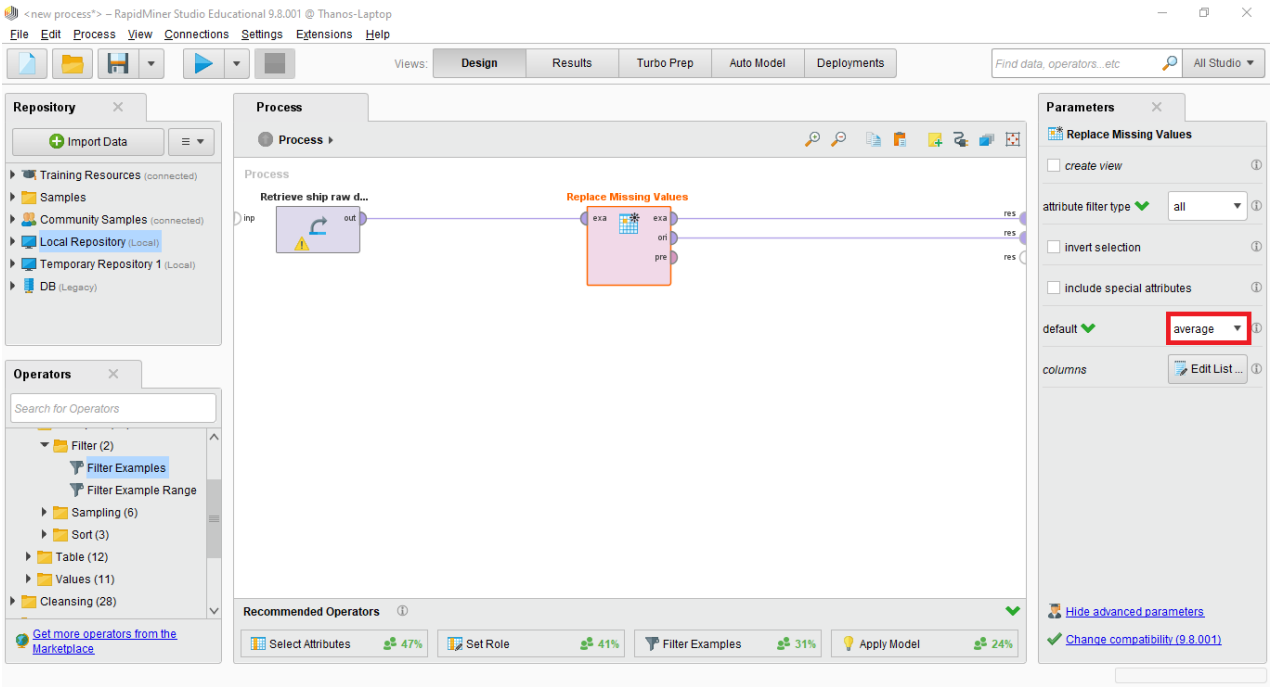
Feature	**Weight**
Longitudinal_Water_Speed	0.999
Longitudinal_Ground_Speed	0.998
Speed_Over_Ground	0.993
ME_Rpm	0.958
ME_Rpm_AMS	0.957
ME_CYL2_EXH_GAS_OULET_TEMP_AMS	0.940
ME_EXH_GAS_MEAN_TEMP_AMS	0.940
ME_CYL5_EXH_GAS_OULET_TEMP_AMS	0.939
ME_CYL1_EXH_GAS_OULET_TEMP_AMS	0.938
ME_CYL3_EXH_GAS_OULET_TEMP_AMS	0.937
ME_CYL4_EXH_GAS_OULET_TEMP_AMS	0.937
ME_Torque	0.922
ME_TC_RPM_AMS	0.919
ME_CYL6_EXH_GAS_OULET_TEMP_AMS	0.917
ME_PCO6_OUTLET_TEMP_AMS	0.898
ME_PCO2_OUTLET_TEMP_AMS	0.897
ME_TC_LO_INLET_TEMP_AMS	0.897
ME_PCO1_OUTLET_TEMP_AMS	0.895
ME_FO_Flow_Mass	0.894
ME_PCO4_OUTLET_TEMP_AMS	0.894
ME_PCO3_OUTLET_TEMP_AMS	0.892
ME_PCO5_OUTLET_TEMP_AMS	0.883
ME_Power	0.877

Πίνακας 7.1: Βάρη μεταβλητών που χρησιμοποιήθηκαν

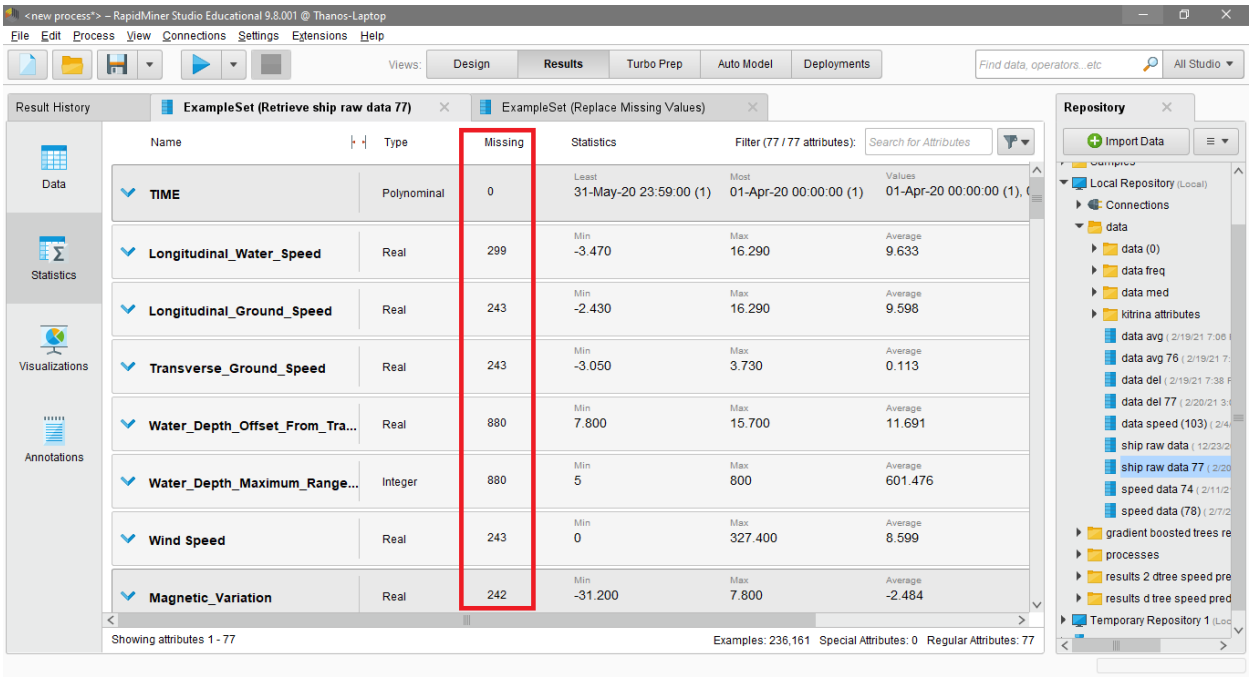
7.3 Τεχνικές συμπλήρωσης των χαμένων τιμών (missing values)

7.3.1 Αντικατάσταση με την μέση τιμή της μεταβλητής (Feature mean value)

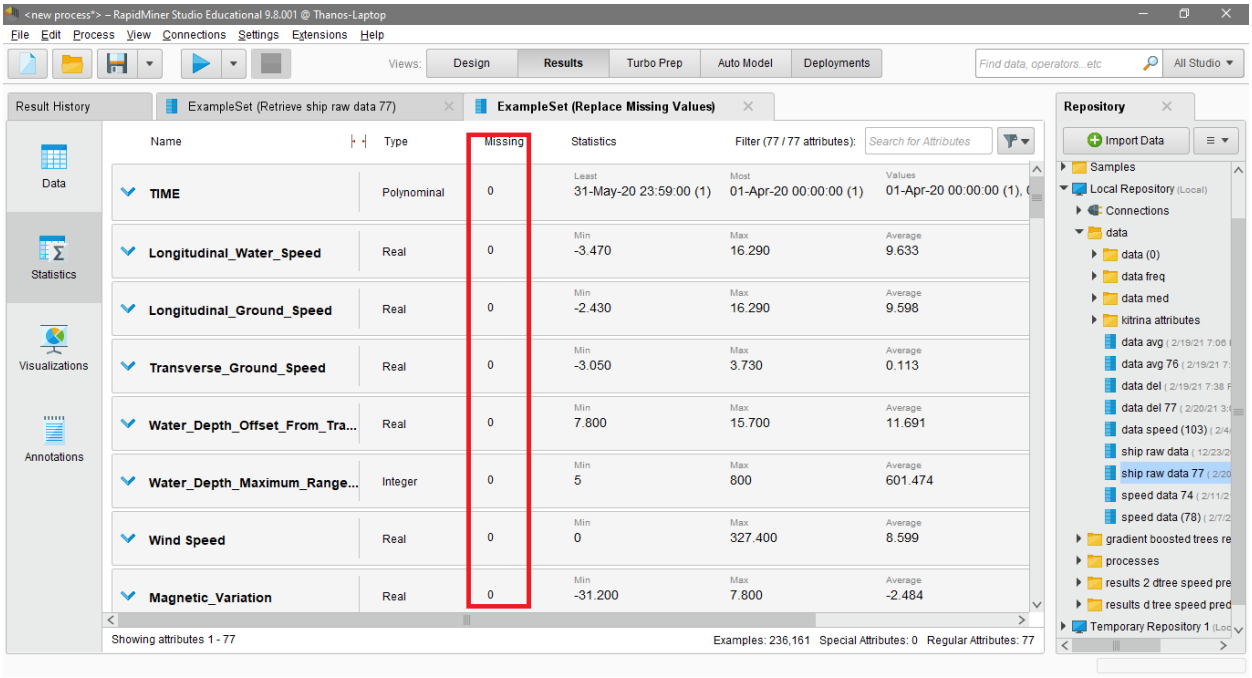
Μετά την μείωση των δεδομένων ακολούθησε η αντιμετώπιση του προβλήματος των missing values. Η πρώτη τεχνική, που εφαρμόστηκε είναι η αντικατάσταση των κενών στα δεδομένα με την μέση τιμή κάθε μεταβλητής – στήλης. Στις παρακάτω εικόνες (Εικόνα 7.1, Εικόνα 7.2 και Εικόνα 7.3) φαίνονται τα αποτελέσματα και η διαδικασία της επεξεργασίας.



Εικόνα 7.1: Διαδικασία συμπλήρωσης missing values



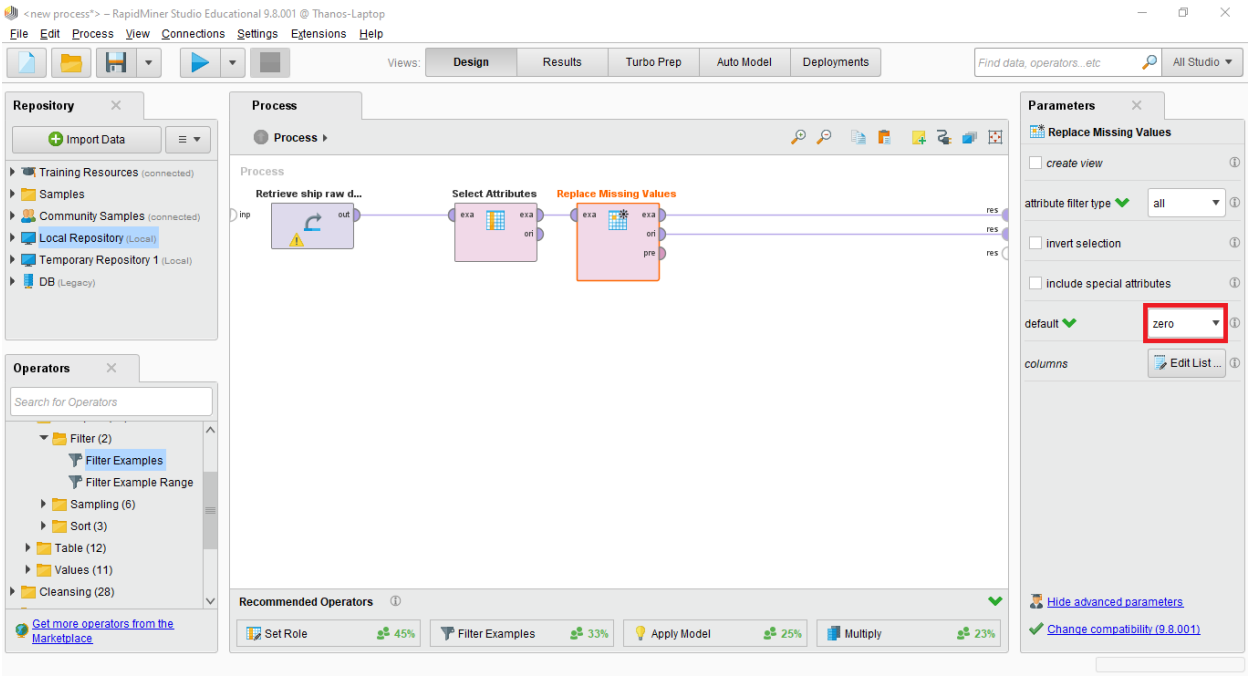
Εικόνα 7.2: Αρχικά δεδομένα μετά μείωσης



Εικόνα 7.3: Επεξεργασμένα δεδομένα μετά μείωσης

7.3.2 Αντικατάσταση με μηδενική τιμή (Zero value)

Σε αυτήν την περίπτωση η διαδικασία είναι ίδια με την προηγούμενη μόνο που αντί για την μέση τιμή αντικαταστήσαμε τις εσφαλμένες καταχωρήσεις του dataset με την τιμή μηδέν. Η μόνη διαφορά εδώ είναι ότι, εξαιρέθηκε η μεταβλητή **TIME,** η οποία είναι ονομαστικής μορφής και αντιπροσωπεύει την ημερομηνία και την ώρα που έγινε η καταχώρηση της μέτρησης. Αυτό έπρεπε να γίνει διότι, το πρόγραμμα, παρουσίαζε σφάλμα κατά την εκτέλεση. Επίσης δεν επηρεάζει καθόλου την διαδικασία αφού όπως φαίνεται και στην Εικόνα 7.4, η συγκεκριμένη μεταβλητή δεν έχει κενά στις καταχωρήσεις της.

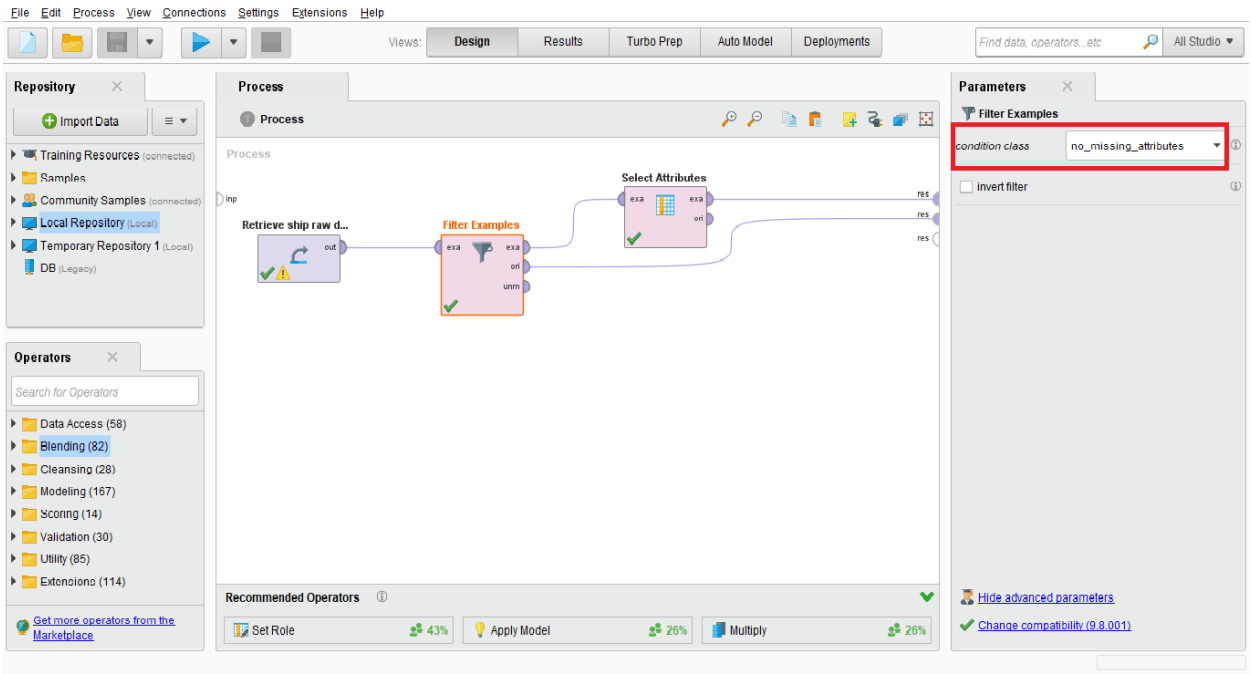


Εικόνα 7.4: Διαδικασία συμπλήρωσης missing values με μηδενικές τιμές

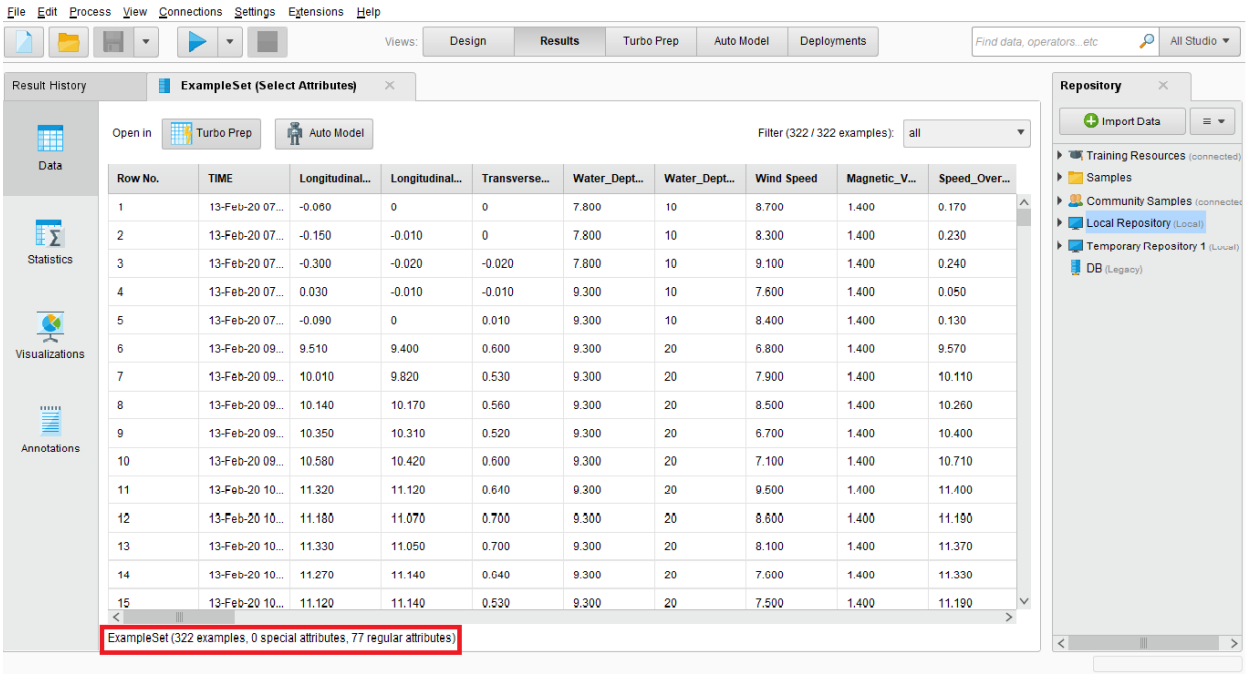
7.3.3 Διαγραφή των σειρών όπου υπάρχει κενή καταχώρηση

Με αυτή την τεχνική εντοπίζονται όλες οι τιμές που λείπουν (missing values) σε κάθε μεταβλητή και κάθε φορά διαγράφεται ολόκληρη η σειρά από όλες τις μεταβλητές, ασχέτως αν έχουν σωστή ή λάθος καταχώρηση. Η διαδικασία ξεκίνησε την αφαίρεση των γραμμών στις οποίες εντοπίστηκε κάποια λάθος καταχώρηση και στη συνέχεια έγινε η επιλογή των μεταβλητών οποίες συμμετείχαν στην εκπαίδευση. Αποτέλεσμα αυτής, είναι η ακόμη μεγαλύτερη μείωση του dataset, όχι μόνο σε

σχέση με το αρχικό αλλά και σε σχέση με την ήδη μικρότερη εκδοχή , καθώς αυτή η εκδοχή του dataset περιέχει μόλις 322 καταχωρήσεις έναντι των 236,161 του μειωμένου dataset.



Εικόνα 7.5: Διαγραφή σειρών με missing values



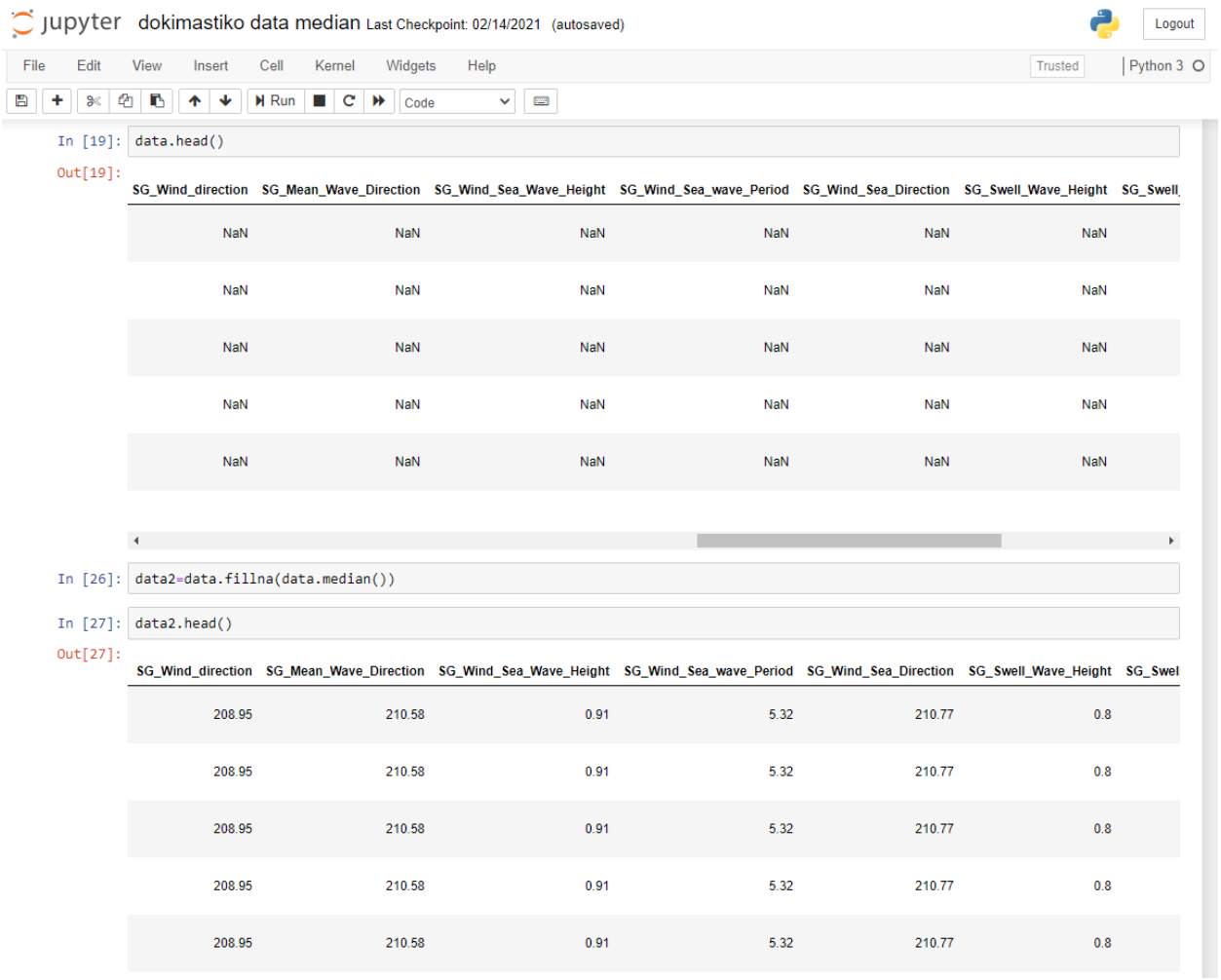
Εικόνα 7.6: Dataset μετά την διαγραφή

7.3.4 Αντικατάσταση με την μεσαία τιμή της μεταβλητής (Feature median value)

Η συγκεκριμένη τεχνική εφαρμόστηκε στο Jupyter Notebook και όπως στην περίπτωση της μέσης τιμής έτσι και εδώ οι χαμένες τιμές αντικαταστάθηκαν από την μεσαία τιμή της στήλης – μεταβλητής, δηλαδή από την τιμή η οποία αποτελεί την «μέση» στην σειρά των καταχωρήσεων, του dataset. Ο κώδικας φαίνεται στην Εικόνα 6.10 παρακάτω.

```
import pandas as pd
import chardet
rawdata = open('data\ship_raw_data.csv', 'rb').read()
result = chardet.detect(rawdata)
charenc = result['encoding']
print(charenc)data=pd.read_csv('data\ship_raw_data.csv', ';', encoding= 'ascii')
data.info()
data.median()
import numpy as np
data.head()
data2=data.fillna(data.median())
data2.head()
data2.to_csv('data\data_median.csv')
```

Εικόνα 7.7: Κώδικας για την εύρεση της μεσαίας τιμής και συμπλήρωση των missing values



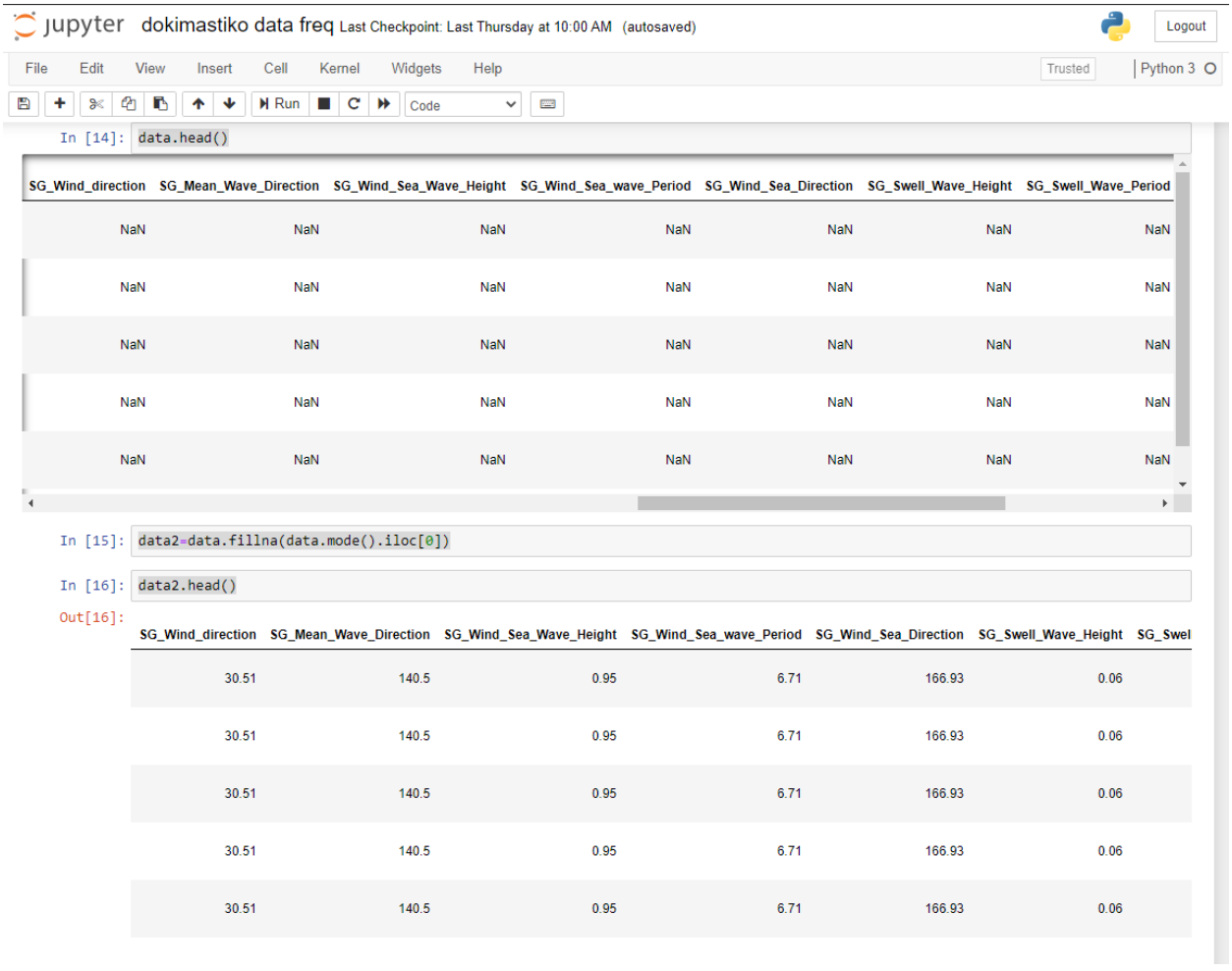
Εικόνες 7.8: Αποτελέσματα με την αντικατάσταση της μεσαίας τιμής

7.3.4 Αντικατάσταση με την πιο συχνά εμφανιζομένη τιμή της μεταβλητής (Most frequent value)

Και σε αυτήν την τεχνική χρησιμοποιήθηκε το Jupyter Notebook. Εδώ οι χαμένες τιμές αντικαταστάθηκαν από την πιο συχνά εμφανιζόμενη τιμή της κάθε στήλης.

```
import pandas as pd
import chardet
rawdata = open('data\ship_raw_data.csv', 'rb').read()
result = chardet.detect(rawdata)
charenc = result['encoding']
print(charenc)data=pd.read_csv('data\ship_raw_data.csv', ';', encoding= 'ascii')
data.info()
import numpy as np
data.head()
data2=data.fillna(data.mode().iloc[0])
data2.head()
data2.to_csv('data\data_freq.csv')
```

Εικόνα 7.9: Κώδικας για την εύρεση της πιο συχνά εμφανιζόμενης τιμής και συμπλήρωση των missing values



SG_Wind_direction	SG_Mean_Wave_Direction	SG_Wind_Sea_Wave_Height	SG_Wind_Sea_wave_Period	SG_Wind_Sea_Direction	SG_Swell_Wave_Height	SG_Swell_Wave_Period
NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN

SG_Wind_direction	SG_Mean_Wave_Direction	SG_Wind_Sea_Wave_Height	SG_Wind_Sea_wave_Period	SG_Wind_Sea_Direction	SG_Swell_Wave_Height	SG_Swel
30.51	140.5	0.95	6.71	166.93	0.06	
30.51	140.5	0.95	6.71	166.93	0.06	
30.51	140.5	0.95	6.71	166.93	0.06	
30.51	140.5	0.95	6.71	166.93	0.06	
30.51	140.5	0.95	6.71	166.93	0.06	

Εικόνες 7.10: Αποτελέσματα με την αντικατάσταση της πιο συχνά εμφανιζόμενης τιμής

Κεφάλαιο 8: Επιλογή, εκπαίδευση και αξιολόγηση του μοντέλου

8.1 Εισαγωγή

Σε αυτό το κεφάλαιο γίνεται ο προσδιορισμός του μοντέλου που χρησιμοποιήθηκε στην εφαρμογή, το κριτήριο απόδοσης που επιλέχθηκε για την απόδοση του και η περιγραφή της διαδικασίας δημιουργίας του μοντέλου. Επίσης παρουσιάζονται, τα αποτελέσματα του μοντέλου, με βάση το επιλεγμένο κριτήριο.

8.2 Το μοντέλο και το κριτήριο απόδοσης που εφαρμόστηκαν

Το μοντέλο που επιλέχθηκε για να υλοποιηθεί η συγκεκριμένη εφαρμογή πρόβλεψης, είναι το Δέντρο Απόφασης (Decision Tree). Όπως έχει αναφερθεί ήδη (Ενότητα 5.5), το μοντέλο του δέντρου απόφασης μπορεί να εφαρμοστεί σε πλήθος εφαρμογών και είναι αρκετά απλό στην λειτουργία του, κάνοντας το εύκολο στην κατανόηση. Επίσης, τα Δέντρα Απόφασης δεν έχουν υψηλές απαιτήσεις στην προεπεξεργασία των δεδομένων, γεγονός που κάνει πολύ πιο εύκολη την εκτέλεση της εφαρμογής που θέλουμε να κάνουμε.

Επίσης, ως μέτρο απόδοσης χρησιμοποιήθηκε το μέσο τετραγωνικό σφάλμα ή Root Mean Squared Error (RMSE). Στα προβλήματα παλινδρόμησης – πρόβλεψης, είναι ένα από τα συχνότερο μέτρα που χρησιμοποιούνται για την αξιολόγηση των μοντέλων και εκφράζει το ποσοστό λάθους που κάνει το μοντέλο στις προβλέψεις του. Η εξίσωση υπολογισμού του RMSE είναι: $RMSE(X, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$ (Εξίσωση 8.1).

8.3 Η εκπαίδευση του Δέντρου Απόφασης

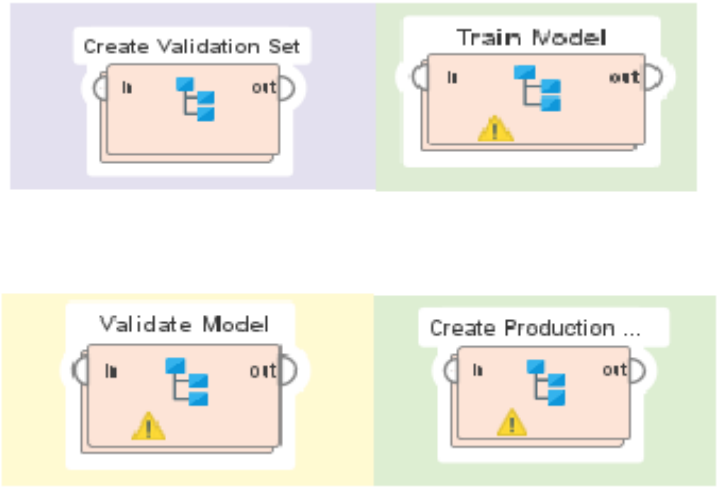
Με την επεξεργασία των δεδομένων και την συμπλήρωση των κενών με τις τεχνικές που αναφέρθηκαν στο προηγούμενο κεφάλαιο, έχουν γίνει όλες οι προετοιμασίες ώστε το μοντέλο μας να μπορεί να εκπαιδευτεί. Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε το εργαλείο Auto Model του RapidMiner, το οποίο διευκολύνει ακόμα πιο πολύ τον χρήστη. Η διαδικασία χωρίζεται σε πέντε βήματα:

- Την φόρτωση των δεδομένων

- Την επιλογή της ενέργειας (π.χ. Prediction)
- Την προετοιμασία για την επίτευξη του στόχου
- Την επιλογή των δεδομένων που θα χρησιμοποιηθούν από το dataset
- Την επιλογή του μοντέλου

Εικόνα 8.1: Βήματα Auto Model

Πιο αναλυτικά τα μπλοκ που κάνουν τα παραπάνω φαίνονται στην Εικόνα 8.2.



Εικόνα 8.2: Blocks – Βήματα δημιουργίας και εκπαίδευσης του Decision Tree

8.4 Απόδοση του μοντέλου

Τα αποτελέσματα του συγκεκριμένου μοντέλου Decision tree για κάθε εκδοχή του dataset, που χρησιμοποιήθηκε, είχε τα αποτελέσματα που φαίνονται στον Πίνακα 8.1.

Τεχνική Αντικατάστασης Missing Values	**RMSE**
Μέση τιμή μεταβλητής	0.219 +/- 0.032
Διαγραφή σειρών όπου υπάρχει χαμένη τιμή	0.362 +/- 0.060
Μεσαία τιμή μεταβλητής	0.480 +/- 0.042
Πιο συχνά εμφανιζόμενη τιμή της μεταβλητής	0.311 +/- 0.018
Αντικατάσταση κενών με μηδενική τιμή	0.311 +/- 0.027

Πίνακας 8.1: Απόδοση μοντέλου για κάθε εκδοχή του dataset

Κεφάλαιο 9: Αποτελέσματα

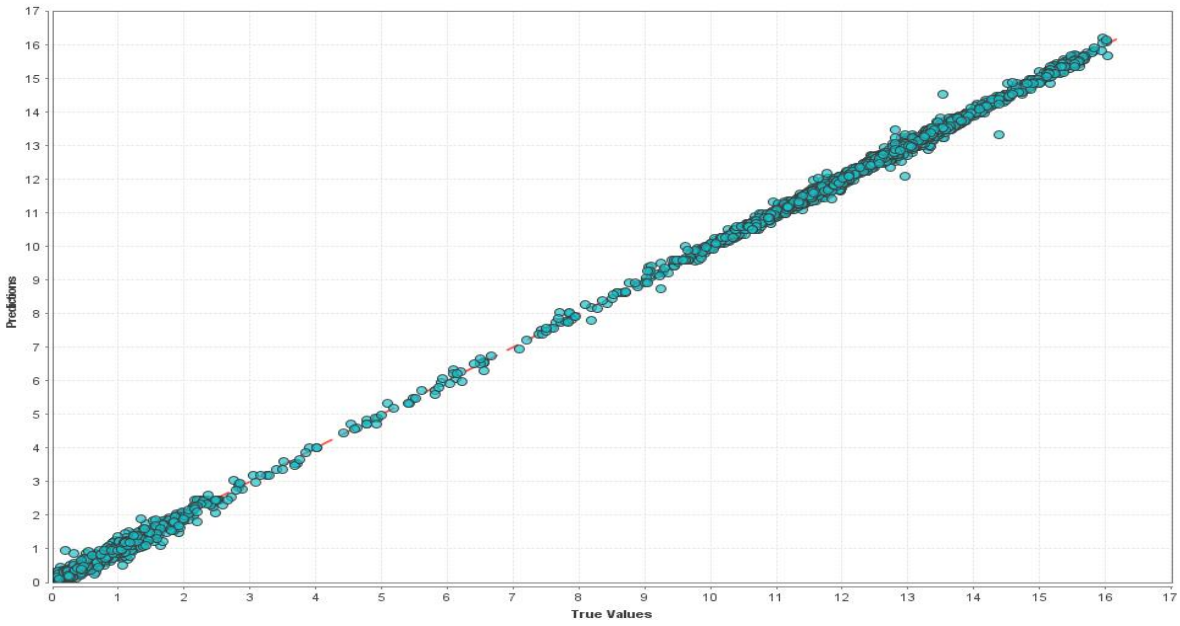
9.1 Παράθεση αποτελεσμάτων

Μετά από την εισαγωγή κάθε εκδοχής των δεδομένων στο μοντέλο παρατηρήθηκαν οι μετρήσεις του Πίνακα 9.1.

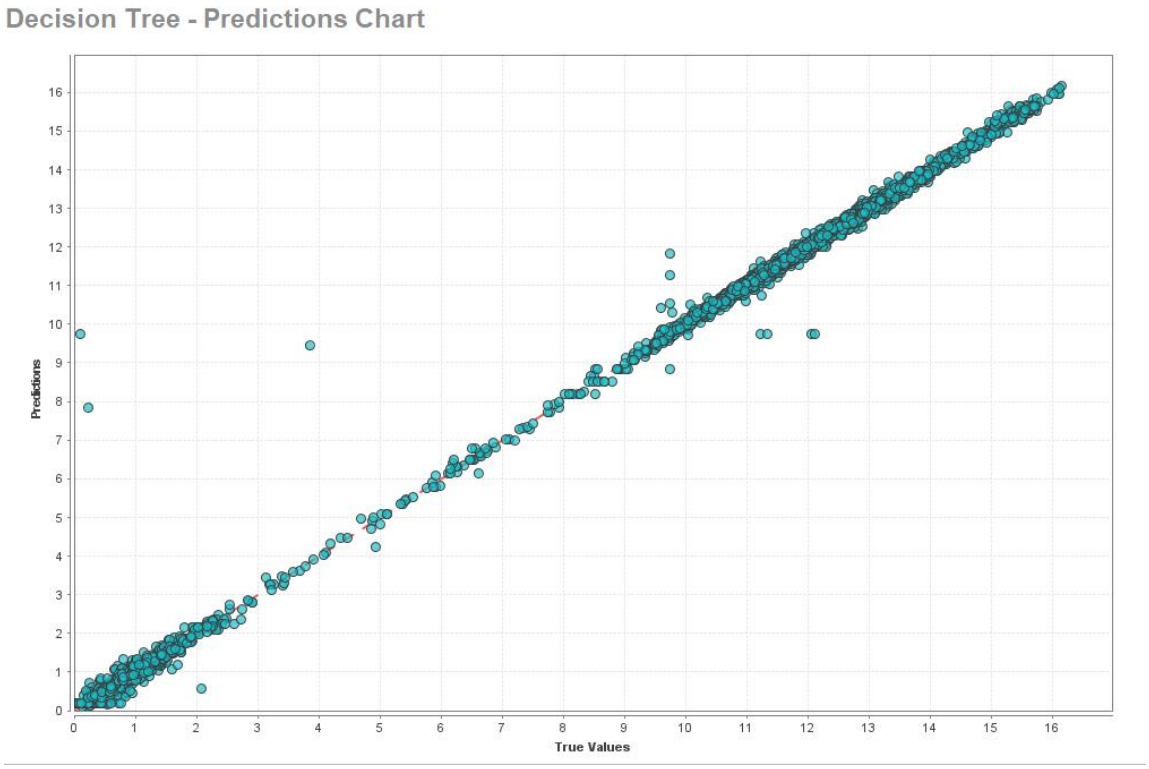
Τεχνική Αντικατάστασης Missing Values	**RMSE**	**Χρόνος**
Αρχικό dataset	0.105 +/- 0.009	01:54
Μέση τιμή μεταβλητής	0.219 +/- 0.032	01:55
Διαγραφή σειρών όπου υπάρχει χαμένη τιμή	0.362 +/- 0.060	00:05
Μεσαία τιμή μεταβλητής	0.48 +/- 0.042	02:13
Πιο συχνά εμφανιζόμενη τιμή της μεταβλητής	0.311 +/- 0.018	01:49
Αντικατάσταση κενών με μηδενική τιμή	0.311 +/- 0.027	02:14

Πίνακας 9.1: Τελικά αποτελέσματα διαδικασιών

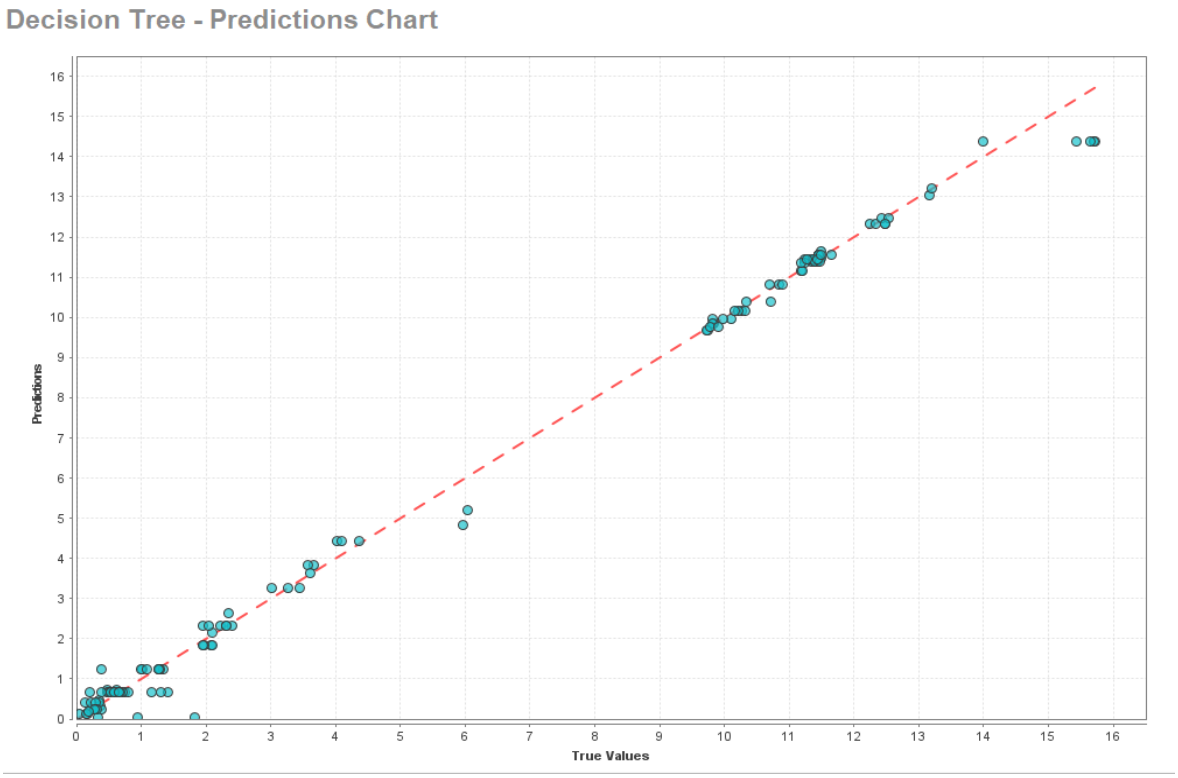
Παρακάτω παρατίθενται, τα γραφήματα προβλέψεων για κάθε περίπτωση ώστε να είναι πιο εύκολη η κατανόηση των αποτελεσμάτων.



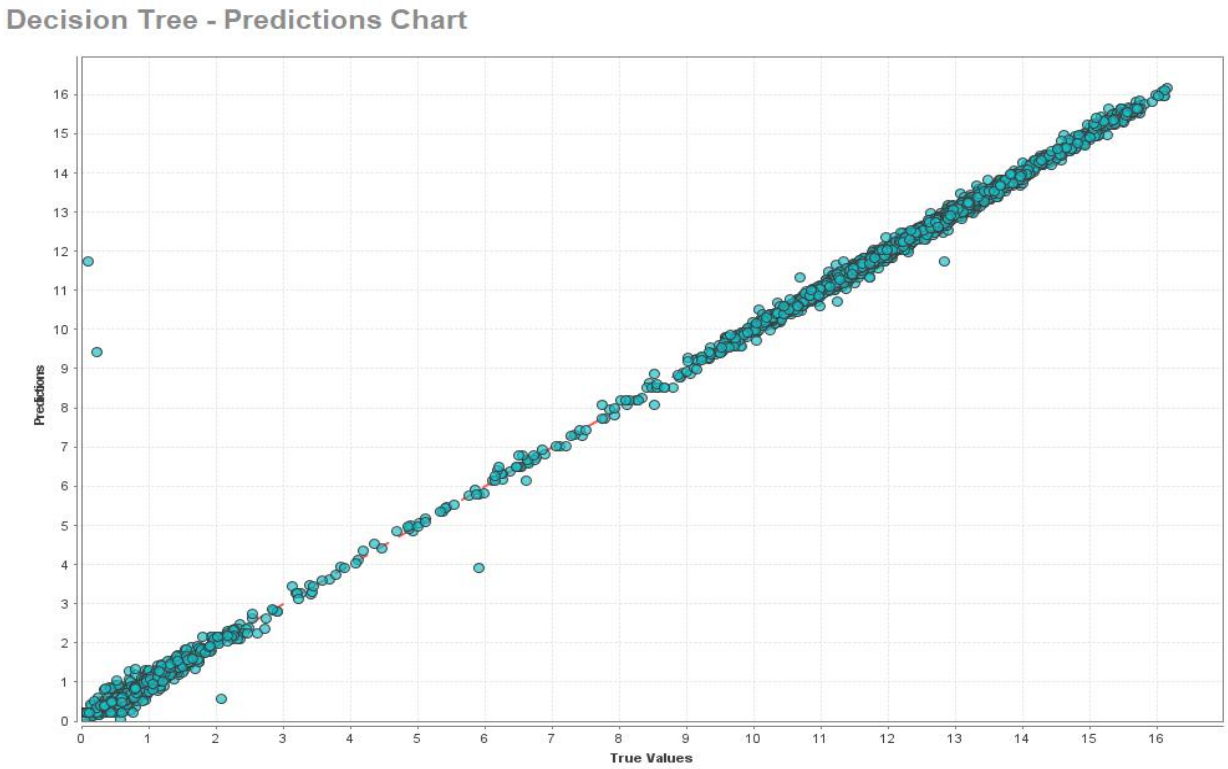
Εικόνα 9.1: Γράφημα προβλέψεων για το αρχικό dataset



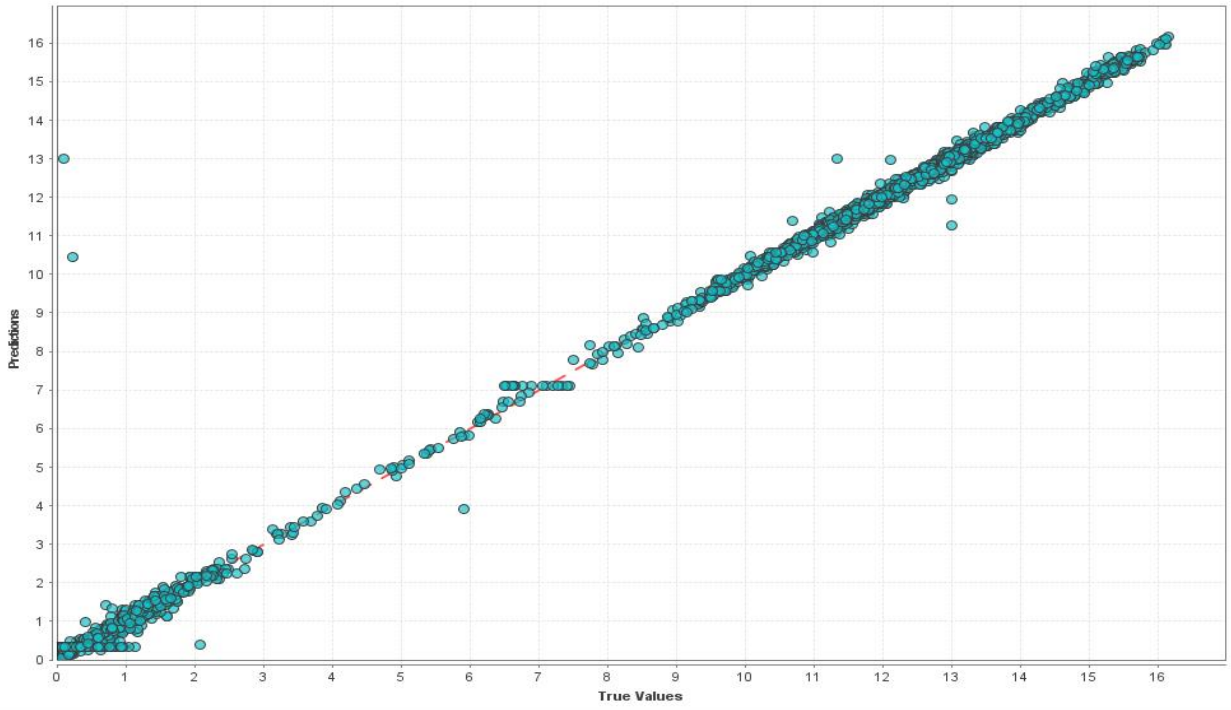
Εικόνα 9.2: Γράφημα προβλέψεων για την μέση τιμή



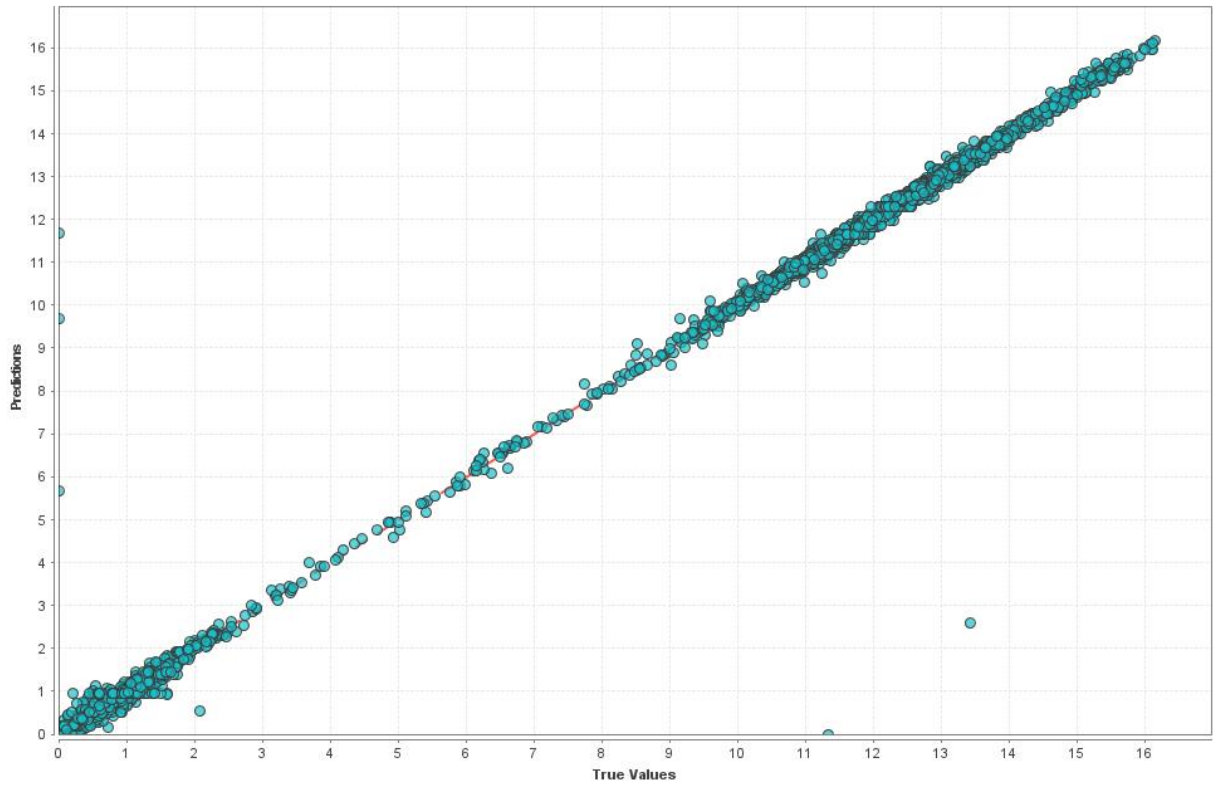
Εικόνα 9.3: Γράφημα προβλέψεων 'Διαγραφή σειρών όπου υπάρχει χαμένη τιμή'



Εικόνα 9.4: Γράφημα προβλέψεων 'Αντικατάσταση με την μεσαία τιμή'



Εικόνα 9.5: Γράφημα προβλέψεων 'Αντικατάσταση με την πιο συχνά εμφανιζόμενη τιμή'



Εικόνα 9.6: Γράφημα προβλέψεων 'Αντικατάσταση με μηδενική τιμή'

9.2 Συμπεράσματα

Ξεκινώντας από το αρχικό dataset παρατηρούμε ότι το ποσοστό σφάλματος και ο χρόνος που χρειάστηκε είναι πολύ μικρότερος συγκριτικά με τις υπόλοιπες τεχνικές και οι προβλέψεις βλέπουμε ότι ταιριάζουν αρκετά με την γραμμή πρόβλεψης παρά τις χαμένες τιμές που υπάρχουν στα δεδομένα. Όμως, αυτό, είναι μια περίπτωση overfitting του μοντέλου στα δεδομένα το οποίο παρατηρείται συχνά όταν το dataset, περιέχει δεδομένα με θόρυβο.

Στη συνέχεια συγκρίνοντας, τα αποτελέσματα των υπόλοιπων μεθόδων μεταξύ τους, βλέπουμε ότι με την χρήση της μέσης τιμής πετύχαμε το χαμηλότερο δυνατό ποσοστό σφάλματος στον καλύτερο δυνατό χρόνο. Επίσης, όπως φαίνεται από το γράφημα οι περισσότερες προβλέψεις συγκλίνουν με την γραμμή πρόβλεψης με μερικές εξαιρέσεις. Οι τιμές που βρίσκονται πιο «μακριά» από την καμπύλη πρόβλεψης υπάρχουν διότι, στην αρχή της εκπαίδευσης οι προβλέψεις του μοντέλου έχουν μεγάλη διαφορά από τις πραγματικές τιμές. Η διαφορά αυτή μαζί με την γνώση που αποκτά το μοντέλο και έπειτα από μικρό χρονικό διάστημα, τελικά γίνεται όλο και μικρότερη έως ότου φτάσει στην τελική τιμή της.

9.3 Προκλήσεις

Τα παραπάνω συμπεράσματα προήλθαν, έπειτα από πολλές ώρες μελέτης των δεδομένων και προσομοιώσεων ώστε να υπάρχει φερεγγυότητα. Για να φτάσουμε σε αυτό το τελικό σημείο προηγήθηκαν πολλές και αρκετά χρονοβόρες δοκιμαστικές προσομοιώσεις στα δεδομένα που χρησιμοποιήθηκαν, ώστε να επιλεχθούν τα απαραίτητα δεδομένα για την εφαρμογή και το κατάλληλο μοντέλο που θα μπορούσε να «χειριστεί» τα δεδομένα αυτά, μέσα σε λογικά χρονικά περιθώρια. Κύριο κριτήριο για την επιλογή του μοντέλου, ήταν και οι υπολογιστικοί πόροι που ήταν διαθέσιμοι καθώς με βάση αυτούς έγιναν οι απαιτούμενες επιλογές μοντέλου, υπερπαραμέτρων και δεδομένων. Με αυτόν τον τρόπο η ζητούμενη εφαρμογή προσαρμόστηκε στις δυνατότητες του υπολογιστή που χρησιμοποιήθηκε για αυτή την μελέτη, η οποία υλοποιήθηκε με τον καλύτερο δυνατό τρόπο.

9.4 Μελλοντικές βελτιώσεις

Στον συγκεκριμένο αλγόριθμο θα μπορούσε να δοθούν περισσότερα δεδομένα ως δεδομένα εισόδου πράγμα το οποίο θα επηρέαζε σημαντικά τα τελικά αποτελέσματα σε κάθε περίπτωση. Επιπλέον αλλάζοντας την υπερπαράμετρο του «βάθους» του δέντρου αλλάζει και η κατηγοριοποίηση των δεδομένων, επομένως και οι προβλέψεις για αυτά. Ακόμη θα μπορούσε να γίνει και χρήση ενός πιο αποτελεσματικού αλγορίθμου για την εφαρμογή αυτή έχοντας βέβαια και

την ανάλογη υπολογιστική δύναμη, επιλογή που θα έφερνε αλλαγές στα αποτελέσματα σε ριζικό επίπεδο, αφού ο νέος αλγόριθμος δεν θα λειτουργεί όπως ο αλγόριθμος που χρησιμοποιήθηκε. Τέλος, αλλάζοντας την έξοδο του αλγορίθμου έτσι ώστε να υπολογίζεται η κατανάλωση καυσίμου, μπορεί να γινεί επαναπροσδιορισμός της πορείας του πλοίου χωρίς απομακρυσμένη βοήθεια, για συγκεκριμένα δεδομένα εισόδου/συνθήκες.

Βιβλιογραφία

[1] Mehmed Kantardzic - Data Mining Concepts, Models, Methods, and Algorithms, 3rd Edition, Wiley-IEEE Press, 2020

[2] Jiawei Han, Micheline Kamber, Jian Pei - Data Mining. Concepts and Techniques, 3rd Edition, Morgan Kaufman, 2012

[3] Aurélien Géron - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow_ Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O'Reilly Media, 2019

[4] Xin-She Yang - Introduction to Algorithms for Data Mining and Machine Learning, Academic Press, 2019

[5] Christos Gkerekos, Iraklis Lazakis, Gerasimos Theotokatos - Machine learning models for predicting shipmain engine Fuel Oil Consumption: A comparative study, Sep 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0029801819304561?via%3Dihub

[6] Andrea Coraddu, Luca Oneto, Francesco Baldi, Davide Anguita - Vessels fuel consumption forecast and trim optimization: A data analytics perspective, January 2017. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0029801816305571?via%3Dihub

[7] Wengang Mao, Igor Rychlik, Jonas Wallin, Gaute Storhaug - Statistical models for the speed prediction of a container ship, November 2016. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0029801816303699?via%3Dihub

[8] Fredrik Ahlgen, Maria E. Mondejar, Marcus Thern. February 2019. Predicting Dynamic Fuel Oil Consumption on Ships with Automated Machine Learning. Presented at 10th International Conference on Applied Energy (ICAE2018), 22-25 August 2018, Hong Kong, China. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610219305223

[9] RapidMiner Studio (9.8), RapidMiner. Accessed: 5/1/2021. [Online]. Available: https://rapidminer.com/

[10] Jupyter Notebook (6.0.1), Fernando Perez, Brian Granger. Accessed: 1/2/2021. [Online]. Available: https://jupyter.org/

[11] Carlota Feliu, The 15 Benefits of Data Mining, February 2019. [Online]. Available: https://blog.datumize.com/the-15-benefits-of-data-mining

[12] Εικόνα 4.1: https://towardsdatascience.com/underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6fe4a8a49dbf

[13] Εικόνα 5.4: https://scipython.com/blog/plotting-the-decision-boundary-of-a-logistic-regression-model/

[14] Εικόνα 5.5: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html

[15] Εικόνα 5.6: https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3