Υλοποίηση Διαδραστικής Εφαρμογής Ιστού για τη Δημιουργία Μοντέλων Μηχανικής Μάθησης

Implementation of an Interactive Web Application to Build Machine Learning Models

ΦΕΒΡΟΥΑΡΙΟΣ 2023

Χατζηαναστασιάδης Μιχαήλ Μάριος

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Business Analytics and Data Science

Matapae

Περίληψη

Η παρούσα διπλωματική εργασία μελετά σε βάθος τις θεμελιώδεις αρχές της Επιστήμης των Δεδομένων, της Επιχειρηματικής Ευφυΐας και της Αναλυτικής των Big Data και στα πλαίσια αυτής σχεδιάστηκε και υλοποιήθηκε η διαδραστική εφαρμογή ιστού “Matapae” για τη δημιουργία μοντέλων Μηχανικής Μάθησης. Πιο συγκεκριμένα, παρουσιάζονται οι έννοιες των εξειδικευμένων πεδίων της Επιστήμης των Δεδομένων, οι δεξιότητες που απαιτούνται και οι στρατηγικές που εφαρμόζονται για την επίλυση των προβλημάτων, αξιοποιώντας τα διαθέσιμα δεδομένα. Επίσης, αναλύονται οι ποικίλοι τύποι μοντέλων Μηχανικής Μάθησης και οι αντίστοιχοι αλγόριθμοι που χρησιμοποιούνται για την αναζήτηση εξειδικευμένων λύσεων.

Τα τελευταία χρόνια, η ραγδαία αύξηση της ποσότητας των δεδομένων καθιστά δύσκολη την ανάλυσή τους καθώς υπάρχει έλλειψη στους διαθέσιμους ανθρώπινους και υπολογιστικούς πόρους, γεγονός που ενισχύει την αξία της αυτοματοποίησης των διαδικασιών της Μηχανικής Μάθησης. Συνεπώς, όλο και περισσότερο οι επιχειρήσεις εμπιστεύονται πλατφόρμες και εργαλεία που προσφέρουν ολοκληρωμένες λύσεις διαχείρισης και ανάλυσης δεδομένων, καθώς επίσης δημιουργίας μοντέλων μηχανικής μάθησης χωρίς τη χρήση κώδικα, όπως είναι το “Google Cloud Platform” ή το “H2O.ai”.

Σε αυτή τη λογική στηρίχθηκε η ιδέα για την ανάπτυξη της διαδικτυακής εφαρμογής “Matapae”, η οποία υλοποιήθηκε σε γλώσσα προγραμματισμού R προσφέροντας ένα “User-friendly” περιβάλλον. Στον τελικό χρήστη παρέχεται η δυνατότητα να εισάγει, αποθηκεύσει και εξερευνήσει σύνολα δεδομένων, να εκπαιδεύει μοντέλα Μηχανικής Μάθησης μέσω του αλγορίθμου XGBoost και ανάλογα με το είδος του προβλήματος που θέλει να επιλύσει, να τα αποθηκεύει και να τα επεξηγεί, να τα βελτιώνει και εν τέλει να προβλέπει πάνω σε νέα δεδομένα.

Το κύριο και τελευταίο μέρος της διπλωματικής εργασίας επικεντρώνεται στο εγχειρίδιο χρήσης της εφαρμογής, δηλαδή στην αναλυτική περιγραφή και την επεξήγηση των καθορισμένων ενεργειών που μπορεί να κάνει ο τελικός χρήστης. Επιπλέον, παρατίθενται και τεκμηριώνονται τα βασικά τμήματα R κώδικα για τις κύριες λειτουργίες της εφαρμογής όπως η επικοινωνία με τη βάση δεδομένων, η αναπαράσταση διαφόρων γραφημάτων, η διαχείριση Datasets και τέλος η μεθοδολογία πρόβλεψης.

Λέξεις Κλειδιά: Διερευνητική Ανάλυση Δεδομένων, Αυτοματοποιημένη Μηχανική Μάθηση, R Shiny, XGBoost, Εποπτευόμενη Μηχανική Μάθηση

ΧΑΤΖΗΑΝΑΣΤΑΣΙΑΔΗΣ ΜΙΧΑΗΛ ΜΑΡΙΟΣ
Π.Μ.Σ. στην Αναλυτική των Επιχειρήσεων
και Επιστήμη των Δεδομένων
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Abstract

The present thesis thoroughly studies the principles of Data Science, Business Intelligence, and Big Data Analytics and based on that an interactive Web Application “Matapae” was implemented which builds Machine Learning Models. In particular, the meanings of specialized fields of Data Science have been presented as well as the required capabilities and the strategies that are applied to resolve problems using the available data. Moreover, the various model types of Machine Learning and the respective algorithms, which are used to investigate the optimal solution, have been analyzed in detail.

During the latest years, the rapid increase of the data quantity makes data analysis difficult since there is a lack of both available human and computational resources. This fact strengthens the value of automated Machine Learning processes. Hence businesses trust platforms and tools that provide data management and analysis, as well as the building of Machine learning models such as “Google Cloud Platform” or “H2O.ai”.

Based on that need, the user-friendly web application “Matapae” was developed using R programming language. Through this platform, the end-user is able to import, save and manage datasets and train Machine Learning Models using the XGBoost algorithm. Finally, the end-user can save, clarify, and improve the aforementioned models depending on the kind of problem and he is able to predict based on new datasets as well.

The main and most important part of the document addresses the user manual of the web application, in which the predefined actions that the user is able to perform are described and explained comprehensively. Furthermore, parts of the R code are documented referring to the main functions of the application such as the connection with the database, the various graphs, the management of datasets, and the methodology of prediction.

Keywords: Exploratory Data Analysis (EDA), Automated Machine Learning (AutoML), R Shiny, XGBoost, Supervised Machine Learning

CHATZIANASTASIADIS MICHAIL MARIOS
Master in Business Analytics and Data Science
UNIVERSITY OF MACEDONIA

Περιεχόμενα

Περίληψη 2
Abstract 3
1. Εισαγωγή στην Επιστήμη των Δεδομένων (Data Science) 7
1.1 Τι είναι η Επιστήμη των Δεδομένων; 7
1.2 Τι είναι η Μηχανική Μάθηση; 8
1.3 Ερευνητικές Δεξιότητες στην Επιστήμη των Δεδομένων 8
1.4 Δεξιότητες Λογισμικού στην Επιστήμη των Δεδομένων 8
1.5 Βασικές Στρατηγικές Επίλυσης Προβλημάτων 9
1.6 Big Data & 5V's Μοντέλο 10
2. Επιχειρηματική Ευφυΐα και Big Data Αναλυτική (Business Intelligence and Big Data Analytics) 12
2.1 Τι είναι Επιχειρηματική Ευφυΐα και Big Data Αναλυτική; 12
2.2 Οι Τέσσερις Τύποι Αναλυτικής 12
2.2.1 Περιγραφική Αναλυτική 14
2.2.2 Διαγνωστική Αναλυτική 14
2.2.3 Προβλεπτική Αναλυτική 15
2.2.4 Καθοδηγητική Αναλυτική 15
2.3 Παράγοντες Υιοθέτησης Πληροφοριακών Συστημάτων Επιχειρηματικής Ευφυΐας και Big Data Αναλυτικής 16
3. Προηγμένη Αναλυτική και Αλγόριθμοι Μηχανικής Μάθησης (Advanced Analytics and Machine Learning Algorithms) 17
3.1 Βασικοί Τύποι Μοντέλων Μηχανικής Μάθησης 17
3.2 Εφαρμογές Αλγορίθμων Μηχανικής Μάθησης 18
3.2.1 Αλγόριθμοι Εποπτευόμενης Μηχανικής Μάθησης (Supervised Machine Learning Algorithms) 18
3.2.1.1 Η εξέλιξη των Tree-based Αλγορίθμων 22
3.2.2 Αλγόριθμοι μη-Εποπτευόμενης Μηχανικής Μάθησης (Unsupervised Machine Learning Algorithms) 25
3.2.3 Αλγόριθμοι Ενισχυτικής Μηχανικής Μάθησης (Reinforcement Machine Learning) 28
3.3 Αυτοματοποιημένη Μηχανική Μάθηση (Automated Machine Learning – AutoML) 30
3.3.1 Τι είναι AutoML; 30

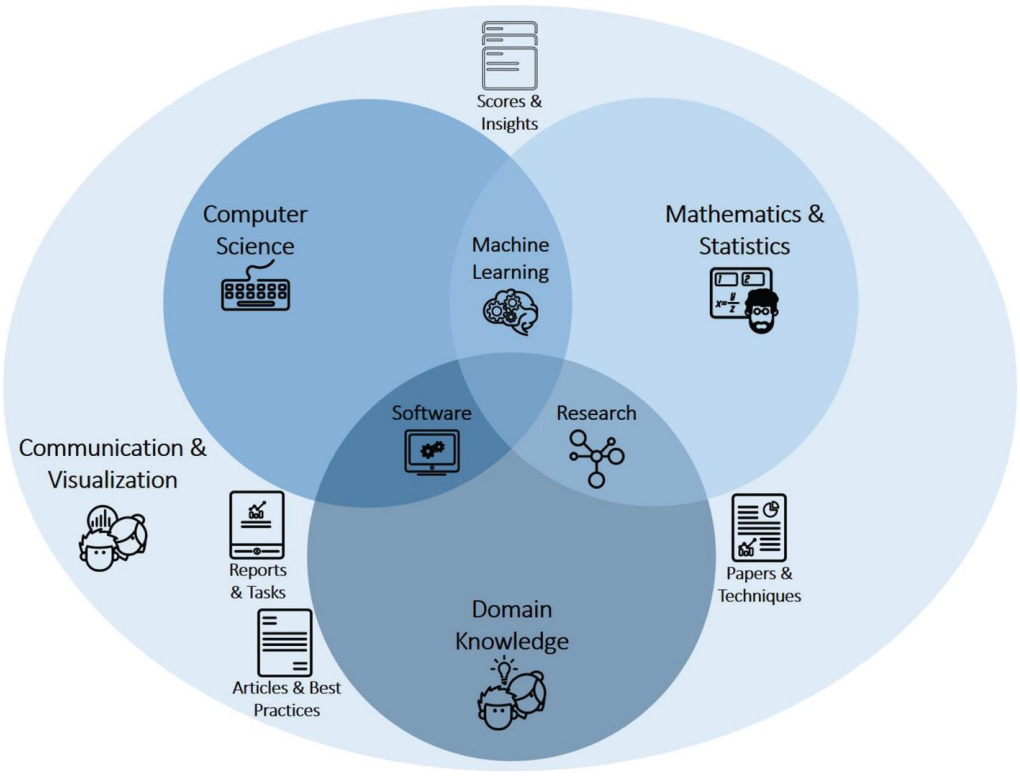
3.3.2 Βασικά Στάδια στη Δημιουργία AutoML Pipelines (AutoML Lifecycle) 31
4. Υλοποίηση και Οδηγός Χρήσης “Matapae” AutoML Web Application 35
4.1 Σκοπός της Εφαρμογής 35
4.2 RStudio IDE, GitHub και Shiny Web App Framework 35
4.3 Σχήμα Βάσης Δεδομένων – Oracle MySQL Database Server 36
4.4 Σύστημα Επαλήθευσης Ταυτότητας και Εξουσιοδότησης Χρηστών (Auth0) 38
4.4.1 R Κώδικας για την Εγγραφή Χρήστη στη Βάση Δεδομένων 39
4.5 Dashboard 40
4.6 Data Manager 41
4.6.1 File Uploader 41
4.6.1.1 R Κώδικας για τον Έλεγχο Εγκυρότητας του Ονόματος Αποθήκευσης 45
4.6.1.2 R Κώδικας για την Αποθήκευση του Dataset στη Βάση Δεδομένων 46
4.6.2 Datasets’ Storage 47
4.6.2.1 R Κώδικας για την Ανάκτηση της Λίστας των Αποθηκευμένων Datasets 50
4.6.2.2 R Κώδικας για τη Διαγραφή Αποθηκευμένου Dataset 51
4.6.2.3 R Κώδικας για την Ανάκτηση Αποθηκευμένου Dataset 52
4.7 Exploratory Data Analysis (EDA) 53
4.7.1 Dataset Dimension 54
4.7.2 Basic Information 55
4.7.3 Summary Statistics 56
4.7.4 Descriptive Statistics 57
4.7.5 Missing Values 58
4.7.6 Histograms & Density Plots για τις Continuous Μεταβλητές 59
4.7.7 Multivariate Analysis 61
4.7.8 Bar Plots για τις Categorical Μεταβλητές 62
4.7.9 Quantile-Quantile Plots – Κανονική Κατανομή (Normal Distribution) 63
4.7.10 Box Plots 65
4.7.11 Scatter Plots 66
4.7.12 Principal Component Analysis (PCA) 67
4.7.13 R Κώδικας για τη διαχείριση των EDA Tabs 69
4.8 ML Prediction Modeling 73
4.8.1 Supervised ML 74
4.8.1.1 Step 1 – Dataset and Partitions 75

4.8.1.2 Step 2 – Prediction Type: Regression or Classification 76
4.8.1.3 Step 3 – Dependent and Independent Variables 77
4.8.1.4 Step 4 – Cross Validation and Hyperparameter Tuning 78
4.8.1.4.1 R Κώδικας για την Εκπαίδευση XGBoost Regression Models 82
4.8.2 Learning Process and Results 84
4.8.2.1 R Κώδικας για την Αποθήκευση ML Model στη Βάση Δεδομένων 88
4.8.3 ML Models' Storage 90
4.8.3.1 Feature Importance 91
4.8.3.2 Prediction Plot 92
4.8.3.3 Prediction in New Data 93
4.8.3.4 R Κώδικας για την Ανάκτηση Αποθηκευμένου ML Model 94
Βιβλιογραφία 95

1. Εισαγωγή στην Επιστήμη των Δεδομένων (Data Science)

1.1 Τι είναι η Επιστήμη των Δεδομένων;

Η Επιστήμη των Δεδομένων δεν αποτελεί έναν αυτοτελή επιστημονικό κλάδο, καθώς συμπεριλαμβάνει μια σειρά διαφορετικών τομέων τεχνογνωσίας και δεξιοτήτων που συνδυάζονται για την επίλυση προβλημάτων και τη βελτιστοποίηση των διαδικασιών. Οι πιο σημαντικές δεξιότητες που απαιτούνται είναι τα μαθηματικά και η στατιστική, η επιστήμη των υπολογιστών, τις εξειδικευμένες γνώσεις σχετικά με τον εκάστοτε τομέα που θα ερευνηθεί (Domain Knowledge), την οπτικοποίηση (Visualization) και επικοινωνία των αποτελεσμάτων, όπως αναφέρεται στο γράφημα στην Εικόνα 1.



Εικόνα 1: Περιοχές Εξειδίκευσης στην Επιστήμη των Δεδομένων

Οι Data Scientists χρειάζονται τα μαθηματικά και τη στατιστική για να κατανοήσουν τα δεδομένα που δημιουργούνται στο επιχειρηματικό σενάριο που θα κληθούν να μοντελοποιήσουν ώστε να προκύψουν χρήσιμα Insights, κατηγοριοποιήσεις ή εκτιμήσεις για μελλοντικά γεγονότα. Επίσης, για την αξιολόγηση των μοντέλων που αναπτύχθηκαν και κατά πόσο το παραγόμενο μοντέλο που αναπτύχθηκε είναι αντιπροσωπευτικό του προβλήματος αλλά και πώς μπορεί να χρησιμοποιηθεί για την επίλυση ή τη βελτίωση μιας συγκεκριμένης διαδικασίας [1].

1.2 Τι είναι η Μηχανική Μάθηση;

Η μηχανική μάθηση (Machine Learning) είναι το πεδίο που διασταυρώνει τα μαθηματικά, τη στατιστική και την επιστήμη των υπολογιστών. Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης (Artificial Intelligence) που βασίζεται στην ιδέα ότι τα συστήματα μπορούν να μάθουν από δεδομένα, να αναγνωρίζουν πρότυπα, να αναγνωρίζουν συμπεριφορές και να λαμβάνουν αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση. Είναι μια μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί την προετοιμασία δεδομένων (Data Preparation), τη δημιουργία νέων μεταβλητών (Feature Engineering), την εκπαίδευση μοντέλων και τελικά την ανάπτυξη μοντέλων. Η μηχανική μάθηση επιτρέπει στους Data Scientists να εφαρμόζουν πολύ περίπλοκα μοντέλα, όπως νευρωνικά δίκτυα (Neural Network) ή μηχανές διανυσματικής υποστήριξης (Support Vector Machines), και ένα σύνολο απλών μοντέλων όπως δέντρα αποφάσεων (Decision Trees), Gradient Boosting και Random Forest. Αυτά τα πολύπλοκα μοντέλα μπορούν να ανιχνεύσουν πολύ ασυνήθιστες σχέσεις μεταξύ των ανεξάρτητων input μεταβλητών (Independent Variables) και της εξαρτημένης target μεταβλητής (Dependent Variables) [1].

1.3 Ερευνητικές Δεξιότητες στην Επιστήμη των Δεδομένων

Η διασταύρωση των μαθηματικών, της στατιστικής και της εξειδικευμένης γνώσης στους τομείς μελέτης (Domain Knowledge) αποτελούν το ερευνητικό πεδίο. Οι ερευνητικές δεξιότητες επιτρέπουν στους Data Scientists να εφαρμόζουν νέες τεχνικές στη δημιουργία μοντέλων. Είναι σημαντικό να γνωρίζουμε όχι μόνο ποιες μεταβλητές θα συμπεριληφθούν στο μοντέλο, αλλά και τις κατάλληλες μορφές συναρτήσεων (Functional Forms) των μαθηματικών εξισώσεων [2]. Αυτός ο συνδυασμός επιτρέπει την ανάπτυξη πολύ περίπλοκων μοντέλων που είναι πιο ακριβή και λιγότερο εξαρτώμενα από τη functional form. Οι ερευνητικές δεξιότητες (Research Skills) μπορούν να επιταχύνουν τη διαδικασία ανάπτυξης των μοντέλων, ειδικά όταν χρειάζονται να ληφθούν υπόψιν λιγότερες υποθέσεις σχετικά με την κατανομή του στόχου και τη σχέση των ανεξάρτητων και εξαρτημένων μεταβλητών. Οι δεξιότητες λογισμικού στην επιστήμη δεδομένων αναφέρονται συνήθως στη διασταύρωση της επιστήμης των υπολογιστών και της γνώσης τομέα.

1.4 Δεξιότητες Λογισμικού στην Επιστήμη των Δεδομένων

Οι δεξιότητες λογισμικού (Software Skills) στην επιστήμη δεδομένων αναφέρονται συνήθως στη διασταύρωση της επιστήμης των υπολογιστών και της εξειδικευμένης γνώσης του εκάστοτε τομέα μελέτης (Domain Knowledge). Οι δεξιότητες λογισμικού, όπως η εξοικείωση στις γλώσσες ανοιχτού κώδικα και άλλων γλωσσών προγραμματισμού, βοηθούν τους Data Scientists να δημιουργήσουν νέα μοντέλα. Ο συνδυασμός δεξιοτήτων Computer Science, Software και Domain Knowledge μπορεί να

βοηθούν τους Data Scientists να λύσουν ένα επιχειρηματικό πρόβλημα ή να βελτιώσουν μια συγκεκριμένη επιχειρηματική διαδικασία [1].

1.5 Βασικές Στρατηγικές Επίλυσης Προβλημάτων

Η Επιστήμης των Δεδομένων αποτελεί μια μεθοδολογία μέσω της οποίας δίνεται η δυνατότητα να τεκμηριωθούν απόψεις οι οποίες βασίζονται στα δεδομένα και να αποφευχθεί η λήψη αποφάσεων βάσει βέλτιστων πρακτικών ή ακόμα και τη διαίσθηση. Με αυτό τον τρόπο αξιοποιείται η δυνατότητα εφαρμογής όλης της επιστημονικής γνώσης που έχουμε σχετικά με την εξαγωγή γνώσεων από τα δεδομένα. Γενικά, η Επιστήμη των Δεδομένων επιτρέπει την υιοθέτηση τεσσάρων βασικών στρατηγικών εξερεύνησης των δεδομένων [3]:

1. **Probing Reality:** Τα δεδομένα μπορούν να συλλέγονται ως αποτέλεσμα παθητικών ή ενεργητικών μεθόδων. Η τελευταία περίπτωση αφορά τα δεδομένα που αντιπροσωπεύουν την ανταπόκριση του κοινού σε στοχευμένες ενέργειες. Η ανάλυση των απαντήσεων μπορεί να είναι εξαιρετικά πολύτιμη για τις αποφάσεις των μελλοντικών ενεργειών. Ένα τέτοιο παράδειγμα που αντιπροσωπεύει αυτή τη στρατηγική είναι το A/B Testing. Πρόκειται για μία σειρά διαδοχικών ελέγχων σχετικά με το ποιο μέγεθος, σχήμα ή χρώμα είναι πιο αποδοτικό για τον σκοπό που χρησιμοποιείται. Η καλύτερη απάντηση δίνεται από τους ίδιους τους χρήστες.
2. **Pattern Discovery:** Κατά την επίλυση προβλημάτων που αφορούν «ψηφιοποιημένες» διαδικασίες, δίνεται η δυνατότητα να αναλυθούν αυτόματα τα δεδομένα τους και να ανακαλυφθούν χρήσιμα μοτίβα και Clusters, ικανά να ανάγουν υψηλής πολυπλοκότητας προβλήματα σε απλούστερα. Από τη χρήση αυτής της τεχνικής προκύπτει η κατάρτιση προφίλ χρηστών τα οποία αποτελούν καταλυτικό παράγοντα για την αποτελεσματικότητα διαφόρων πεδίων, όπως τα Programmatic Ads ή το Digital Marketing.
3. **Predicting Future Events:** Από τις αρχές της στατιστικής επιστήμης, ένα από τα πιο σημαντικά επιστημονικά ερωτήματα ήταν πώς θα δημιουργηθούν ισχυρά μοντέλα δεδομένων που να είναι ικανά να προβλέψουν τις μελλοντικές τιμές των μεταβλητών. Η προγνωστική αναλυτική (Predictive Analytics) επιτρέπει τη λήψη αποφάσεων σε μελλοντικά γεγονότα εκ των προτέρων. Καθώς το μέλλον είναι αδύνατον να προβλεφθεί και πάντα θα υπάρχουν απρόβλεπτα γεγονότα, η γνώση που προκύπτει από τον εντοπισμό γεγονότων που είναι δυνατόν να προβλεφθούν είναι πολύτιμη. Για παράδειγμα, η προγνωστική αναλυτική μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση του προγράμματος εργασιών για το προσωπικό καταστημάτων λιανικής κατά τη διάρκεια της επόμενης εβδομάδας, αναλύοντας δεδομένα όπως ο καιρός, η ιστορικότητα των πωλήσεων, οι συνθήκες κυκλοφορίας κ.λπ.
4. **Understanding People and the World:** Αποτελεί την πιο μεγαλεπήβολη στρατηγική η οποία αυτή τη στιγμή είναι μακριά από το πεδίο δραστηριοτήτων των περισσότερων εταιρειών. Ωστόσο, μεγάλες εταιρείες άρχισαν να επενδύουν στην έρευνα για την κατανόηση της φυσικής γλώσσας (Natural Language), την όραση των υπολογιστών (Computer Vision), την ψυχολογία και τη νεύρο-επιστήμη (Neuroscience). Η επιστημονική κατανόηση αυτών των πεδίων έχει μεγάλη σημασία στον τομέα της Επιστήμης των Δεδομένων αφού μέσω αυτών τεκμηριώνονται

οι πραγματικές διαδικασίες και η συμπεριφορά των ανθρώπων, οι οποίες συμβάλουν στην βέλτιστη λήψη αποφάσεων. Η ανάπτυξη μεθόδων βαθιάς μάθησης για την κατανόηση φυσικής γλώσσας και για την οπτική αναγνώριση αντικειμένων είναι ένα καλό παράδειγμα αυτού του είδους έρευνας.

1.6 Big Data & 5V's Μοντέλο

Τα Big Data, συχνά λαμβάνονται με περιορισμένη γνώση των συνθηκών των οποίων δημιουργήθηκαν, συλλέχθηκαν και προετοιμάστηκαν για ανάλυση. Πρόκειται για συλλογές κοινωνικοοικονομικών μεταβλητών όπως φύλο, εκπαίδευση, εισόδημα, αριθμό παιδιών κλπ. και ονομάζονται empirical observational και χρησιμοποιούνται στην επιστήμη των δεδομένων. Τα πειράματα που βασίζονται σε τέτοιου τύπου δεδομένα δεν μπορούν να αξιολογηθούν ως επιστημονικά, καθώς σε αυτή την περίπτωση θα απαιτούνταν η διενέργεια ελέγχων ακριβείας. Σε αντίθεση με τα Empirical Scientific ή Experimental δεδομένα, τα Empirical Observational δεδομένα, συνήθως βρίσκονται υπό κλίμακα (Data at Scale) κατά τάξη μεγέθους σε μία ή περισσότερες από τις βασικές αρχές που διέπουν τα Big Data και εν συντομία αναφέρονται ως το μοντέλο των 5V's (Volume, Velocity, Variety, Veracity και Value). Ως εκ τούτου, απαιτούν διαχείριση (Scale Management) και μεθόδους ανάλυσης (Analytic Methods) που σπάνια απαιτούνται στην εμπειρική επιστήμη [4] [5].

Πιο επεξηγηματικά, μελετώντας την αρχιτεκτονική των Big Data, το μοντέλο 5V's αναλύεται στις παρακάτω θεμελιώδεις αρχές [6] [7] [8]:

- **Volume:** Το μέγεθος, η περιπλοκότητα της επεξεργασίας και ο σκοπός συλλογής των δεδομένων. Παγκοσμίως υπολογίζεται ότι καθημερινά δημιουργούνται 2.5 Quintillions Bytes δεδομένων.
- **Velocity:** Ο ρυθμός ταχύτητας με τον οποίο τα δεδομένα δημιουργούνται, αυξάνονται, συγκεντρώνονται και διαμοιράζονται μέσω των διαφόρων διαθέσιμων πόρων, όπως μέσω των εκατοντάδων Sensors που διαθέτουν οι συσκευές IoT, δισεκατομμυρίων συνδέσεων δικτύου και υποδομών πληροφοριακών συστημάτων. Η ραγδαία αύξηση του ρυθμού δημιουργίας των δεδομένων καθιστά περίπλοκη τη γρήγορη επεξεργασία και μεταβίβαση τους.
- **Variety:** Η ποικιλία στη μορφή και στους τύπους των δεδομένων, όπου διακρίνονται σε:
    - Δομημένα Δεδομένα (Structured Data) σε σειρές και στήλες τα οποία συνήθως αποθηκεύονται σε σχεσιακές βάσεις δεδομένων. Συλλέγονται από μεγάλο αριθμό αντιπροσωπευτικών περιπτώσεων και αναλύονται στατιστικά (π.χ. αριθμητικά δεδομένα).
    - Αδόμητα Δεδομένα (Unstructured Data) που περιλαμβάνουν διαφορετικούς τύπους δεδομένων οι οποίοι δεν μπορούν να οργανωθούν σε κατάλληλη μορφή όπως τα αρχεία κειμένων, ηχητικά αρχεία και βίντεο, XML κλπ. Συλλέγονται από μικρό αριθμό μη αντιπροσωπευτικών περιπτώσεων και δεν αναλύονται στατιστικά.
    - Ημι-Δομημένα Δεδομένα (Semi-Structured Data) αποτελούν τα δεδομένα που δεν είναι πλήρως δομημένα ή αδόμητα δηλαδή είναι μερικώς δομημένα και συνδυασμένα με αδόμητης μορφής δεδομένα.

- **Veracity:** Η ακρίβεια, η σημασία και η επιβεβαίωση της ορθότητας των δεδομένων αποτελούν σημαντικό παράγοντα για την εξαγωγή σωστής πληροφορίας από ένα μεγάλο σύνολο διαφορετικών τύπων δεδομένων. Οποιαδήποτε ανακρίβεια μπορεί να οδηγήσει σε ζημία στα έσοδα ή στα απαιτούμενα αποτελέσματα.
- **Value:** Η αξία, που αποτελεί τον σημαντικότερο πυλώνα των Big Data καθώς μέσω αυτής μετριέται είτε η χρησιμότητα των δεδομένων για τη λήψη αποφάσεων, είτε το «επίπεδο» της αξίας που προσδίδουν τα δεδομένα μετά την επικείμενη επεξεργασία τους. Δηλαδή, από το στάδιο της συλλογή των δεδομένων μέχρι την ανάλυση και πρόβλεψη των διερευνητικών υποθέσεων που βασίζονται στα αντίστοιχα δεδομένα.

2. Επιχειρηματική Ευφυΐα και Big Data Αναλυτική (Business Intelligence and Big Data Analytics)

2.1 Τι είναι Επιχειρηματική Ευφυΐα και Big Data Αναλυτική;

Η Επιχειρηματική Ευφυΐα και Big Data Αναλυτική (BIBDA) αποτελεί ένα σύνολο τεχνικών, τεχνολογιών, συστημάτων, αρχιτεκτονικών, μεθοδολογιών, εφαρμογών και διαδικασιών προηγμένων μορφών επεξεργασίας μεγάλης ποσότητας (Big Data) κρίσιμων επιχειρησιακών δεδομένων μέσω στατιστικών και ποσοτικών αναλύσεων. Οι αναλύσεις αυτές οδηγούν στην εξαγωγή χρήσιμων επεξηγηματικών αναφορών (Reports) και προγνωστικών μοντέλων για την απόκτηση νέων γνώσεων που υποστηρίζουν τη λήψη αποφάσεων [9] [10] [11] [12].

2.2 Οι Τέσσερις Τύποι Αναλυτικής

Η έννοια της Επιχειρηματικής Ευφυΐας και Big Data Αναλυτικής εξατομικεύεται περαιτέρω, βάσει του σκοπού που διενεργείται, των εργαλείων και τεχνικών που χρησιμοποιούνται σε τέσσερις γενικούς τύπους που οι επιχειρήσεις και οι οργανισμοί μπορούν να εφαρμόσουν. Κατά το στάδιο της Περιγραφικής Αναλυτικής (Descriptive Analytics) αναπτύσσονται συνοπτικές αναφορές των παρελθοντικών επιχειρηματικών δραστηριοτήτων και συναλλαγών. Με τις επαυξημένες δυνατότητες των εργαλείων της Διαγνωστικής Αναλυτικής (Diagnostic Analytics), όπως η εξόρυξη νέων δεδομένων, η διασύνδεσης τους, η εύρεση συσχετίσεων μεταξύ τους, η λεπτομερέστερη ανάλυση (Drill-Down) και η εφαρμογή απλών στατιστικών μοντέλων, κατανοείται ο λόγος που έχουν συμβεί τα γεγονότα μέσω της χρήσης των προηγμένων αναφορών Επιχειρηματικής Ευφυΐας (Business Intelligence). Τα επόμενα στάδια προηγμένης Big Data Αναλυτικής (ή Advances Analytics) σχετίζονται με τις μελλοντικές εξελίξεις και καταναλωτικές συμπεριφορές μέσω της Προβλεπτικής Αναλυτικής (Predictive Analytics), καθώς επίσης, την Καθοδηγητική Αναλυτική (Prescriptive Analytics) βάσει της οποίας προτείνονται εναλλακτικές επιλογές λύσεων ώστε να επιλεγούν οι βέλτιστες για μελλοντικές ενέργειες [10] [13].

Ο κάθε γενικός τύπος αναλυτικής δίνει τις αντίστοιχες απαντήσεις στα στελέχη των επιχειρήσεων σχετικά με το «Τι συνέβη;», «Γιατί συνέβη;», «Τι θα συμβεί;» και «Πως μπορούμε να το κάνουμε να συμβεί;» έπειτα από αναλύσεις που διενεργούνται με τη χρήση εξειδικευμένων εργαλείων όπως απεικονίζονται στον Πίνακας 1 [14].

Τύπος Αναλυτικής	Ερωτήματα	Εργαλεία	Αποτελέσματα	Σκοπός
Περιγραφική (εκ των υστέρων γνώση)	Τι συνέβη;	• Μοντελοποίηση Δεδομένων (Data Modeling) • Business Reporting (KPIs, Metrics) • Αυτοματοποιημένη Παρακολούθηση / Προειδοποιητικά Μηνύματα (Προκαθορισμένα Thresholds) • Οπτικοποιήσεις (Visualizations) • Dashboards • Scorecards • Παλινδρόμηση (Regression)	Σαφώς ορισμένα επιχειρηματικά προβλήματα ή ευκαιρίες	Αποκάλυψη μοτίβων και τάσεων που προσφέρουν στα στελέχη την επίγνωση της υπάρχουσας κατάστασης
	Τι συμβαίνει;			
Διαγνωστική (Γνώση της υπάρχουσας κατάστασης)	Γιατί συνέβη;	• Data Warehouses • OLAP (Cubes, Slice & Dice, Drill-Down/Roll-Up) • Modelling Statistics • Data Discovery (Ad-hoc Queries) • Εξόρυξη Δεδομένων (Data Mining) • Correlations	Ακριβείς “προβολές” των μελλοντικών συνθηκών και καταστάσεων	Αναγνώριση μοτίβων του παρελθόντος για πρόβλεψη του μέλλοντος
Προβλεπτική (Προνοητικότητα)	Τι θα συμβεί;	• Εξόρυξη Δεδομένων (Data Mining) • Text/Media Mining • Προβλεπτικά Μοντέλα Μηχανικής Μάθησης • Artificial Neural Networks (ANN)		
	Τι είναι πιθανό να συμβεί;			
	Γιατί θα συμβεί;			
Καθοδηγητική (Αυτοματοποίηση)	Πως μπορούμε να το κάνουμε να συμβεί;	• Μοντελοποίηση Αποφάσεων (Decision Modeling) • Optimization • Προσομοιώσεις (Simulations)	Βελτιστοποίηση – Καλύτερη δυνατή επιχειρηματική λύση	Εστίαση στη λήψη αποφάσεων και την αποδοτικότητα
	Τι πρέπει να γίνει ώστε να συμβεί;			
	Γιατί πρέπει να συμβεί;			

Πίνακας 1: Σύνοψη των Τύπων Επιχειρηματικής Ευφυΐας και Big Data Αναλυτικής

2.2.1 Περιγραφική Αναλυτική

Είναι η πιο βασική μορφή αναλυτικής κατά την οποία γίνεται χρήση ιστορικών δεδομένων ή δεδομένων σε πραγματικό χρόνο ώστε να ανιχνευθούν τάσεις από καταναλωτικές συμπεριφορές, δραστηριότητες και συναλλαγές του παρελθόντος. Αποτελεί προαπαιτούμενο βήμα για τους μεταγενέστερους τύπους αναλυτικής. Η Περιγραφική Στατιστική (Descriptive Statistics) περιλαμβάνει αθροίσματα, μέσους όρους και ποσοστιαίες μεταβολές [9].

Με την κατηγοριοποίηση, τον χαρακτηρισμό των μεταβλητών, τη συλλογή και την κατάταξη των δεδομένων, η Περιγραφική Αναλυτική μετατρέπει τα δεδομένα σε χρήσιμες πληροφορίες για την ανάλυση των επιχειρηματικών αποφάσεων και των αποτελεσμάτων τους. Οι συνοπτικές πληροφορίες που προκύπτουν μπορούν να εμφανιστούν είτε ως γραφήματα και αναφορές είτε απλά ως απαντήσεις που προκύπτουν από SQL ερωτήματα [15].

Για παράδειγμα, μέσα από την Περιγραφική Αναλυτική μπορεί να γίνει η διαπίστωση ότι υπάρχει μια εποχική αύξηση στις πωλήσεις για ένα από τα προϊόντα της εταιρείας. Δηλαδή, ότι το προϊόν «Α» παρουσιάζει αύξηση πωλήσεων τον Οκτώβριο, τον Νοέμβριο και τις αρχές Δεκεμβρίου κάθε έτους [16].

2.2.2 Διαγνωστική Αναλυτική

Προχωρώντας την ανάλυση ένα βήμα παρακάτω, με τη Διαγνωστική Αναλυτική (Diagnostic Analytics) απαντάται το ερώτημα πώς και γιατί συνέβη ή συμβαίνει ένα γεγονός. Για να απαντηθούν αυτά τα ερωτήματα και να δημιουργηθούν οι αντίστοιχες πληροφορίες από τις συλλογές Big Data, χρησιμοποιούνται ορισμένες τεχνικές, όπως το Drill-Down, η εξόρυξη δεδομένων (Data Mining) και οι συσχετίσεις (Correlations) για την εύρεση μοτίβων και τάσεων. Η Διαγνωστική Αναλυτική βοηθάει στην κατανόηση των Cross-Functional δεδομένων, δηλαδή των δεδομένων που προκύπτουν από τα διαφορετικά τμήματα των εταιρειών, επιτρέποντας έτσι στις επιχειρήσεις ή τους οργανισμούς να αναπτύσσουν στρατηγική γύρω από τις πωλήσεις, τα έσοδα, το κόστος, το κέρδος, τον κίνδυνο που πιθανώς να εμφανιστεί καθώς και τη μέτρηση της απόδοσης. Αυτή η δυνατότητα προσφέρεται μέσω εξατομικευμένων διαδραστικών αναφορών με δυνατότητα φιλτραρίσματος και λεπτομερούς εστίασης ή ιεραρχικής επέκτασης στα δεδομένα [9].

Συνεχίζοντας τη διερεύνηση για το προϊόν «Α», βάσει των δημογραφικών δεδομένων για το κοινό στο οποίο απευθύνεται, διαπιστώνεται ότι είναι μεταξύ 8 έως 18 ετών. Οι πελάτες, ωστόσο, είναι μεταξύ 35 και 55 ετών. Η ανάλυση των δεδομένων της καταναλωτικής έρευνας αποκαλύπτει ότι υπάρχει ένα βασικό κίνητρο για τους πελάτες της ηλικιακής κατηγορίας των γονέων ώστε να αγοράσουν το «Α» για τα παιδιά τους. Έτσι λοιπόν, η άνοδος των πωλήσεων κατά τους μήνες πριν και κατά τη διάρκεια της εορταστικής περιόδου πιθανώς οφείλεται στην αγοραστική ανάγκη που προκύπτει για την αγορά παιδικών δώρων [16].

2.2.3 Προβλεπτική Αναλυτική

Καθώς ολοκληρωθούν τα βήματα της Περιγραφικής και Διαγνωστικής Αναλυτικής, οι Data Scientists μπορούν να συνδυάσουν τα ιστορικά δεδομένα με προβλεπτικούς αλγόριθμους μηχανικής μάθησης (Machine Learning) ώστε να πραγματοποιήσουν προβλέψεις για μελλοντικά γεγονότα και τα πιθανά αποτελέσματά τους ώστε να συμβάλουν στον προγραμματισμό και τον καθορισμό των εταιρικών στόχων. Οι προβλέψεις που προκύπτουν εκφράζονται από τις πιθανότητες βάσει των οποίων μπορεί να πραγματοποιηθούν ή όχι τα γεγονότα. Για την επικαιροποίηση της ακρίβειας των αποτελεσμάτων χρησιμοποιούνται τεχνικές μέτρησης όπως είναι: Accuracy Confusion Matrix (ACC), Area Under the ROC Curve (AUC), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Sum Squared Error (SSE) ανάλογα με τον τύπο του αλγορίθμου μηχανικής μάθησης που έχει χρησιμοποιηθεί για την πρόβλεψη (π.χ. Classification, Regression, Clustering) [9].

Για παράδειγμα, γνωρίζοντας ότι οι πωλήσεις του προϊόντος «Α» εμφανίζουν αύξηση τον Οκτώβριο, το Νοέμβριο και στις αρχές Δεκεμβρίου κάθε έτους την τελευταία δεκαετία, παρέχονται άφθονα δεδομένα για να γίνει η πρόβλεψη ότι η ίδια τάση είναι πιθανό να συμβεί και το επόμενο έτος [16].

2.2.4 Καθοδηγητική Αναλυτική

Η Καθοδηγητική Αναλυτική (Prescriptive Analytics) συμπεριλαμβάνει τη μέθοδο του Πειραματικού Σχεδιασμού (Experimental Design), όπου δίνονται απαντήσεις γιατί συνέβη κάτι με τη διεξαγωγή πειραμάτων, και τη μέθοδο της Βελτιστοποίησης (Optimization) μέσω της οποίας διερευνώνται στρατηγικές επιλογές δράσης των επιχειρήσεων. Όσον αφορά τον Πειραματικό Σχεδιασμό, για να αιτιολογηθούν τα συμπεράσματα, απαιτείται οι Data Scientists να εξετάσουν μια ή περισσότερες ανεξάρτητες μεταβλητές για να ελέγξουν αποτελεσματικά άλλες νέες μεταβλητές, εκτός από αυτές που μελέτησαν στα προηγούμενα βήματα της Προβλεπτικής Αναλυτικής. Στην περίπτωση που η ομάδα μεταβλητών που δοκιμάζεται είναι περισσότερο βιώσιμη από την ομάδα των ανεξάρτητων μεταβλητών που έχει ήδη εξεταστεί, τα στελέχη που είναι επιφορτισμένα με τη λήψη αποφάσεων θα εφαρμόσουν τη λύση με τη μεγαλύτερη βιωσιμότητα. Η μέθοδος της Βελτιστοποίησης προσδιορίζει το ιδανικό επίπεδο μιας συγκεκριμένης μεταβλητής σε σχέση με κάποια άλλη με τη χρήση μαθηματικών τεχνικών για την εξεύρεση βέλτιστων λύσεων βάσει ορισμένων κριτηρίων και περιορισμών [17].

Ολοκληρώνοντας το παράδειγμα του προϊόντος «Α», σε αυτό το στάδιο της ανάλυσης, η ομάδα θα πρέπει να πάρει στρατηγικές αποφάσεις δεδομένης την προβλεπόμενης αύξησης των πωλήσεων λόγω της τάσης εποχικότητας που παρουσιάζει η αγορά δώρων. Στο τμήμα Marketing πιθανώς να αποφασιστεί η εφαρμογή A/B Testing σε δυο groups διαφημίσεων με εξατομικευμένο περιεχόμενο, το ένα εκ των οποίων θα απευθύνεται στους τελικούς χρήστες του προϊόντος (παιδιά) και το άλλο στους πελάτες (γονείς). Τα αποτελέσματα από το A/B Testing θα κατατοπίσουν τα στελέχη ακόμα περισσότερο ώστε να αξιοποιήσουν καλύτερα την εποχική άνοδο των πωλήσεων, αναγνωρίζοντας περισσότερους παράγοντες που την προκαλούν και την επηρεάζουν. Ειδικότερα,

μια απόφαση για την έναρξη των προωθητικών ενεργειών Marketing με εορταστικό περιεχόμενο κατά ένα μηνά νωρίτερα από το συνηθισμένο (Σεπτέμβριο) θα επεκτείνει την εορταστική περίοδο, συμβάλλοντας στην εποχική αύξηση των πωλήσεων από το μήνα Σεπτέμβρη [16].

2.3 Παράγοντες Υιοθέτησης Πληροφοριακών Συστημάτων Επιχειρηματικής Ευφυΐας και Big Data Αναλυτικής

Τα Big Data είναι κάτι περισσότερο από μεγάλου όγκου δεδομένα. Η διερεύνηση και ανάλυση ημι-δομημένων και αδόμητων δεδομένων αποτελεί ένα νέο εργαλείο για τις επιχειρήσεις, το οποίο μπορεί να ενισχύσει το επιχειρηματικό τους πλεονέκτημα έναντι των ανταγωνιστών τους. Λόγω της εκθετικής αύξησης της ημερήσιας δημιουργίας δεδομένων, και καθώς υπολογίζεται ότι το 90% εξ αυτών είναι πρόσφατα και ανεξερεύνητα, γίνεται αντιληπτή η σπουδαιότητα της υιοθέτησης Πληροφοριακών Συστημάτων (IS) Επιχειρηματικής Ευφυΐας και Big Data Αναλυτικής (BIBDA) [9].

Σύμφωνα με τη μελέτη [10], οι επιχειρήσεις οι οποίες για την υποστήριξη των καθημερινών τους λειτουργιών, έχουν αναπτύξει και χρησιμοποιούν εκτενώς Επιχειρησιακά Συστήματα (ES) όπως Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) και Supply Chain Management (SCM), μέσω των οποίων καταγράφονται μεγάλες ποσότητες δεδομένων, προχωρούν στην υιοθέτηση συστημάτων BIBDA. Έναν ακόμα σημαντικό παράγοντα υιοθέτησης συστημάτων BIBDA, αποτελούν οι ικανότητες διαχείρισης Information Communication Technologies (ICT Management Capabilities). Δηλαδή, οι ικανότητες που βασίζονται στην καλή επικοινωνία του ICT προσωπικού και των υπόλοιπων τμημάτων που χρησιμοποιούν ICT, την καλή συνεργασία, την εμπιστοσύνη και την ανταλλαγή πληροφοριών με τους προμηθευτές ICT (Hardware, Software, Networks), όπως επίσης τις ICT στρατηγικές και τα επιχειρηματικά πλάνα τα οποία είναι συνυφασμένα με τη συνολική στρατηγική των επιχειρήσεων.

3. Προηγμένη Αναλυτική και Αλγόριθμοι Μηχανικής Μάθησης (Advanced Analytics and Machine Learning Algorithms)

Στο πλαίσιο των επιχειρηματικών διαδικασιών ανάλυσης των Big Data, όπως αναλύθηκε στο Κεφάλαιο 2.2, τα στάδια της Προβλεπτικής και Καθοδηγητικής Αναλυτικής (Predictive and Prescriptive Analytics) αναφέρονται με τον όρο Προηγμένη Αναλυτική (Advanced Analytics), καθώς αφορούν κάτι περισσότερο από απλά στατιστικά και μαθηματικά μοντέλα. Η Προηγμένη Αναλυτική συμπεριλαμβάνει τεχνικές Μηχανικής Μάθησης (Machine Learning), Προβλέψεων (Forecasting) και Βελτιστοποίησης (Optimization). Από το συνδυασμό τους, οι Data Scientists δημιουργούν μοντέλα ικανά να δώσουν εξειδικευμένες επιχειρηματικές λύσεις σε κάθε εφικτό σενάριο [1].

3.1 Βασικοί Τύποι Μοντέλων Μηχανικής Μάθησης

Υπάρχουν δύο βασικοί τύποι μοντέλων μηχανική μάθησης: η Εποπτευόμενη Μάθηση (Supervised Learning), όταν η εξαρτημένη μεταβλητή που αποτελεί τον στόχο της πρόβλεψης (Target) είναι γνωστή και χρησιμοποιείται στο μοντέλο και η μη-Εποπτευομένη Μάθηση (Unsupervised Learning), όταν ο στόχος είναι άγνωστος ή δεν χρησιμοποιείται στο μοντέλο πρόβλεψης. Οι μεταβλητές εισόδου (Input Variables), οι οποίες είναι γνωστές ως Ανεξάρτητες Μεταβλητές (Dependent Variables) στον τομέα της στατιστικής ή ως γνωρίσματα (Features) στον τομέα της μηχανικής μάθησης, περιέχουν πληροφορίες σχετικά με το προφίλ των πελατών, δηλαδή πώς καταναλώνουν το προϊόν ή την υπηρεσία, πώς πληρώνουν για αυτό, για πόσο καιρό είναι πελάτες, από ποιο «κανάλι» ήρθαν και πως περιηγήθηκαν σε ένα ηλεκτρονικό κατάστημα κ.α.

Ο στόχος (Target ή Label) σε ένα μοντέλο μηχανικής μάθησης είναι η πρόβλεψη, κατηγοριοποίηση ή εκτίμηση σημαντικών επιχειρηματικών γεγονότων τα οποία έχουν αξία για την απόδοση μιας εταιρείας. Για παράδειγμα, πότε ένας πελάτης θα σταματήσει να αγοράζει ένα προϊόν (Customer Churn), πότε θα αγοράσει ένα προϊόν και θα κάνει μια πληρωμή ή απλά θα δείξει ενεργό ενδιαφέρον για ένα προϊόν με μια τηλεφωνική επικοινωνία. Η μεταβλητή στόχος (Target Variable) αφορά τα Supervised Learning Models. Τα Unsupervised Models δεν απαιτούν στόχο, καθώς χρησιμοποιούνται για τη δημιουργία χρήσιμών πληροφοριών σχετικά με τα δεδομένα, την αγορά ή τους πελάτες, για την αξιολόγηση πιθανών τάσεων ή την καλύτερη κατανόηση συγκεκριμένων επιχειρηματικών σεναρίων. Αυτά τα μοντέλα δεν στοχεύουν σε κατηγοριοποίηση, πρόβλεψη ή εκτίμηση μελλοντικών επιχειρηματικών γεγονότων.

Υπάρχουν ακόμα δύο συμπληρωματικοί τύποι μοντέλων μηχανικής μάθησης. Ο ένας από αυτούς τους τύπους είναι τα μοντέλα ημι-Εποπτευόμενης Μάθησης (Semi-Supervised Learning), τα όποια είναι παρόμοια με τα Supervised, όμως περιλαμβάνουν μικρό όγκο δεδομένων με γνωστό στόχο (Labeled Data) και μεγάλο όγκο δεδομένων όπου ο στόχος είναι άγνωστος (Unlabeled Data). Χρησιμοποιώντας αυτό το συνδυασμό δεδομένων και τη χρήση αλγορίθμων μηχανικής μάθησης θα επιτευχθεί η εκμάθηση των Labels στα Unlabeled Data. Τα ημι-Εποπτευόμενα μοντέλα γίνονται όλο και πιο διαδεδομένα και συχνά εφαρμόζονται σε εφαρμογές τεχνητής νοημοσύνης (Artificial Intelligence). Υπάρχουν επίσης τα μοντέλα Ενισχυτικής Μάθησης (Reinforcement Learning), όπου ο αλγόριθμος εκπαιδεύεται χρησιμοποιώντας ένα σύστημα για να «επιβραβεύει» το βήμα της εκμάθησης όταν το μοντέλο πηγαίνει στη σωστή κατεύθυνση εκμάθησης και να το «τιμωρεί» όταν το μοντέλο πηγαίνει στη λάθος κατεύθυνση εκμάθησης (Trial and Error). Για παράδειγμα, η ενισχυτική μάθηση μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου ώστε να μαθαίνει και να αναλαμβάνει ενέργειες σε αυτοκίνητα αυτόνομης οδήγησης. Κατά τη διάρκεια της εκπαίδευσης, εάν το αυτοκίνητο οδηγεί με ασφάλεια στο δρόμο, το βήμα εκμάθησης «ανταμείβεται» επειδή πηγαίνει στη σωστή κατεύθυνση. Εάν το αυτοκίνητο βγει από το δρόμο, το βήμα εκμάθησης «τιμωρείται» επειδή η εκπαίδευση πηγαίνει σε λάθος κατεύθυνση [1].

3.2 Εφαρμογές Αλγορίθμων Μηχανικής Μάθησης

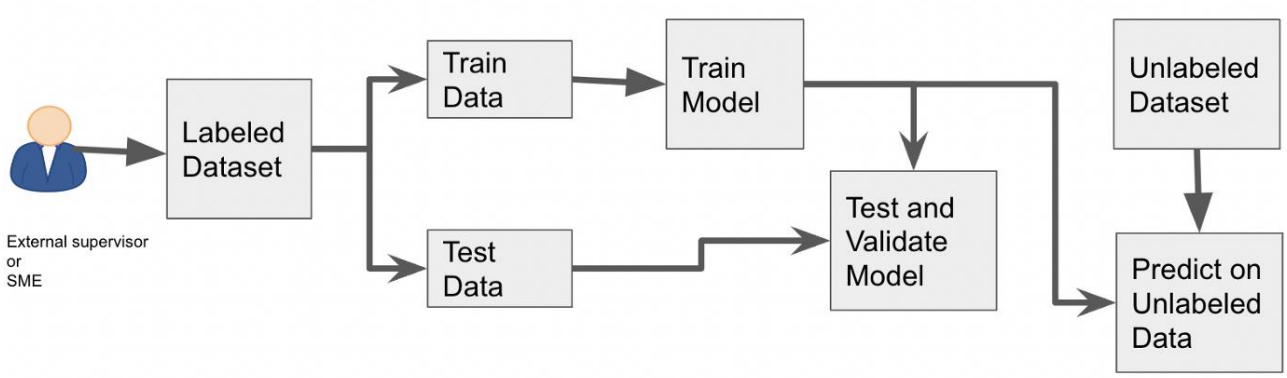
Η Μηχανική Μάθηση είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης, που ορίζεται ως η δυνατότητα ενός υπολογιστικού συστήματος να «μαθαίνει» από τα δεδομένα χρησιμοποιώντας αλγόριθμους με σκοπό να προσομοιάσει την ευφυΐα της ανθρώπινης συμπεριφοράς στη λήψη αποφάσεων και προβλέψεων. Μια συνήθεις ερώτηση που προκύπτει είναι: «Ποιος είναι ο κατάλληλος τύπος αλγορίθμου που θα πρέπει να εφαρμοστεί για την επίλυση των επικείμενων προβλημάτων;» Ακόμα και ένας έμπειρος Data Scientist, δεν είναι δυνατόν να γνωρίζει ποιος είναι ο ποιο αποδοτικός αλγόριθμος εάν δεν κάνει δοκιμές με διαφορετικούς αλγορίθμους, όμως σίγουρα γνωρίζει την κατηγορία αλγορίθμων που θα πρέπει να επιλέξει ανάλογα με το είδος του προβλήματος που θέλει να επιλύσει.

Στις παρακάτω υποενότητες, θα αναλυθούν λεπτομερώς οι αλγόριθμοι, κατηγοριοποιημένοι βάσει του τύπου μηχανικής μάθησης που απαιτεί η επίλυση του εκάστοτε προβλήματος: Εποπτευόμενης, μη-Εποπτευόμενης και Ενισχυτικής Μάθησης.

3.2.1 Αλγόριθμοι Εποπτευόμενης Μηχανικής Μάθησης (Supervised Machine Learning Algorithms)

Οι Supervised Learning αλγόριθμοι χρησιμοποιούν Labeled Datasets και τα μοντέλα μηχανικής μάθησης, που προκύπτουν, «μαθαίνουν» από ένα σύνολο παραδειγμάτων σχετικά με τον τρόπο που δρουν τα Features στην εξαρτημένη μεταβλητή (που είναι γνωστή και ως Label ή Target). Ο Data Scientist θα πρέπει να διαχωρίσει το σύνολο δεδομένων (Dataset) σε Train για τη διαδικασία

εκμάθησης και Test για την επικύρωση του μοντέλου που έχει εκπαιδευτεί. Κατά την Εποπτευόμενη Μηχανική Μάθηση, ο αλγόριθμος λαμβάνει ένα Dataset γνωστών παραδειγμάτων με καθορισμένα Features και Labels για να προσδιοριστούν μοτίβα στα δεδομένα, «μαθαίνοντας» από τις παρατηρήσεις. Βάσει της γενίκευσης στα δεδομένα της εκπαίδευσης, το μοντέλο γνωρίζει πώς για κάθε άγνωστο παράδειγμα τιμών εισόδου, θα προβλέπονται οι κατάλληλες τιμές εξόδου με υψηλό επίπεδο ακρίβειας/απόδοσης. Η διαδικασία εκπαίδευσης ενός μοντέλου με τη χρήση Supervised Learning αλγορίθμων περιγράφεται σχηματικά στην Εικόνα 2 [18] [19].



Εικόνα 2: Διαδικασία Εκπαίδευσης Supervised Learning Model

Οι Supervised Learning αλγόριθμοι μπορούν να επιλύσουν τρεις κατηγορίες προβλημάτων [1] [18] [19]:

1. **Κατηγοριοποίησης (Classification):** Όταν από τα δεδομένα θα προκύψει η πρόβλεψη μιας Κατηγορικής Μεταβλητής (Categorical Variable). Δηλαδή, στις περιπτώσεις που το Label θα έχει διακριτές καταστάσεις. Πιο συγκεκριμένα, όταν τα Labels έχουν μόνο δύο καταστάσεις (π.χ. “Cat” ή “Dog” σε ένα πρόβλημα κατηγοριοποίησης εικόνας, “Churn” ή “non-Churn” σε ένα K-Nearest Neighbors Classification Model), τότε το πρόβλημα ονομάζεται Binary Classification, ενώ όταν υπάρχουν περισσότερες από δύο τότε ονομάζεται Multi-Class Classification.
2. **Παλινδρόμησης (Regression):** Όταν η πρόβλεψη αφορά Συνεχείς Μεταβλητές (Continuous Variables), όπως για παράδειγμα η πρόβλεψη των πωλήσεων, οι τιμές σπιτιών κ.α.
3. **Πρόβλεψη Χρονοσειρών (Time-Series Forecasting):** Είναι η διαδικασία πρόβλεψης μελλοντικών γεγονότων από μια παρατηρούμενη χρονοσειρά, μελετώντας τους παράγοντες που έχουν επιφέρει αλλαγές κατά τη διάρκεια των παρελθοντικών γεγονότων και μπορούν να επηρεάσουν τις μελλοντικές τιμές. Οι παράγοντες αυτοί μπορούν να σχετίζονται με [20]:
    a. **Εποχικότητα (Seasonal):** Είναι ένα εποχιακό μοτίβο που εμφανίζεται όταν μια χρονοσειρά επηρεάζεται από εποχιακούς παράγοντες, όπως την περίοδο του έτους ή την ημέρα της εβδομάδας. Η εποχικότητα είναι πάντα με σταθερή και γνωστή συχνότητα.
    b. **Τάση (Trend):** Η τάση μπορεί να υφίσταται σε μακροπρόθεσμη αύξηση ή μείωση των δεδομένων. Δεν χρειάζεται να είναι γραμμική. Πολλές φορές αναφέρεται ως

«αλλαγή κατεύθυνσης» όταν, για παράδειγμα, η χρονοσειρά από μια αυξητική τάση μεταβαίνει σε μια φθίνουσα τάση.

c. **Κυκλικότητα (Cyclic):** Ένας κύκλος εμφανίζεται όταν τα δεδομένα παρουσιάζουν αυξήσεις και μειώσεις που δεν έχουν σταθερή συχνότητα. Αυτές οι διακυμάνσεις οφείλονται συνήθως σε οικονομικές συνθήκες και συχνά σχετίζονται με τον «Επιχειρηματικό Κύκλο». Η διάρκεια των διακυμάνσεων είναι συνήθως μεγαλύτερη των δύο ετών.

Στον Πίνακας 2, παρατίθενται οι αλγόριθμοι Εποπτευόμενης Μηχανικής Μάθησης, καθώς επίσης τα πλεονεκτήματα και μειονεκτήματα τους αντίστοιχα:

Αλγόριθμος	Περιγραφή/Βασικά Χαρακτηριστικά	Τύπος Επίλυσης	Πλεονεκτήματα/Μειονεκτήματα
Naïve Bayes Classifier	Βασίζεται στο θεώρημα του Bayes και κατηγοριοποιεί κάθε τιμή ως ανεξάρτητη από οποιαδήποτε άλλη τιμή. Προβλέπει μια κατηγορία, με βάση ένα δεδομένο σύνολο Features, χρησιμοποιώντας πιθανότητες.	Classification	+ Μικρή περίοδος εκπαίδευσης + Απλή εφαρμογή + Ταιριάζει καλύτερα σε κατηγορικές μεταβλητές − Υποθέτει ότι όλα τα Features είναι ανεξάρτητα (γεγονός που σπάνια συμβαίνει σε πραγματικά σενάρια) − Μηδενική συχνότητα − Η εκτίμηση μπορεί να είναι λανθασμένη σε κάποιες από περιπτώσεις
Support Vector Machine (SVM)	Αναλύει δεδομένα που έχουν χρησιμοποιηθεί για Classification και Regression αναλύσεις. Συγκεκριμένα, τα δεδομένα φιλτράρονται σε κατηγορίες, κάτι το οποίο επιτυγχάνεται με ένα σύνολο εκπαιδευμένων παραδειγμάτων. Το κάθε σύνολο χαρακτηρίζεται ότι ανήκει σε μια από τις δύο κατηγορίες. Στο τελευταίο στάδιο του αλγορίθμου, δημιουργείται ένα μοντέλο ώστε να διαχωρίζει τις νέες τιμές στη μια ή την άλλη κατηγορία.	Classification	+ Αποτελεσματικότητα σε πολυδιάστατα δεδομένα + Καλή απόδοση σε μικρά Datasets + Ικανός να επιλύσει μη-γραμμικά προβλήματα − Μη-αποτελεσματικότητα σε Big Data − Απαιτεί τη σωστή επιλογή Kernel μεθόδου
Logistic Regression	Επικεντρώνεται στην εκτίμηση της πιθανότητας να συμβεί ένα γεγονός βάσει προηγούμενων δεδομένων. Χρησιμοποιείται για την κατηγοριοποίηση μίας δυαδικής εξαρτημένης μεταβλητής όπου μόνο δύο τιμές, το 0 και το 1, αντιπροσωπεύουν τα αποτελέσματα (π.χ. No/Yes, Sad/Happy).	Classification	+ Δεν αποτελεί αλγόριθμο επιρρεπή στο Overfitting (εκτός από τις περιπτώσεις πολυδιάστατων Datasets) + Είναι αποτελεσματικός όταν το Dataset έχει Features που είναι γραμμικά διαχωρίσιμα + Εύκολος στην εφαρμογή και αποτελεσματικός στη διαδικασία εκπαίδευσης − Δεν πρέπει να εφαρμόζεται όταν ο αριθμός των παρατηρήσεων είναι μικρότερος από εκείνον των Features

			– Εφαρμόζεται η υπόθεση της γραμμικότητας (κάτι που στην πραγματικότητα σπάνια συμβαίνει) – Χρησιμοποιείται μόνο για την πρόβλεψη διακριτών καταστάσεων
Linear Regression	Χρησιμοποιείται για να εξηγήσει τη γραμμικότητα της σχέσης μιας εξαρτημένης μεταβλητής και μίας ή πολλών ανεξάρτητων μεταβλητών.	Regression	+ Εύκολη κατανόηση + Σαφής αντίληψη των παραγόντων που επηρεάζουν περισσότερο το μοντέλο – Δυσκολία ανίχνευσης περίπλοκων σχέσεων μεταξύ των μεταβλητών – Τάση για Overfitting
Decision Trees and Ensemble Methods	Αποτελούν μια δενδροειδή δομή, σαν διαγράμματα ροής, που χρησιμοποιούν διακλαδώσεις για να απεικονίσουν κάθε πιθανό αποτέλεσμα απόφασης και τις συνέπειες του. Κάθε κόμβος του δέντρου αντιπροσωπεύει μια δοκιμή μιας συγκεκριμένης μεταβλητής, ενώ κάθε διακλάδωση είναι το αποτέλεσμα αυτής της δοκιμής. Παραδείγματα που βασίζονται σε Decision Trees είναι οι αλγόριθμοι: Classification and Regression Tree (CART), Random Forest και Gradient Boosting.	Classification Regression Time-Series Forecasting	+ Επίλυση μη-γραμμικών προβλημάτων + Επιτυγχάνεται άριστη ακρίβεια σε πολυδιάστατα δεδομένα + Εύκολη απεικόνιση και επεξήγηση – Παρουσιάζονται προβλήματα Overfitting, που ίσως προσπερνιόνται με Random Forest και Gradient Boosting (Ensemble Methods) – Μία μικρή αλλαγή στα δεδομένα μπορεί να επιφέρει μια μεγάλη αλλαγή στη δομή του βέλτιστου δέντρου απόφασης – Οι υπολογισμοί μπορεί να είναι αρκετά περίπλοκοι
Random Forests	Είναι μια μέθοδος εκμάθησης συνόλου, που συνδυάζει πολλούς αλγόριθμους για να παράγει καλύτερα αποτελέσματα για προβλήματα κατηγοριοποίησης και παλινδρόμησης. Κάθε μεμονωμένος Classifier είναι αδύναμος, αλλά όταν συνδυάζεται με άλλους, μπορεί να παράγει εξαιρετικά αποτελέσματα. Ο αλγόριθμος ξεκινά με ένα «δέντρο αποφάσεων» (ένα δέντρο τύπου γραφήματος ή μοντέλου αποφάσεων) και ένα input στην κορυφή του. Στη συνέχεια, προσπελαύνει το δέντρο με αποτέλεσμα τα δεδομένα να τμηματοποιούνται σε όλο και μικρότερα σύνολα, βάσει συγκριμένων μεταβλητών.	Classification Regression Time-Series Forecasting	+ Ανθεκτικότητα σε ακραίες τιμές (Outliers) + Θετική ανταπόκριση σε μη-γραμμικά δεδομένα + Μικρή πιθανότητα για Overfitting + Αποτελεσματική εκτέλεση σε μεγάλα Datasets + Παρέχεται καλύτερη ακρίβεια σε σχέση με τους υπόλοιπους αλγορίθμους κατηγοριοποίησης – «Προκατάληψη» στις κατηγορικές μεταβλητές – Καθυστερήσεις στην εκμάθηση – Ακαταλληλότητα για γραμμικές μεθόδους με πολλά Sparse Features

Πίνακας 2: Πλεονεκτήματα και Μειονεκτήματα Αλγορίθμων Supervised ML

3.2.1.1 Η εξέλιξη των Tree-based Αλγορίθμων

Με την πάροδο των χρόνων, οι αλγόριθμοι που βασίζονται σε δέντρα αποφάσεων έχουν παρουσιάσει ιδιαίτερη εξέλιξη. Ο XGBoost αλγόριθμος ο οποίος έχει χρησιμοποιηθεί για την υλοποίηση των ML μοντέλων της εφαρμογής "Matapae" ανήκει σε αυτή τη κατηγορία αλγορίθμων και αποτελεί την πιο πρόσφατη και ολοκληρωμένη προσέγγιση. Η πορεία της εξέλιξης τους ανά τα έτη, περιγράφεται παρακάτω [21] [22] [23] [24]:

- **Decision Trees:** Πρόκειται για τη γραφική αναπαράσταση των πιθανών επιλογών επίλυσης για τη λήψη μίας απόφαση, που λαμβάνεται βάσει καθορισμένων συνθηκών ή κριτηρίων.
- **Bagging (ή Bootstrap Aggregating):** Απλή και πολύ ισχυρή Ensemble Meta-algorithm μέθοδος, η οποία συνδυάζει προβλέψεις που προκύπτουν από πολλαπλά δέντρα αποφάσεων μέσω ενός πλειοψηφικού μηχανισμού.
    - Αποτελεί την εφαρμογή της διαδικασία του **Bootstraping** σε έναν Machine Learning αλγόριθμο υψηλής διακύμανσης (High-variance):
        - Αναφέρεται σε τυχαία δειγματοληψία με αντικατάσταση.
        - Επιτρέπει την καλύτερη κατανόηση της προκατάληψης (Bias) και της διακύμανσης (Variance) με το σύνολο των δεδομένων.
        - Είναι μια τεχνική δειγματοληψίας στην οποία δημιουργούμε υποσύνολα παρατηρήσεων από το αρχικό σύνολο δεδομένων με αντικατάσταση, όπου το μέγεθος των υποσυνόλων που προκύπτουν είναι ίδιο με το μέγεθος του αρχικού συνόλου.
        - Η επιλογή όλων των παραδειγμάτων στο σύνολο δεδομένων έχει ίσες πιθανότητες. Ως εκ τούτου, επιτυγχάνεται η καλύτερη κατανόηση του μέσου όρου και της τυπικής απόκλισης του συνόλου των δεδομένων.
    - Η γενική ιδέα του Bagging είναι ο συνδυασμός των αποτελεσμάτων από πολλαπλά ML μοντέλα, ώστε να προκύψει ένα γενικευμένο αποτέλεσμα.
    - Η τεχνική Bagging χρησιμοποιεί τα υποσύνολα για να λάβει μια αντιπροσωπευτική αντίληψη της κατανομής ολόκληρου του Dataset. Το μέγεθος των υποσυνόλων που δημιουργήθηκαν πιθανώς να είναι μικρότερο από το πραγματικό Dataset.
    - Στην τεχνική αυτή κάθε μοντέλο τρέχει ανεξάρτητα και τα αποτελέσματα συγκεντρώνονται στο τέλος χωρίς προτίμηση σε κάποιο μοντέλο.
- **Random Forest:** Αλγόριθμος βασισμένος στη τεχνική Bagging όπου μόνο τα υποσύνολα των Features επιλέγονται τυχαία ώστε να δημιουργήσουν μια συλλογή από δέντρα αποφάσεων.
- **Boosting:** Σε αυτή την τεχνική, τα μοντέλα που έχουν αναπτυχθεί διαδοχικά μειώνουν τα Errors από τα προηγούμενα μοντέλα, ενώ παράλληλα αυξάνεται η επιρροή των μοντέλων με την καλύτερη απόδοση στο τελικό μοντέλο.
    - Το κάθε μοντέλο προσπαθεί να διορθώσει τα Errors του επόμενου.
    - Η επιτυχία κάθε μοντέλου εξαρτάται από αυτή του προηγούμενου.

- Τα εκπαιδευόμενα μοντέλα μαθαίνουν διαδοχικά με την εφαρμογή πρώιμων απλών μοντέλων στα δεδομένα εκπαίδευσης και έπειτα αναλύοντας τα δεδομένα για Errors. Με άλλα λόγια, εφαρμόζονται διαδοχικά δέντρα από ένα τυχαίο δείγμα και σε κάθε βήμα ο σκοπός είναι να βελτιστοποιηθεί η διαφορά του Error από το προηγούμενο δέντρο εκπαίδευσης.
- Όταν ένα Input δεν μπορεί να κατηγοριοποιηθεί από κάποια υπόθεση, το βάρος (Weight) του Input αυξάνεται, έτσι ώστε η επόμενη υπόθεση να είναι πιο πιθανό να το κατηγοριοποιήσει κατάλληλα.
- Ο συνδυασμός των διαδοχικών εκπαιδεύσεων μετατρέπει, στο τέλος της διαδικασίας, τα αδύναμα εκπαιδευμένα μοντέλα (Weak Models) στο καλύτερα εκπαιδευμένο μοντέλο.
- Πρόκειται για μια Ensemble τεχνική που δημιουργεί μια συλλογή μοντέλων.

- **Gradient Boosting:** Βασισμένο στη χρήση αλγορίθμων Gradient Descent για την ελαχιστοποίηση των Errors μέσω της ανάπτυξης διαδοχικών μοντέλων.
  - Αφορά μια Machine Learning τεχνική για Regression και Classification προβλήματα, η οποία παράγει μοντέλα πρόβλεψης μέσω της λογικής του συνδυασμού Weak Models (Decision Trees).
  - Το τελικό μοντέλο πρόβλεψης δομείται σε στάδια και βελτιστοποιείται βάσει της Loss Functions που έχει οριστεί.
  - Όπως σε κάθε περίπτωση Supervised Machine Learning, έτσι και στην τεχνική Gradient Boosting, ορίζεται η Loss Function με σκοπό την ελαχιστοποίηση της τιμής της. Συνήθως, ως Loss Function ορίζεται είτε το Mean Squared Error (MSE), είτε το Root Mean Squared Error (RMSE).
- **XGBoost:** Πρόκειται για έναν βελτιστοποιημένο Gradient Boosting αλγόριθμο μέσω παράλληλης επεξεργασίας μείωσης του μεγέθους του δέντρου αποφάσεων (Tree Pruning). Η εν λόγω μείωση, επιτυγχάνεται με τη διαγραφή των τμημάτων του δέντρου που δεν είναι σημαντικά και είναι περιττά για τη κατηγοριοποίηση των περιπτώσεων. Επιπροσθέτως, διαχειρίζεται αποτελεσματικά τις ελλιπείς τιμές (Missing Values), αποφεύγοντας περιπτώσεις Overfitting ή Bias κατά τη διαδικασία εκπαίδευσης των μοντέλων μηχανική μάθησης.
  - Αποτελεί έναν τέλειο συνδυασμό τεχνικών βελτιστοποίησης λογισμικού και υλικού, ώστε να προκύψουν εξαιρετικά αποτελέσματα χρησιμοποιώντας λιγότερους υπολογιστικούς πόρους και σε όσο το δυνατόν συντομότερο χρονικό διάστημα.
  - Βελτιώνει το Gradient Boosting Machine (GBM) Framework, καθώς υποστηρίζει τόσο παρεμβάσεις βελτιστοποίησης στους πόρους των υπολογιστικών συστημάτων στα οποία εφαρμόζεται (System Optimization – Parallelization, Tree Pruning, Hardware Optimization), αλλά και τεχνικές βελτιστοποίησης στον ίδιο τον αλγόριθμο (Algorithmic Enhancements – Regularization, Sparsity Awareness, Weighted Quantile Sketch, Cross-validation).
    - **Parallelization:** Η προσέγγιση αυτή αφορά τη δημιουργία διαδοχικών δέντρων με παράλληλη υλοποίηση. Η εκπαίδευση πραγματοποιείται με

έναν εξωτερικό βρόχο που απαριθμεί τα φύλλα των κόμβων που περιλαμβάνει το δέντρο και με έναν εσωτερικό βρόχο ο οποίος υπολογίζει τα Features. Η εμφώλευση των βρόχων περιορίζει την παράλληλη επεξεργασία καθώς απαιτείται η ολοκλήρωση του εσωτερικού βρόχου, ο οποίος είναι και πιο απαιτητικός υπολογιστικά από τους δύο, με αποτέλεσμα να μη μπορεί να ξεκινήσει η εκτέλεση του εξωτερικού. Επομένως, για να βελτιωθεί ο χρόνος εκτέλεσης, η σειρά εκτέλεσης των βρόχων εναλλάσσεται βάσει ενός αρχικού ελέγχου σε όλες τις περιπτώσεις και της ταξινόμησης τους χρησιμοποιώντας παράλληλα Threads.

- **Tree Pruning:** Κατά τη διάσπαση των δέντρων βάσει του GBM Framework, το κριτήριο διακοπής της διάσπασης είναι «άπληστο» και εξαρτάται από το κριτήριο Negative Loss στο σημείο της διάσπασης. Πιο συγκεκριμένα, όταν ένας αλγόριθμος αποφασίζει μέσω της άπληστης επιλογής, σημαίνει ότι επιλέγει αυτό που φαίνεται ως η καλύτερη λύση της τρέχουσας κατάστασης με κάποιο απλό κριτήριο. Επομένως, βάσει της χρήσης της παραμέτρου “Max_depth” η διαδικασία Tree Pruning ξεκινά από το “Max_depth” και προς τα πίσω, βελτιώνοντας σημαντικά την υπολογιστική απόδοση.
- **Hardware Optimization:** Ο αλγόριθμος αυτός είναι σχεδιασμένος με τέτοιο τρόπο ώστε να κάνει αποτελεσματική χρήση των υπολογιστικών πόρων. Αυτό επιτυγχάνεται με το “Cache Awareness”, δηλαδή με την κατανομή των εσωτερικών Buffers σε κάθε Thread για την αποθήκευση των Gradient Statistics. Επιπλέον, βελτιστοποιεί τον διαθέσιμο χώρο στον δίσκο όσο διαχειρίζεται Big Data που δεν χωρούν στη μνήμη.
- **Regularization:** Οι τεχνικές της κανονικοποίησης (Regularization), επιβάλλουν «ποινή» στις περιπτώσεις που τα μοντέλα εκπαιδεύονται με Datasets που παρουσιάζουν μεγάλο αριθμό Features, γεγονός που τα καθιστά αρκετά περίπλοκα. Με τις επιμέρους τεχνικές LASSO (Least Absolute Shrinkage and Selection Operator – L1) και Ridge (L2) Regularization, δημιουργούνται λιγότερο περίπλοκα μοντέλα καθώς διευθετούνται ζητήματα Overfitting και επιλογής των κατάλληλων Features. Όταν σε ένα μοντέλο παλινδρόμησης (Regression Model) χρησιμοποιείται L1 τότε ονομάζεται LASSO Regression, ενώ όταν χρησιμοποιείται L2 ονομάζεται Ridge Regression. Η βασική διαφορά των προαναφερθέντων προσεγγίσεων είναι το μέγεθος του συντελεστή που προστίθεται ως «ποινή» στη Loss Function. Πιο συγκεκριμένα, στην περίπτωση L1 προστίθεται η απόλυτη τιμή μεγέθους με σκοπό να αποδυναμώνει τα λιγότερο σημαντικά Features, ενώ στην L2 προστίθεται το τετράγωνο μεγέθους του συντελεστή, συμβάλλοντας στην αποφυγή ζητημάτων Overfitting [25].

- **Sparsity Awareness:** Στον XGBoost υπάρχει η δυνατότητα εισαγωγής Sparse Features με σκοπό την κωδικοποίηση των Missing Values και των μηδενικών τιμών. Επομένως, ο αλγόριθμος έχει τη δυνατότητα να αναγνωρίζει διαφορετικούς τύπους από Sparsity Patterns και κατά τη διαδικασία εκμάθησης, να προσπελαύνει πιο αποδοτικά κάθε κόμβο του δέντρου που δεν παρουσιάζει ελλείπεις τιμές (Non-missing Values), βάσει της Training Loss Function.
- **Weighted Quantile Sketch:** Αφορά έναν ενσωματωμένο και κατανεμημένο αλγόριθμο για την κατά προσέγγιση εκπαίδευση του δέντρου. Μέσω του αλγορίθμου, προτείνονται υποψήφια σημεία διαχωρισμού έτσι ώστε να βρεθούν, με αποτελεσματικό τρόπο, τα καλύτερα μεταξύ των σταθμισμένων (Weighted) Datasets.
- **Cross-validation:** Πρόκειται για μια ενσωματωμένη μέθοδο επικύρωσης των αποτελεσμάτων κάθε επανάληψης, εξαλείφοντας την ανάγκη να προγραμματιστεί ξεχωριστά αυτή η αναζήτηση και να καθοριστεί ο ακριβής αριθμός των επαναλήψεων Boosting που απαιτούνται σε μία μόνο εκτέλεση.

3.2.2 Αλγόριθμοι μη-Εποπτευόμενης Μηχανικής Μάθησης (Unsupervised Machine Learning Algorithms)

Οι αλγόριθμοι μη-Εποπτευόμενης Μάθησης (Unsupervised Learning), στοχεύουν στον εντοπισμό εγγενών μοτίβων σε Unlabeled δεδομένα. Δεν υπάρχει καθορισμένο «κλειδί» απάντησης ή ο ανθρώπινος παράγοντας που να παρέχει οδηγίες και να επηρεάζει την εκπαίδευση του μοντέλου. Αντιθέτως, ο αλγόριθμος καθορίζει τις συσχετίσεις και τις σχέσεις αναλύοντας τα διαθέσιμα δεδομένα, ώστε να τα οργανώσει κατάλληλα βάσει των χαρακτηριστικών τους και να τα ομαδοποιήσει σε Clusters ή να τα διασυνδέσει με τέτοιο τρόπο ώστε να είναι οργανωμένα [26]. Η διαδικασία εκπαίδευσης ενός μοντέλου με χρήση Unsupervised Learning αλγορίθμου περιγράφεται σχηματικά στην Εικόνα 3 [19].

Εικόνα 3: Διαδικασία Εκπαίδευσης Unsupervised Learning Model

Οι Supervised Learning αλγόριθμοι, μπορούν να επιλύσουν τέσσερις κατηγορίες προβλημάτων [18] [19]:

1. **Ομαδοποίησης (Clustering):** Κατά την ομαδοποίηση παρατηρήσεων ενός συνόλου δεδομένων, δημιουργούνται ομάδες των οποίων οι παρατηρήσεις είναι περισσότερο όμοιες μεταξύ τους, βάσει κριτηρίων, σε σχέση με τις παρατηρήσεις άλλων ομάδων. Η διαδικασία αυτή χρησιμοποιείται συχνά για να διαχωρίσει ολόκληρο το Dataset σε διάφορες ομάδες. Η ανάλυση κάθε ομάδας συμβάλει στην ανίχνευση εγγενών μοτίβων (Patterns) που πιθανώς να υπάρχουν.
2. **Κανόνων Συσχέτισης (Association Rules):** Οι κανόνες συσχέτισης επιτρέπουν τον καθορισμό συσχετίσεων μεταξύ οντοτήτων που εμπεριέχονται σε συλλογές Big Data, προσδιορίζοντας σχέσεις μεταξύ των μεταβλητών που τις περιγράφουν. Για παράδειγμα, σε μια ανάλυση για το καλάθι αγορών σε ένα ηλεκτρονικό κατάστημα, η ανάλυση μπορεί να αφορά τον συνδυασμό των προϊόντων που συνήθως αγοράζονται μαζί. Σκοπός της, να γίνονται οι κατάλληλες προτάσεις στον χρήστη, όπου θα τον παροτρύνουν να προσθέσει στο καλάθι του, επιπλέον προϊόντα που είναι πιθανό να αγοράσει στην τρέχουσα παραγγελία, ώστε να επιτευχθεί η αύξηση του μέσου όρου της αξίας στο καλάθι αγορών, κάτι το οποίο τελικά σημαίνει, την αύξηση των πωλήσεων.
3. **Ανίχνευσης Ανωμαλιών (Anomaly Detection):** Χρησιμοποιείται για τη διάγνωση σφαλμάτων. Ο αλγόριθμος εκπαιδεύεται ώστε να μπορεί να αντιλαμβάνεται την εμφάνιση ασυνήθιστων τιμών δεδομένων, τα οποία δεν συμβαδίζουν με την ορθή λειτουργία των συστημάτων βάσει των προδιαγραφών που έχουν οριστεί. Ενδεικτικές περιπτώσεις είναι:
    a. Διάγνωση σφαλμάτων
    b. Μη-εξουσιοδοτημένη πρόσβαση σε συστήματα
    c. Εντοπισμός συμβάντων απάτης
4. **Μείωσης Διαστάσεων (Dimensionality Reduction):** Πρόκειται για αλγορίθμους ή στατιστικές τεχνικές μετατροπής δεδομένων που μειώνουν τον αριθμό των χαρακτηριστικών σε ένα Dataset πολλών διαστάσεων. Το νέο σύνολο δεδομένων που προκύπτει, διατηρεί την ακεραιότητα των δεδομένων και των χαρακτηριστικών του, αλλά με λιγότερες διαστάσεις. Ορισμένες από τις πιο δημοφιλείς τεχνικές είναι:
    a. Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis)
    b. Αποσύνθεση Μοναδιαίας Τιμής (Singular Value Decomposition)
    c. Λανθάνουσα Κατανομή Dirichlet (Latent Dirichlet Allocation)

Στον Πίνακας 3, παρατίθενται οι αλγόριθμοι μη-Εποπτευόμενης Μηχανικής Μάθησης, καθώς επίσης τα πλεονεκτήματα και μειονεκτήματα τους αντίστοιχα:

Αλγόριθμος	Περιγραφή/Βασικά Χαρακτηριστικά	Τύπος Επίλυσης	Πλεονεκτήματα/Μειονεκτήματα
k-Means Clustering	Χρησιμοποιούνται για να κατηγοριοποιήσουν Unlabeled Data. Επομένως, λαμβάνουν και αναλύουν σύνολα δεδομένων τα οποία δεν έχουν προσδιορισμένες κατηγορίες. Λειτουργούν ανιχνεύοντας ομάδες μέσα στα δεδομένα. Το πλήθος των ομάδων αντιπροσωπεύεται από τη μεταβλητή k. Στη συνέχεια, ο αλγόριθμος αναθέτει επαναληπτικά κάθε στοιχείο των δεδομένων σε μία από τις ομάδες που ταιριάζει περισσότερο βάσει των χαρακτηριστικών του. Στόχος είναι οι ομάδες που θα προκύψουν να έχουν νόημα και χρησιμότητα.	Clustering	+ Ευκολία στην υλοποίηση και την προσαρμογή σε νέα παραδείγματα + Διαχείριση μεγάλων Datasets + Εγγυημένη σύγκλιση (Convergence) + Δυνατότητα γενίκευσης σε Clusters διαφορετικών σχημάτων και μεγεθών − Ευαισθησία σε ακραίες τιμές (Outliers) − Δυσκολία στη «χειροκίνητη» επιλογή των k τιμών − Εξάρτηση από τις αρχικές τιμές − Μειωμένη επεκτασιμότητα και δυνατότητα Scalability σε μαζικά δεδομένα (όσο οι διαστάσεις τους αυξάνονται)
Apriori	Αλγόριθμος που χρησιμοποιείται για την εξόρυξη συχνά εμφανιζόμενων στοιχείων και τη θέσπιση κανόνων συσχέτισης από μια βάση δεδομένων που καταγράφει όλες τις συναλλαγές. Οι κανόνες αξιολογούνται με τις παραμέτρους: Support που αναφέρεται στη συχνότητα εμφάνισης των στοιχείων, Confidence που αφορά την πιθανότητα να συμβεί ο κάθε συνδυασμός και Lift που δηλώνει την ισχύ του κανόνα. Τα στοιχεία (π.χ. προϊόντα) σε μια συναλλαγή σχηματίζουν ένα σύνολο. Ο αλγόριθμος ξεκινά προσδιορίζοντας τη συχνότητα μεμονωμένων στοιχείων (στοιχεία συχνότητας μεγαλύτερης ή ίσης με το Support) και συνεχίζει να επεκτείνεται σε μεγαλύτερα και συχνότερα εμφανιζόμενα σύνολα [27] [28].	Association Rules	+ Ευκολία στην κατανόηση + Απλή εφαρμογή των διαδικασιών Merge και Squash σε μεγάλα σύνολα στοιχείων τεράστιων βάσεων δεδομένων − Απαιτείται μεγάλο πλήθος υπολογισμών στην περίπτωση που τα σύνολα στοιχείων είναι μεγάλα και η ελάχιστη τιμή της παραμέτρου Support διατηρείται στο ελάχιστο − Απαιτείται πλήρης προσπέλαση της βάσης δεδομένων [29]
Principal Component Analysis (PCA)	Μέθοδος για τη μείωση των διαστάσεων στα μεγάλα Datasets, μετασχηματίζοντας τα σε μικρότερης διάστασης και διατηρώντας όσο το δυνατόν περισσότερες πληροφορίες. Ουσιαστικά, ο αλγόριθμος ανακαλύπτει έναν υποχώρο που διατηρεί τη διακύμανση των δεδομένων. Ο υποχώρος ορίζεται από τα κύρια ιδιοδιανύσματα του πίνακα συνδιακύμανσης των δεδομένων.	Dimensionality Reduction	+ Μειωμένα Features + Βελτίωση απόδοσης + Μειωμένο Overfitting − Λιγότερο κατανοητό αποτέλεσμα − Πιθανότητα να χαθεί πληροφορία − Προηγείται Data Standardization ή Normalization πριν την εφαρμογή του αλγορίθμου

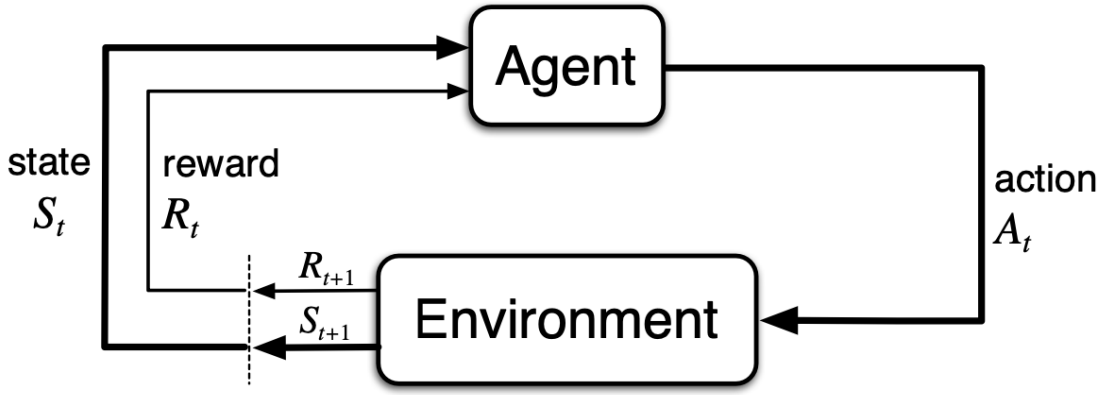
Isolation Forest	Για το σκοπό αυτό χρησιμοποιείται ο αλγόριθμος Random Forest για να ανιχνεύσει Outliers στο Dataset. Κατά την εφαρμογή του, προσπαθεί να χωρίσει ή να διαιρέσει τα Data Points των παρατηρήσεων έτσι ώστε κάθε μια από αυτές, να είναι απομονωμένη από τις υπόλοιπες. Οι ελλείψεις ομαλότητας που ανιχνεύονται, συνήθως, βρίσκονται μακριά από τα Clusters των Data Points, επομένως είναι ευκολότερη η απομόνωση τους σε σύγκριση με τα Regular Data Points [30].	Anomaly Detection	+ Δεν απαιτείται Normalization των Features + Αποτελεσματικότητα όταν η κατανομή των Features δεν είναι θεωρητική/υποθετική (π.χ. Κανονική Κατανομή) + Χρησιμοποιώντας ελάχιστα μόνο παραδείγματα το μοντέλο γίνεται ισχυρό (Robust) + Ευκολία στη διαδικασία βελτιστοποίησης + Διαθέτει λεπτομερή τεκμηρίωση επομένως εφαρμόζεται εύκολα − Περιπλοκότητα στην οπτικοποίηση των αποτελεσμάτων − Στις περιπτώσεις που δεν βελτιστοποιηθεί σωστά, η εφαρμογή μπορεί να είναι μια μακρά διαδικασία και ίσως απαιτηθεί μεγαλύτερη υπολογιστική ισχύς [31]

Πίνακας 3: Πλεονεκτήματα και Μειονεκτήματα Αλγορίθμων Unsupervised ML

3.2.3 Αλγόριθμοι Ενισχυτικής Μηχανικής Μάθησης (Reinforcement Machine Learning)

Οι αλγόριθμοι Ενισχυτικής Μάθησης (Reinforcement Learning), χρησιμοποιούνται κυρίως σε προβλήματα που απαιτούν συνεχή και διαδοχική λήψη αποφάσεων. Σε αυτή την περίπτωση μάθησης δεν χρειάζεται να έχουμε δεδομένα εκ των προτέρων, παρά μόνο καθορισμένους κανόνες βάσει των οποίων ο Learning Agent θα προσπαθήσει να εξερευνήσει διαφορετικές επιλογές και δυνατότητες, εξετάζοντας και αξιολογώντας κάθε αποτέλεσμα για να προσδιορίσει ποιο είναι το βέλτιστο. Συγκεκριμένα, σε κάθε χρονικό βήμα, αντιλαμβάνεται την κατάσταση (State) του περιβάλλοντος (Environment) που διερευνά, επιλέγει να κάνει μια ενέργεια και βάσει του feedback που λαμβάνει, «μαθαίνει» να δρα με τέτοιο τρόπο ώστε να μεγιστοποιεί την ανταμοιβή (Reward) μακροπρόθεσμα. Παράλληλα ο Learning Agent πρέπει να διατηρεί την ισορροπία μεταξύ εξερεύνησης και αξιοποίησης εκτελώντας μια ποικιλία “Trial and Error Actions” με σκοπό να ευνοούνται οι ενέργειες που θα επιφέρουν μελλοντικά τη μέγιστη ανταμοιβή. Ο αλγόριθμος χαρτογραφώντας τις αποφάσεις (State or Reward) των διαδοχικών ενεργειών που έχει κάνει, ενημερώνει την πολιτική λήψης αποφάσεων ώστε ο Learning Agent να βελτιστοποιήσει την ικανότητα αποφάσεων στις μελλοντικές ενέργειες. Με την εξέλιξη του Deep Learning τα τελευταία χρόνια, εφαρμόζονται κυρίως στην ρομποτική, τα Self-driving αυτοκίνητα, στην αξιολόγηση στρατηγικών στον τομέα των επενδύσεων και τους προσαρμοστικούς ελέγχους (Adaptive Controls).

Η διαδικασία εκπαίδευσης ενός μοντέλου με χρήση Reinforcement Learning αλγορίθμου περιγράφεται σχηματικά στην Εικόνα 4 [1] [18] [19]:



Εικόνα 4: Διαδικασία Εκπαίδευσης Reinforcement Learning Model

Στον Πίνακας 4, παρατίθενται τα χαρακτηριστικά των αλγορίθμων όπως διαμορφώνονται στα Τεχνητά Νευρωνικά Δίκτυα, ως παράδειγμα Ενισχυτικής Μηχανικής Μάθησης, καθώς επίσης τα πλεονεκτήματα και μειονεκτήματα τους αντίστοιχα:

Αλγόριθμος	Περιγραφή/Βασικά Χαρακτηριστικά	Τύπος Επίλυσης	Πλεονεκτήματα/Μειονεκτήματα
Artificial Neural Networks (ANN)	Περιλαμβάνουν εγγραφές (Units) διατεταγμένες σε μια σειρά από στρώσεις (Layers), κάθε μια από τις οποίες συνδέεται με τα Layers εκατέρωθεν. Σχεδιασμένοι να λειτουργούν όπως ο εγκέφαλος βάσει του τρόπου που επεξεργάζεται τις πληροφορίες. Πρακτικά, αποτελούνται από ένα μεγάλο αριθμό διασυνδεδεμένων στοιχείων, που λειτουργούν από κοινού για την επίλυση προβλημάτων. Επίσης, εκπαιδεύονται με τα παραδείγματα και την εμπειρία. Ικανοί να μοντελοποιήσουν μη-γραμμικές σχέσεις σε πολυδιάστατα δεδομένα (High-Dimensional Data) ή όπου η σχέση μεταξύ των Input Variables είναι δύσκολο να κατανοηθεί.	Reinforcement	+ Ανεκτικότητα σφαλμάτων + Ικανότητα εκμάθησης και μοντελοποίησης περίπλοκων μη-γραμμικών σχέσεων + Δυνατότητα γενίκευσης σε άγνωστα δεδομένα – Αρκετός χρόνος εκμάθησης – Μη εγγυημένη σύγκλιση (Convergence) – Δύσκολη επεξήγηση της λύσης (Black Box) – Εξαρτάται από το Hardware – Απαιτείται ικανότητα του χρήστη να «μεταφράσει» το πρόβλημα

Πίνακας 4: Πλεονεκτήματα και Μειονεκτήματα Αλγορίθμου Reinforcement ML

3.3 Αυτοματοποιημένη Μηχανική Μάθηση (Automated Machine Learning – AutoML)

Η εξαγωγή γνώσης από την επεξεργασία των δεδομένων με αλγορίθμους Machine Learning (ML) είναι πολύπλοκη διεργασία πολλών βημάτων και συνήθως απαιτεί σημαντικούς επιχειρηματικούς πόρους και χρόνο. Για να μειωθεί αυτό το κόστος ανάπτυξης, προέκυψε μια νέα ιδέα αυτοματοποίησης ολόκληρου του ML Pipeline, η αυτοματοποιημένη μηχανική μάθηση (AutoML) [32].

3.3.1 Τι είναι AutoML;

Υπάρχουν διάφοροι ορισμοί του AutoML. Σύμφωνα με τη μελέτη [33], έχει σχεδιαστεί για να μειώνει τις ανάγκες ζήτησης των Data Scientists και να δίνει τη δυνατότητα σε Domain Experts να δημιουργούν μοντέλα ML με αυτοματοποιημένο τρόπο, χωρίς την απαίτηση να έχουν πολλές γνώσεις σε στατιστικές μεθόδους και μηχανική μάθηση (ML). Στη μελέτη [34], AutoML ορίζεται ως ένας συνδυασμός αυτοματισμού και ML, όπου εξηγείται ως η αυτοματοποίηση των βημάτων που απαιτούνται για την επίλυση προβλημάτων με τη χρήση μεθόδων ML (ML Pipelines), όταν υπάρχουν περιορισμένοι υπολογιστικών και ανθρωπίνων πόρων. Με τη ραγδαία αύξηση της υπολογιστικής ισχύος, το AutoML βρίσκει πεδίο εφαρμογής, τόσο σε επιχειρηματικό, όσο και ακαδημαϊκό επίπεδο.

Ένα πλήρες σύστημα AutoML, θα πρέπει να συνδυάζει διάφορες τεχνικές για να σχηματίσει ένα εύχρηστο End-to-End ML Pipeline που να περιλαμβάνει τις εξής διαδικασίες:

- Εισαγωγή και προετοιμασία των δεδομένων.
- Επιλογή, εξαγωγή και επεξεργασία των Features (Feature Engineering).
- Επιλογή του κατάλληλου ML αλγορίθμου που να ταιριάζει στις ανάγκες της επίλυσης του προβλήματος.
- Εκπαίδευση (Training) και δοκιμή (Testing) του μοντέλου.
- Αναζήτηση του αποδοτικότερου συνδυασμού παραμετροποίησης του μοντέλου (Hyperparameter Tuning).
- Διαχείριση των προκλήσεων που εγείρει η εφαρμογή των ML μοντέλων σε Large-scale συστήματα.
- Συνεχείς παρακολούθηση και βελτιστοποίηση των ML μοντέλων σε νέα δεδομένα, τα οποία προκύπτουν μετά από την ανάπτυξη και εφαρμογή τους.

Πολλές εταιρείες τεχνητής νοημοσύνης έχουν δημιουργήσει συστήματα για να βοηθήσουν στελέχη με λίγη ή καθόλου γνώση ML, να δημιουργήσουν υψηλής ποιότητα μοντέλα προσαρμοσμένα στις ανάγκες τους (π.χ. Cloud AutoML από την Google, H2O.ai, Akkio).

3.3.2 Βασικά Στάδια στη Δημιουργία AutoML Pipelines (AutoML Lifecycle)

Όπως αναφέρθηκε επιγραμματικά στην υποενότητα 3.3.1, κάθε ML Project και κατ' επέκταση ένα AutoML Pipeline, πρέπει να περιλαμβάνει συγκεκριμένες διεργασίες ώστε να είναι αποτελεσματικό για το πρόβλημα που δημιουργήθηκε να επιλύσει. Πιο αναλυτικά, τα στάδια που πρέπει να υποστηρίζει ένα AutoML Pipeline, συμπεριλαμβανομένης της προεργασίας και της τεκμηρίωσης που απαιτεί ένα τέτοιου είδους Project, είναι τα εξής [35] [36]:

1. **Σχεδιασμός Προσέγγισης (Planning):** Το στάδιο του σχεδιασμού περιλαμβάνει την εκτίμηση (ή αξιολόγηση) του σκοπού, τη μέτρηση της επιτυχίας και τη σκοπιμότητα της εφαρμογής τεχνολογιών ML. Χρειάζεται όμως προηγουμένως, να έχει κατανοηθεί πλήρως το έργο (δραστηριότητα) της επιχείρησης και πως μπορεί να χρησιμοποιηθεί η εκάστοτε εφαρμογή ML ώστε να βελτιώσει την υπάρχουσα επιχειρηματική διαδικασία. Επίσης, να έχει αναλυθεί και κατανοηθεί η ισορροπία μεταξύ κόστους και οφέλους που πιθανολογείται ότι θα επιφέρει η λύση, ποια επίπεδα θα επηρεάσει και με ποιον τρόπο. Επιπλέον, να οριστούν σαφώς οι αποδεκτές τιμές μετρήσεων της επιτυχίας των ML μοντέλων για την ίδια την επιχείρηση και το συνολικό οικονομικό όφελος που θα προκύψει. Στο τέλος της μελέτης, τα αποτελέσματα καταγράφονται αναλυτικά στην αναφορά σκοπιμότητας, η οποία καταρτίζεται με τις παρακάτω πληροφορίες:
    a. **Διαθεσιμότητα Δεδομένων (Data Availability):** Περιγραφή και καθορισμός των διαθέσιμων δεδομένων, καθώς είναι αναγκαίο να υπάρχουν αρκετά δεδομένα για την εκμάθηση των μοντέλων.
    b. **Εφαρμογή:** Τεκμηρίωση της ικανότητας που αναμένεται να έχει η εφαρμογή ML, που σχεδιάζεται να υλοποιηθεί, έτσι ώστε να επιλύσει ή να βελτιώσει την τρέχουσα κατάσταση.
    c. **Νομικοί Περιορισμοί:** Αναφέρεται στη διαδικασία έγκρισης από τους αρμόδιους φορείς (στην περίπτωση που κρίνεται απαραίτητη), κατά πόσο η συλλογή δεδομένων γίνεται ηθικά και τι επιπτώσεις θα επιφέρει η εφαρμογή τεχνολογιών ML στην επιχείρηση ή την κοινωνία.
    d. **Ανθεκτικότητα και Επεκτασιμότητα:** Παρατίθενται στοιχεία σχετικά με το πόσο «ισχυρή» είναι η εφαρμογή ML που υλοποιείται. Ακόμα, εξετάζεται η δυνατότητα περαιτέρω επέκτασης της, ώστε να αντιμετωπίσει νέες προκλήσεις που πιθανώς να προκύψουν μελλοντικά.
    e. **Επεξηγησιμότητα (Explainability):** Αφορά την τεκμηρίωση του τρόπου που προέκυψαν τα παραγόμενα αποτελέσματα από τα ML μοντέλα, καθώς και την επεξήγηση της λειτουργίας των εσωτερικών Neural Networks που δημιουργήθηκαν.
    f. **Διαθεσιμότητα Πόρων:** Ανάλυση της διαθεσιμότητας της υπολογιστικής ισχύος του συστήματος, της μνήμης, του δικτύου και των ανθρώπινων πόρων που αφορά τους καταρτισμένους επαγγελματίες που έχουν την ικανότητα να εξυπηρετήσουν τους σκοπούς ενός ML Project.
2. **Προετοιμασία των Δεδομένων (Data Preparation):** Το στάδιο της προετοιμασίας των δεδομένων απαρτίζεται από τις επιμέρους διαδικασίες:

a. **Λήψη και Επισήμανση Δεδομένων (Data Ingestion & Labeling):** Αρχικά, θα πρέπει να αποφασιστεί ο τρόπος βάσει του οποίου θα συλλεχθούν τα δεδομένα (π.χ. εσωτερικά δεδομένα, Open Data, προμήθεια δεδομένων από εξωτερικές πηγές ή δημιουργία νέων). Κάθε μία από τις μεθόδους συλλογής δεδομένων έχει πλεονεκτήματα και μειονεκτήματα, γι' αυτό σε κάποιες περιπτώσεις συλλέγονται δεδομένα με όλες τις μεθόδους. Μετά τη συλλογή τους, πρέπει να επισημανθούν (Labeling). Στην περίπτωση της αγοράς δεδομένων, «καθαρά» και «επισημασμένα» δεδομένα είναι αρκετά δύσκολο να βρεθούν. Για το λόγο αυτό, ίσως χρειαστεί να διαμορφωθούν κατάλληλα. Μόλις αποφασιστεί ποια θα είναι η συλλογή δεδομένων με την οποία θα προχωρήσει η ανάλυση και η κατάρτιση των AutoML μοντέλων, γίνεται η εισαγωγής τους (π.χ. Upload ενός αρχείου.csv). Το Uploaded αρχείο υποβάλλεται σε επεξεργασία και μετατρέπεται σε Machine-readable Format όπως είναι ένα DataFrame στις γλώσσες προγραμματισμού Python ή R.

b. **Καθαρισμός Δεδομένων (Data Cleaning):** Σε αυτό το στάδιο, ο καθαρισμός των δεδομένων πραγματοποιείται με τις τεχνικές ανάθεσης τιμών στις κενές τιμές (Missing Values), την αφαίρεση των Outliers και τη μείωση του θορύβου (Dealing with Noisy Data). Είναι απαραίτητο να δημιουργηθεί ένα Data Pipeline ώστε αυτή η διαδικασία να αυτοματοποιηθεί και να επαληθευτεί η ποιότητα των δεδομένων.

c. **Επεξεργασία Δεδομένων (Data Processing):** Το στάδιο της επεξεργασίας δεδομένων περιλαμβάνει την επιλογή των Features, τη διαχείριση των Imbalanced Classes, Feature Engineering, Data Augmentation και τέλος, Normalizing και Scaling των δεδομένων.

d. **Διαχείριση Δεδομένων (Data Management):** Αφορά τις λύσεις που θα επιλεχθούν για την αποθήκευση των δεδομένων, τη διαχείριση των εκδόσεων που έχουν παραχθεί έπειτα από την επεξεργασία τους (Data Versioning) και το οποίο συμβαίνει για σκοπούς επαναχρησιμοποίησης, την αποθήκευση των Metadata και τελικά τη δημιουργία Extract-Transform-Load (ETL) Pipelines. Τα ETL Pipelines θα εξασφαλίσουν τη συνεχή ροή δεδομένων που θα ανατροφοδοτεί τα μοντέλα για περαιτέρω εκμάθηση.

3. **Επιλογή Μοντέλου (Data Selection):** Περιλαμβάνει την επιλογή από μια ποικιλία μοντέλων για την κατασκευή και την εκπαίδευση ML μοντέλων. Ορισμένα μοντέλα ενδέχεται να παρέχουν καλύτερη ακρίβεια σε ένα σύνολο δεδομένων με συγκεκριμένα χαρακτηριστικά ή για διαφορετικό σκοπό, όπως για παράδειγμα ένα πρόβλημα Classification ή Time-Series Forecasting. Η επιλογή του κατάλληλου μοντέλου είναι δύσκολη όταν υπάρχουν πολλές διαθέσιμες επιλογές. Επομένως, είναι σημαντικό να είναι καλά καθορισμένες οι πληροφορίες που αναμένονται να εξαχθούν από τα σύνολα δεδομένων, όπως επίσης και ο τύπος του μοντέλου που θα ταίριαζε καλύτερα στις ανάγκες αυτού του σκοπού. Τα εργαλεία AutoML καθορίζουν αυτόματα το σωστό μοντέλο.

4. **Μοντελοποίηση (Model Engineering):** Σε αυτό το στάδιο, γίνεται η χρήση όλων των πληροφοριών που καταγράφηκαν στο στάδιο του Planning, ώστε να δημιουργηθεί και να εκπαιδευτεί ένα ή περισσότερα ML μοντέλα. Σε αντίθεση με μια Manual προγραμματιστική προσέγγιση στην οποία θα προχωρούσε ένας ML Expert, τα AutoML Tools περιλαμβάνουν στο Pipeline τους τα επιμέρους βήματα που ακολουθούν:
    a. Δημιουργία της αρχιτεκτονικής του ML μοντέλου έπειτα από εκτενή έρευνα με σκοπό να διασφαλιστεί, όσο το δυνατόν περισσότερο, ότι το μοντέλο που θα προκύψει θα είναι αποτελεσματικό.
    b. Προσδιορισμός των μετρήσεων απόδοσης του ML μοντέλου.
    c. Εκμάθηση και επικύρωση του μοντέλου στο σύνολο των δεδομένων εκμάθησης (Training) και επικύρωσης (Validation).
    d. Βελτιστοποίηση της παραμετροποίησης του ML αλγορίθμου μέσω της διαδικασίας Hyperparameter Tuning.
    e. Παρακολούθηση των πειραμάτων κατά την εκμάθηση, των Metadata που προκύπτουν και των χαρακτηριστικών τους.
    f. Διορθωτικές παρεμβάσεις με αλλαγές στον κώδικα για τις περιπτώσεις της Manual προσέγγισης και των ML Pipelines στις αυτοματοποιημένες προσεγγίσεις.
    g. Εκτέλεση συμπίεσης (Compression) για παραδείγματα λιγότερων ή μικρότερων παραμέτρων.
    h. Συνδυασμό μεθόδων μοντελοποίησης (Ensemble Modeling).
    i. Ερμηνεία των αποτελεσμάτων από εξειδικευμένους επαγγελματίες του εκάστοτε επιχειρηματικού τομέα (Domain Experts).
5. **Βελτιστοποίηση Παραμέτρων (Hyperparameter Tuning):** Ένα AutoML εργαλείο για να λειτουργεί αποτελεσματικά, θα πρέπει στο Pipeline του να μπορεί να βρίσκει τον αποδοτικότερο συνδυασμό παραμετροποίησης του αλγορίθμου μέσω της μεθόδου Hyperparameter Tuning. Ο συνδυασμός που επιλέγεται είναι αυτός όπου μετά από σειρά δοκιμών, στο στάδιο του Training, είναι πιο ακριβής στις προβλέψεις. Συνήθως σε αυτή τη διαδικασία δοκιμάζονται παράμετροι όπως τα βάρη (Weights), ο ρυθμός εκμάθησης (Learning Rate), το μέγιστο βάθος του δέντρου απόφασης (Max Depth of Tree) κ.α.
6. **Αξιολόγηση Μοντέλου (Model Evaluation):** Έπειτα από τη δημιουργία της τελικής έκδοσης του ML μοντέλου, ακολουθεί η αξιολόγησή του βάσει των δεικτών μέτρησης της αποδοτικότητας ώστε να κριθεί εάν είναι κατάλληλο να μπει στην παραγωγή. Σε πρώτη φάση ο έλεγχος βασίζεται στο Test Dataset και με τη συμβολή Domain Experts αναγνωρίζονται τα σφάλματα στην πρόβλεψη. Έπειτα ελέγχεται για την ανθεκτικότητά του σε τυχαία γεγονότα. Πραγματικά ελέγχεται σε νέα δεδομένα που κατά τη διαδικασία εκπαίδευσης ήταν εντελώς άγνωστα για το μοντέλο. Επίσης, σημαντικός είναι ο έλεγχος τήρησης όλων των απαιτήσεων και προδιαγραφών του επιχειρηματικού, ηθικού και νομικού πλαισίου σχετικά με την εφαρμογή τεχνολογιών τεχνητής νοημοσύνης (Artificial Intelligence).

7. **Εφαρμογή/Διανομή Μοντέλου (Model Deployment):** Αποτελεί μια δύσκολη διαδικασία κατά την οποία ένα ML μοντέλο, μετά την εκπαίδευση και τη βελτιστοποίησή του, αποφασίζεται να εφαρμοστεί σε Large-scale συστήματα, τα οποία συνήθως απαιτούν μάλιστα διεργασίες Data Engineering. Γενικά, τα ML μοντέλα μπορούν να εγκατασταθούν στο Cloud και σε Local Servers, σε Web Browsers και ως πακέτα λογισμικού. Το εκάστοτε μοντέλο μπορεί να χρησιμοποιηθεί μέσω API, Web Applications, Plugins ή Dashboards ώστε κάποιος χρήστης να έχει πρόσβαση στις παραγόμενες προβλέψεις. Στο πλαίσιο αυτό, θα πρέπει να εξεταστούν οι δυνατότητες του Hardware. Πρέπει ακόμα να διασφαλιστεί ότι η RAM, ο χώρος αποθήκευσης και η υπολογιστική ισχύς μπορούν να παράγουν γρήγορα αποτελέσματα. Η σχετική απόδοση θα πρέπει να αξιολογείται διαρκώς όσο το μοντέλο χρησιμοποιείται διασφαλίζοντας την αποδοχή του χρήστη. Τέλος, θα πρέπει να διασφαλίζεται ότι οι αλλαγές κατά τη χρήση του είναι εμφανείς και ικανές να επιφέρουν βελτιώσεις βάσει του σκοπού που επιτελεί. Ένα AutoML σύστημα μπορεί να κάνει αυτή τη διαδικασία πιο εύκολη έχοντας τη γνώση να ενσωματώνει μέσω API το παραγόμενο μοντέλο σε διάφορα άλλα συστήματα και Third-party εφαρμογές.
8. **Παρακολούθηση και Συντήρηση (Monitoring and Maintenance):** Μετά τη λειτουργία του μοντέλου, χρειάζεται συνεχής παρακολούθηση και βελτίωση, παρατηρώντας τις μετρήσεις του, την απόδοση τόσο του υλικού (Hardware) όσο και του λογισμικού (Software), καθώς επίσης και την ικανοποίηση των χρηστών. Η παρακολούθηση γίνεται είτε αυτόματα, όπου στέλνονται οι κατάλληλες ειδοποιήσεις για τυχόν μη ομαλά συμβάντα ή την μειωμένη απόδοση του μοντέλου και του συστήματος, είτε μέσω των αρνητικών κριτικών από τους χρήστες. Αφού ληφθεί μια ειδοποίηση χαμηλής απόδοσης, τα ζητήματα που προκύπτουν αξιολογούνται κατάλληλα και στη συνέχεια το μοντέλο πρέπει να εκπαιδευτεί αντίστοιχα με τα νέα δεδομένα ή να γίνουν οι απαραίτητες αλλαγές στην αρχιτεκτονική του. Η διαδικασία που περιεγράφηκε παραπάνω εκτελείται συνέχεια, ενώ σε σπάνιες περιπτώσεις, χρειάζεται η ανανέωση του πλήρους κύκλου ζωής του μοντέλου με σκοπό να βελτιωθεί η επεξεργασία των δεδομένων και οι τεχνικές εκμάθησης του. Τέλος, είναι πιθανή η ανάγκη για εξολοκλήρου ανανέωση του Software και του Hardware που χρησιμοποιείται ή να εισαχθεί ένα νέο Framework για Integration.

4. Υλοποίηση και Οδηγός Χρήσης “Matapae” AutoML Web Application

4.1 Σκοπός της Εφαρμογής

Η αυξανόμενη ανάγκη των οργανισμών και των επιχειρήσεων να αξιοποιούν τα δεδομένα που συλλέγουν για να λαμβάνουν Data-driven αποφάσεις, έχει αυξήσει, τα τελευταία χρόνια, τη ζήτηση για εξειδικευμένα στελέχη με δεξιότητες μηχανικής μάθησης. Σε πολλές περιπτώσεις, είτε η ζήτηση ξεπερνά την προσφορά και οι θέσεις δεν μπορούν να καλυφθούν, είτε οι επιχειρήσεις λόγω περιορισμένων πόρων δεν επιλέγουν να επενδύσουν σε αντίστοιχα τμήματα μηχανικής μάθησης. Έτσι ενισχύεται η ανάγκη παραγωγής ML μοντέλων από μη-εξειδικευμένο προσωπικό μέσω User-friendly διαδικτυακών εφαρμογών που θα καθοδηγούν τον χρήστη να παράγει ML μοντέλα με καθορισμένα βήματα.

Για το σκοπό αυτό, σχεδιάστηκε και αναπτύχθηκε η διαδικτυακή εφαρμογή “Matapae” με λογική αυτοματοποιημένης παραγωγής μοντέλων μηχανικής μάθησης (AutoML Pipeline), χωρίς κώδικα προγραμματισμού. Πρόκειται για μια πλατφόρμα που ο σχεδιασμός της παρέχει στον χρήστη τη δυνατότητα να εισάγει, αποθηκεύει και διερευνά σύνολα δεδομένων, να εκπαιδεύει μοντέλα μηχανικής μάθησης ανάλογα με το είδος του προβλήματος που θέλει να επιλύσει, να τα αποθηκεύει και να τα επεξηγεί, να τα βελτιώνει και εν τέλει να προβλέπει πάνω σε νέα δεδομένα.

4.2 RStudio IDE, GitHub και Shiny Web App Framework

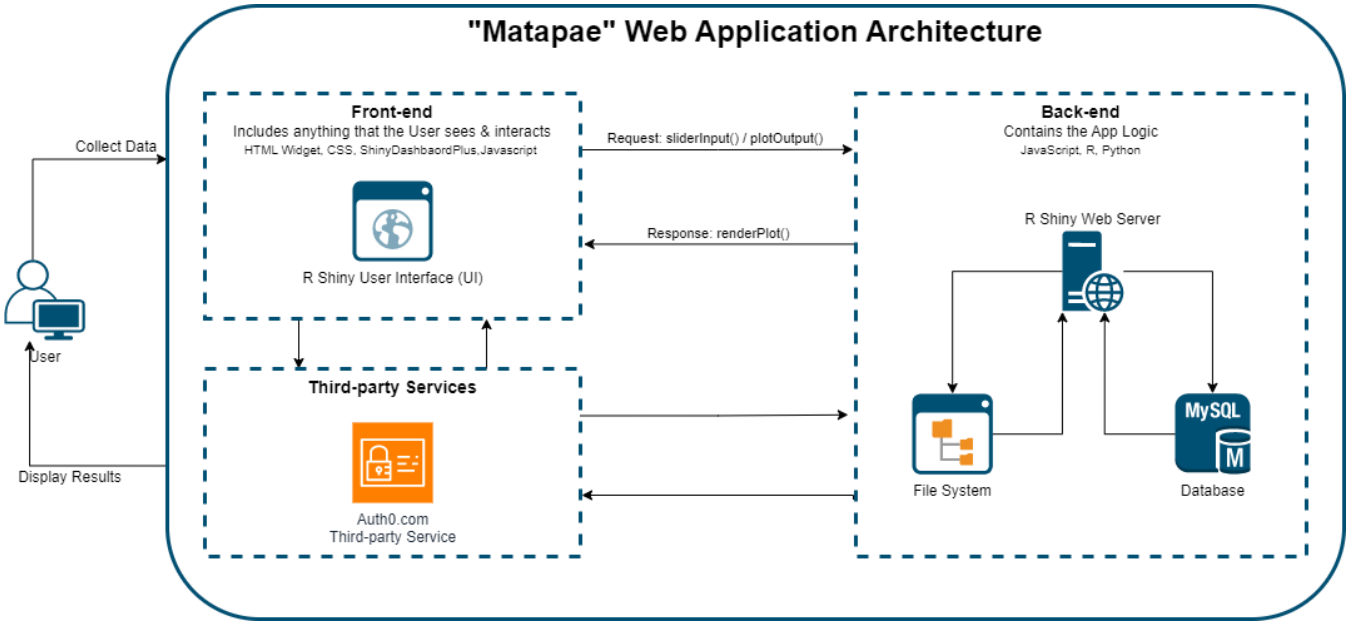
Η ανάπτυξη της εφαρμογής έγινε με τη γλώσσα προγραμματισμού R στην Open-source έκδοση του ολοκληρωμένου περιβάλλοντος ανάπτυξης (Integrated Development Environment) για R και Python, RStudio[1]. Για την αποθήκευση και διαχείριση του πηγαίου κώδικα, όπως επίσης την παρακολούθηση και τον έλεγχο των αλλαγών της εφαρμογής “Matapae”, χρησιμοποιήθηκε το Cloud-based εργαλείο GitHub[2] το όποιο διασυνδέθηκε με το RStudio.

Ένα πλήρως διαδραστικό περιβάλλον χρήσης, βασισμένο στην επικοινωνία μεταξύ User Interface (UI) και Server, εξασφαλίστηκε με τη χρήση του Shiny[3] Web Framework. Το εν λόγω Framework, ακολουθώντας την αρχιτεκτονική των διαδικτυακών εφαρμογών ιστού, δίνει τη δυνατότητα να φιλοξενηθούν αυτόνομες εφαρμογές, που περιλαμβάνουν διαδραστικά εργαλεία ανάλυσης και γραφήματα, σε μια ιστοσελίδα, κατάλληλα δομημένα σε Dashboards. Επιπλέον, ενώ η ανάπτυξη της

[1] https://posit.co/products/open-source/rstudio/
[2] https://github.com/
[3] https://shiny.rstudio.com/

εφαρμογής με χρήση R γίνεται σε επίπεδο UI-Server, το Shiny Package μετατρέπει τον κώδικα σε HTML Widgets, CSS Themes και JavaScript ενέργειες που απαιτούνται για την κατάλληλη εμφάνιση της εφαρμογής σε Web περιβάλλον [37]. Το Shiny Dashboard για την εφαρμογή “Matapae” υλοποιήθηκε με το πακέτο {ShinyDashboardPlus[4]}.

Τα βασικά χαρακτηριστικά της αρχιτεκτονικής δομής της εφαρμογής απεικονίζεται στο διάγραμμα της Εικόνα 5:



Εικόνα 5: Διάγραμμα της R Shiny Αρχιτεκτονικής για την εφαρμογή "Matapae"

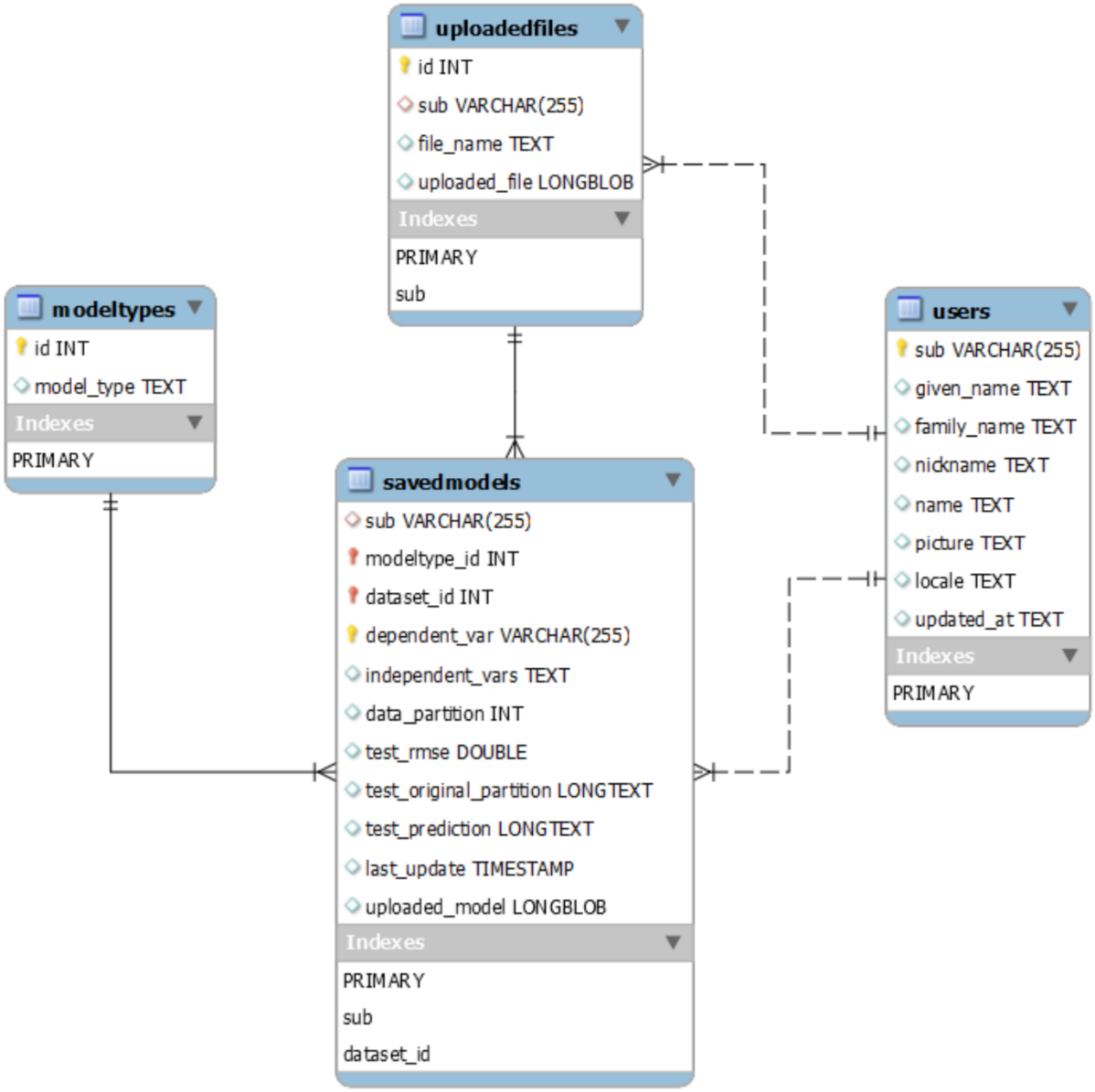
4.3 Σχήμα Βάσης Δεδομένων – Oracle MySQL Database Server

Η αποθήκευση και διαχείριση των συνόλων δεδομένων (Datasets) και των παραγόμενων μοντέλων μηχανικής μάθησης (ML Models) υλοποιήθηκε με σχεσιακή βάση δεδομένων MySQL. Κατά την πρώτη είσοδο του χρήστη (User) στην εφαρμογή, δημιουργείται μια νέα εγγραφή χρήστη στο πίνακα της βάσης δεδομένων “users”. Τα Datasets αντιστοιχίζονται ανά χρήστη και αποθηκεύονται στον πίνακα “uploadedfiles”. Η σχέση των Users με τα Uploaded Files είναι 1:N. Τα ML Models αποθηκεύονται ανά χρήστη στον πίνακα “savedmodels”. Παράλληλα, το κάθε μοντέλο αντιστοιχίζεται με ένα μόνο Dataset και έναν χαρακτηρισμό του τύπου μοντέλου ML (π.χ. Regression, Classification). Η σχέση μεταξύ των Users και των Uploaded Files με τα ML Models είναι 1:N. Επίσης 1:N είναι η σχέση μεταξύ των πινάκων “modeltypes” με τον “savedmodels”. Ιδιαίτερο χαρακτηριστικό του πίνακα “savedmodels” αποτελεί ο συνδυασμός των τριών επιμέρους πεδίων: modeltype_id, dataset_id και dependent_var, ως πρωτεύον

[4] https://rinterface.github.io/shinydashboardPlus/articles/shinydashboardPlus.html

κλειδί του πίνακα (Primary Key). Η αλληλεπίδραση μέσω της γλώσσας προγραμματισμού R με το σύστημα διαχείρισης βάσεων δεδομένων (DBMS) έγινε με το πακέτο {DBI[5]}.

Στο διάγραμμα της Εικόνα 6, απεικονίζονται αναλυτικά οι τέσσερις πίνακες του σχήματος της βάσης δεδομένων όπως διαμορφώνονται με τα πεδία (Fields), τα Primary και Foreign Keys και τις σχέσεις που αναπτύσσονται μεταξύ τους:



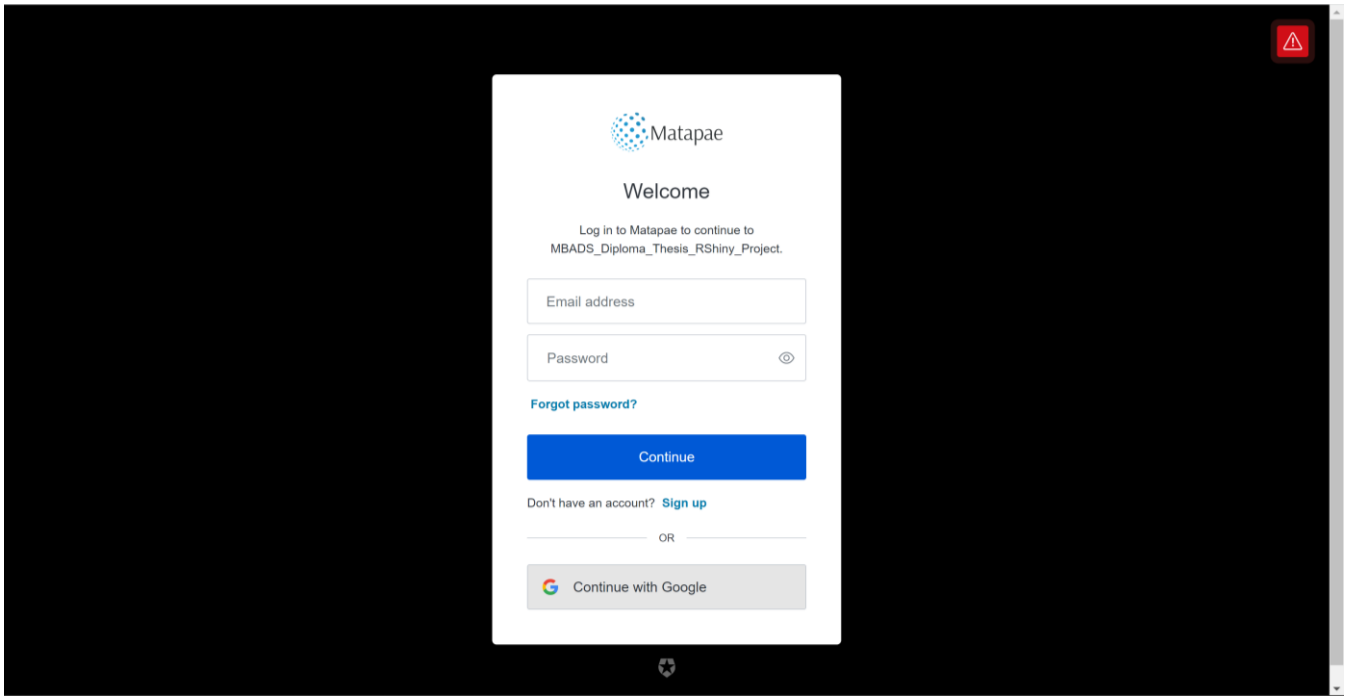
Εικόνα 6: Διάγραμμα Σχήματος Βάσης Δεδομένων από το MySQL Workbench Database Management Tool

[5] https://dbi.r-dbi.org/

4.4 Σύστημα Επαλήθευσης Ταυτότητας και Εξουσιοδότησης Χρηστών (Auth0)

Η υλοποίηση των διαδικασιών επαλήθευσης ταυτότητας (Authentication) και εξουσιοδότησης (Authorization) στις υπηρεσίες της εφαρμογής "Matapae" έγινε μέσω Auth0[6], ένα Third-party API που προσφέρει τεχνολογίες διαχείρισης της ταυτότητας και πρόσβασης των χρηστών (Identity and Access Management – IAM[7]). Πρόκειται για τεχνολογίες που διασφαλίζουν ότι τα κατάλληλα άτομα έχουν πρόσβαση στους κατάλληλους ψηφιακούς πόρους και υπηρεσίες της εφαρμογής, τη σωστή στιγμή και για τους σωστούς λόγους.

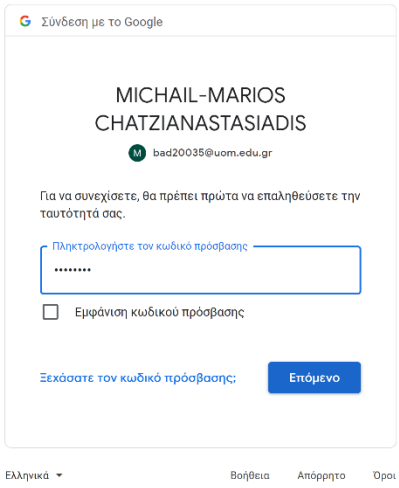
Κατά την είσοδο στην εφαρμογή, ζητείται από τον χρήστη να κάνει εγγραφή (Sign up), εφόσον συνδέεται για πρώτη φορά, ή σύνδεση (Log in) στις περιπτώσεις που είναι ήδη εγγεγραμμένος, όπως φαίνεται στην Εικόνα 7. Η αυθεντικοποίηση των χρηστών πραγματοποιείται με Google Account ή οποιοδήποτε άλλο email. Στις περιπτώσεις που η εγγραφή ολοκληρώνεται με Google Account, η είσοδος στην εφαρμογή πραγματοποιείται με το αντίστοιχο Google Password, ενώ στην άλλη περίπτωση με το Password που έχει ορίσει ο χρήστης κατά την εγγραφή του στην εφαρμογή "Matapae".



Εικόνα 7: Αρχική Σελίδα Sign up/Sign in

[6] https://curso-r.github.io/auth0/
[7] https://auth0.com/docs/get-started/identity-fundamentals/identity-and-access-management

Στην περίπτωση της επιλογής Google Account, ο χρήστης επιλέγει τον αντίστοιχο λογαριασμό του και στη συνέχεια πρέπει να πληκτρολογήσει τον κωδικό του. Πατώντας το κουμπί **Επόμενο** (Εικόνα 8), συνδέεται στην εφαρμογή επιτυχώς.



Εικόνα 8: Εισαγωγή του Κωδικού Πρόσβασης για Σύνδεση μέσω Google Account

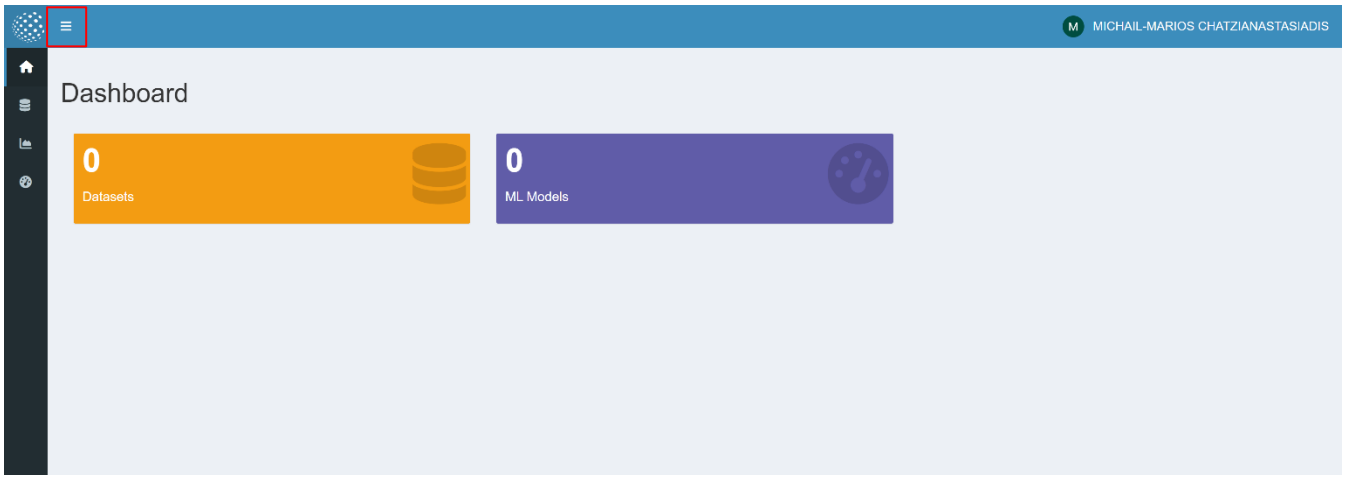
4.4.1 R Κώδικας για την Εγγραφή Χρήστη στη Βάση Δεδομένων

Καθώς ο χρήστης συνδέεται για πρώτη φορά στην εφαρμογή καλείται η συνάρτηση saveUsertoDB(…) ώστε να προστεθεί η νέα εγγραφή χρήστη στη βάση δεδομένων χρησιμοποιώντας τον παρακάτω κώδικα:

```
saveUsertoDB <- function(user) {

 print ("**************saveUsertoDB***************")

 dbcon <- RMySQL::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password)

 user_tibble <- as_tibble(user)

 RMySQL::dbWriteTable(dbcon, "users", user_tibble, append= TRUE,
                      row.names=FALSE)

 RMySQL::dbDisconnect(dbcon)

 }
```

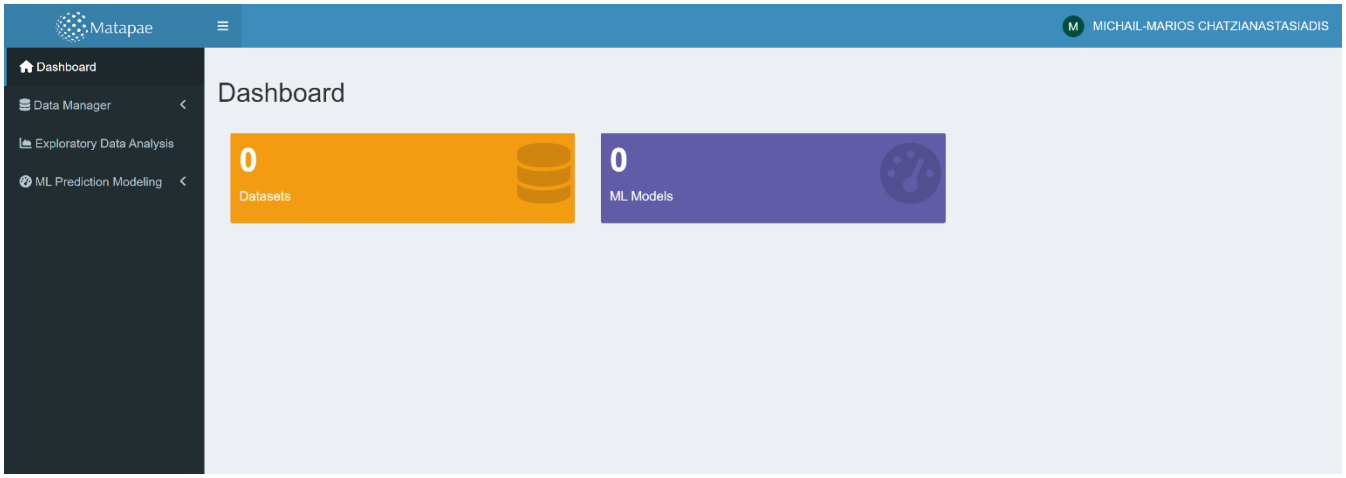
4.5 Dashboard

Αμέσως μετά την επιτυχημένη διαδικασία αυθεντικοποίησης, πραγματοποιείται η είσοδος του χρήστη στην εφαρμογή, όπου μεταφέρεται στην αρχική σελίδα **Dashboard**, στην οποία εμφανίζονται συνοπτικές πληροφορίες μέσω της συνάρτησης valueBox(…) του πακέτου {ShinyDashboard[8]}. Ειδικότερα, στο πορτοκαλί πλαίσιο εμφανίζεται το σύνολο των **Datasets** που έχουν αποθηκευτεί από τον εξουσιοδοτημένο χρήστη, ενώ στο μωβ πλαίσιο το σύνολο των **ML Models** αντίστοιχα (Εικόνα 9).



Εικόνα 9: Είσοδος στην Αρχική Οθόνη της Εφαρμογής (Dashboard)

Καθώς φορτώνει η εφαρμογή, το μενού που βρίσκεται στα αριστερά της οθόνης είναι συμπτυγμένο και ο χρήστης μπορεί να το "ανοίξει" πατώντας το κουμπί με τις **τρεις γραμμές**, το οποίο βρίσκεται δίπλα από το λογότυπο (Εικόνα 9). Με τον τρόπο αυτό, διακρίνονται εύκολα οι διαθέσιμες επιλογές του μενού, όπως απεικονίζεται παρακάτω στην Εικόνα 10:

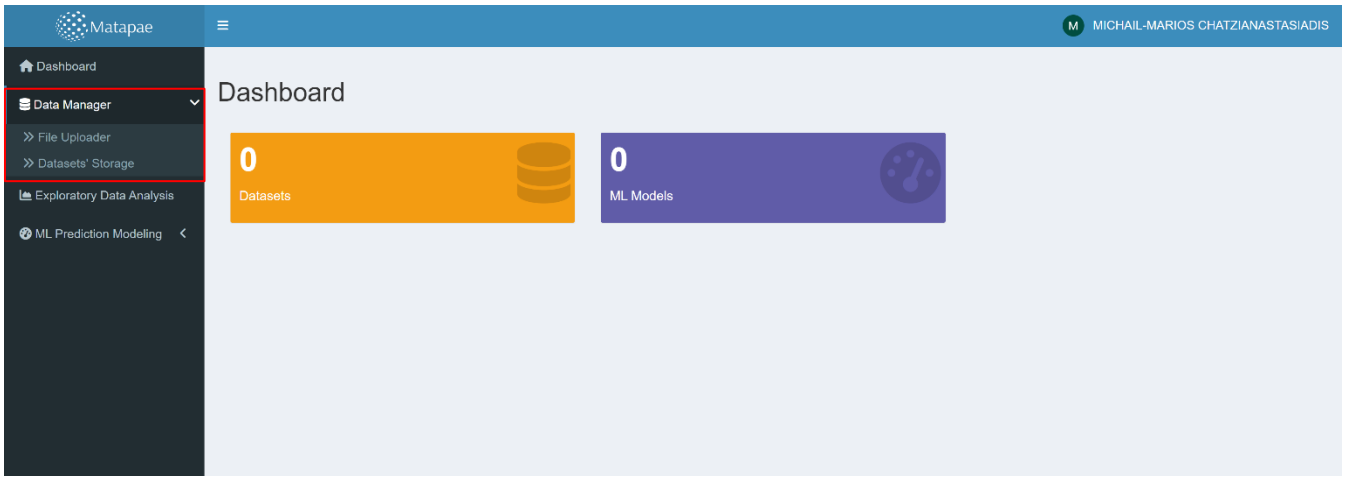


Εικόνα 10: Αρχική Οθόνη της Εφαρμογής (Dashboard) με Ανεπτυγμένο Menu Sidebar

[8] https://rstudio.github.io/shinydashboard/

4.6 Data Manager

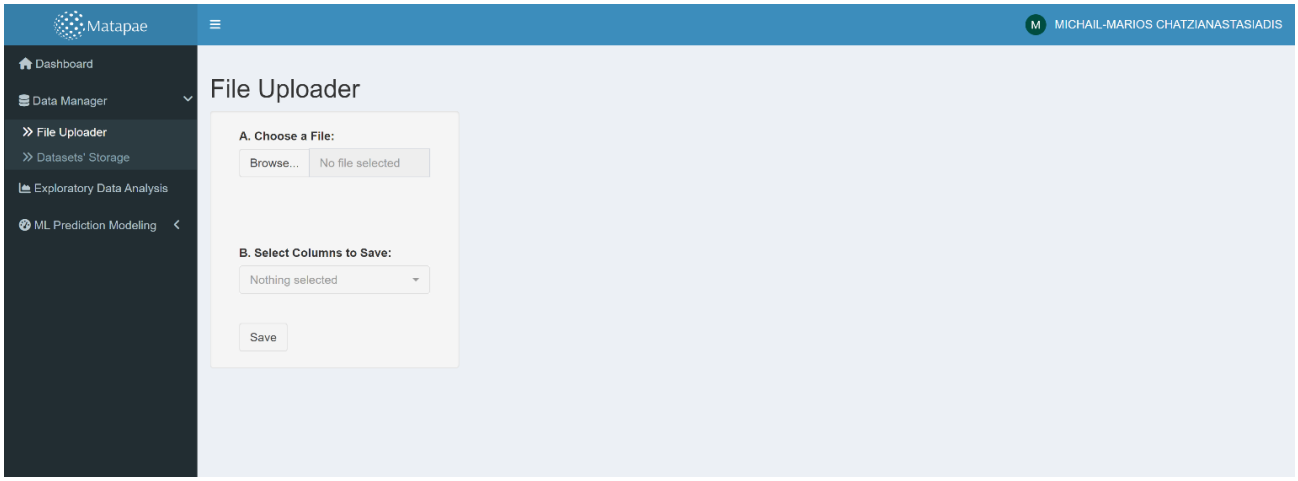
Στην ενότητα του μενού **Data Manager**, υπάρχουν διαθέσιμες οι επιλογές **File Uploader** και **Datasets' Storage** (Εικόνα 11).



Εικόνα 11: Επιλογές Sidebar Menu για την Ενότητα Data Manager

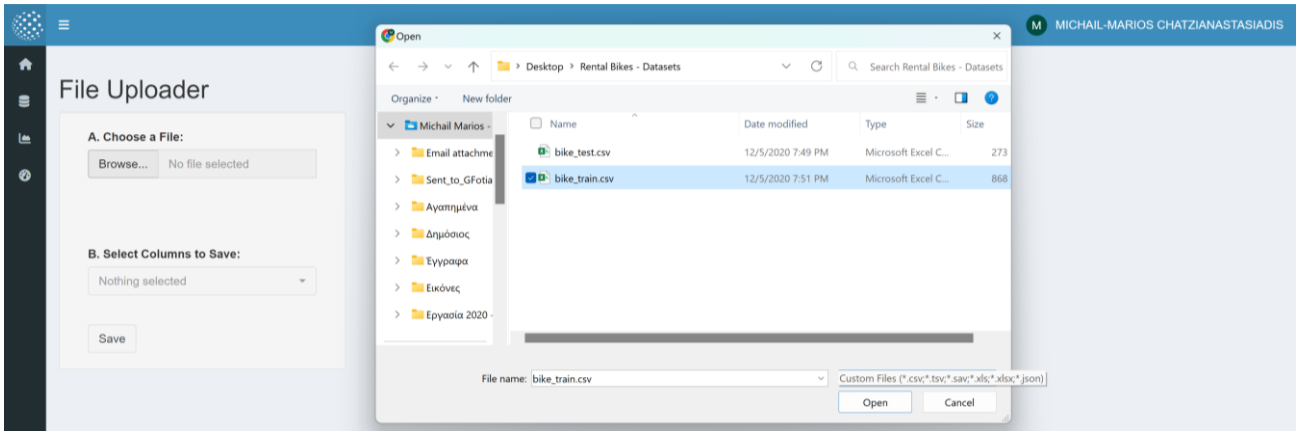
4.6.1 File Uploader

Με την επιλογή του εργαλείου **File Uploader** (Εικόνα 12), δίνεται η δυνατότητα στον χρήστη να εισάγει στην εφαρμογή ένα αρχείο δεδομένων, να δει το περιεχόμενο του και στη συνέχεια να επιλέξει τα πεδία του αρχείου που επιθυμεί να αποθηκευτούν στον «προσωπικό» του χώρο αποθήκευσης. Τα Datasets της βάσης δεδομένων που έχουν αποθηκευτεί μέσω του **File Uploader**, είναι αυτά που θα χρησιμοποιηθούν αργότερα για τις αναλύσεις, τη δημιουργία των μοντέλων μηχανικής μάθησης και τις προβλέψεις.



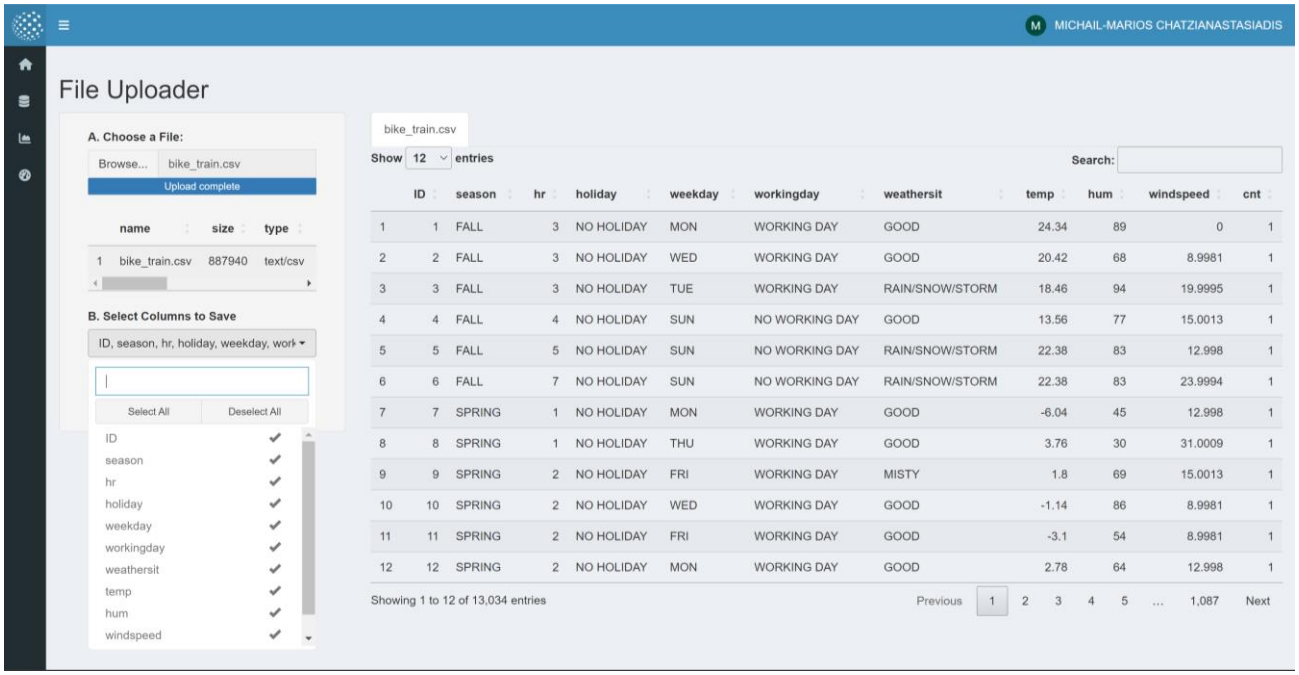
Εικόνα 12: Περιβάλλον File Uploader

«Πατώντας» το κουμπί **Browse**, ανοίγει ένα νέο «παράθυρο» μέσω του οποίου δίνεται πρόσβαση στα αποθηκευμένα αρχεία που βρίσκονται «τοπικά» στον υπολογιστή του χρήστη. Μεταβαίνοντας στον επιθυμητό φάκελο, γίνεται η επιλογή του αρχείου από τη λίστα, η οποία περιλαμβάνει μόνο τους αποδεκτούς τύπους αρχείων (.csv, .tsv, .sav, .xls, .xlsx ή .json). Με το κουμπί **Open**, το αρχείο «ανεβαίνει» στην εφαρμογή (Εικόνα 13).



Εικόνα 13: Παράθυρο Επιλογής Αρχείου για Άνοιγμα στην Εφαρμογή

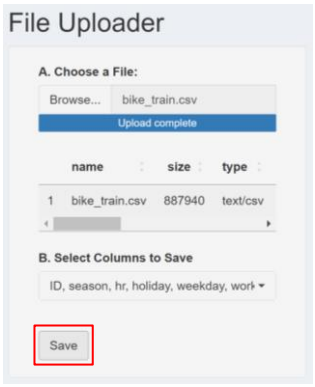
Μόλις το αρχείο «φορτώσει» στο περιβάλλον της εφαρμογής, ο χρήστης έχει τη δυνατότητα να επιλέξει, μέσω Dropdown Menu, τα πεδία του αρχείου που επιθυμεί να δει, σε στήλες, στην προεπισκόπηση περιεχομένου (Εικόνα 14).



	ID	season	hr	holiday	weekday	workingday	weathersit	temp	hum	windspeed	cnt
1	1	FALL	3	NO HOLIDAY	MON	WORKING DAY	GOOD	24.34	89	0	1
2	2	FALL	3	NO HOLIDAY	WED	WORKING DAY	GOOD	20.42	68	8.9981	1
3	3	FALL	3	NO HOLIDAY	TUE	WORKING DAY	RAIN/SNOW/STORM	18.46	94	19.9995	1
4	4	FALL	4	NO HOLIDAY	SUN	NO WORKING DAY	GOOD	13.56	77	15.0013	1
5	5	FALL	5	NO HOLIDAY	SUN	NO WORKING DAY	RAIN/SNOW/STORM	22.38	83	12.998	1
6	6	FALL	7	NO HOLIDAY	SUN	NO WORKING DAY	RAIN/SNOW/STORM	22.38	83	23.9994	1
7	7	SPRING	1	NO HOLIDAY	MON	WORKING DAY	GOOD	-6.04	45	12.998	1
8	8	SPRING	1	NO HOLIDAY	THU	WORKING DAY	GOOD	3.76	30	31.0009	1
9	9	SPRING	2	NO HOLIDAY	FRI	WORKING DAY	MISTY	1.8	69	15.0013	1
10	10	SPRING	2	NO HOLIDAY	WED	WORKING DAY	GOOD	-1.14	86	8.9981	1
11	11	SPRING	2	NO HOLIDAY	FRI	WORKING DAY	GOOD	-3.1	54	8.9981	1
12	12	SPRING	2	NO HOLIDAY	MON	WORKING DAY	GOOD	2.78	64	12.998	1

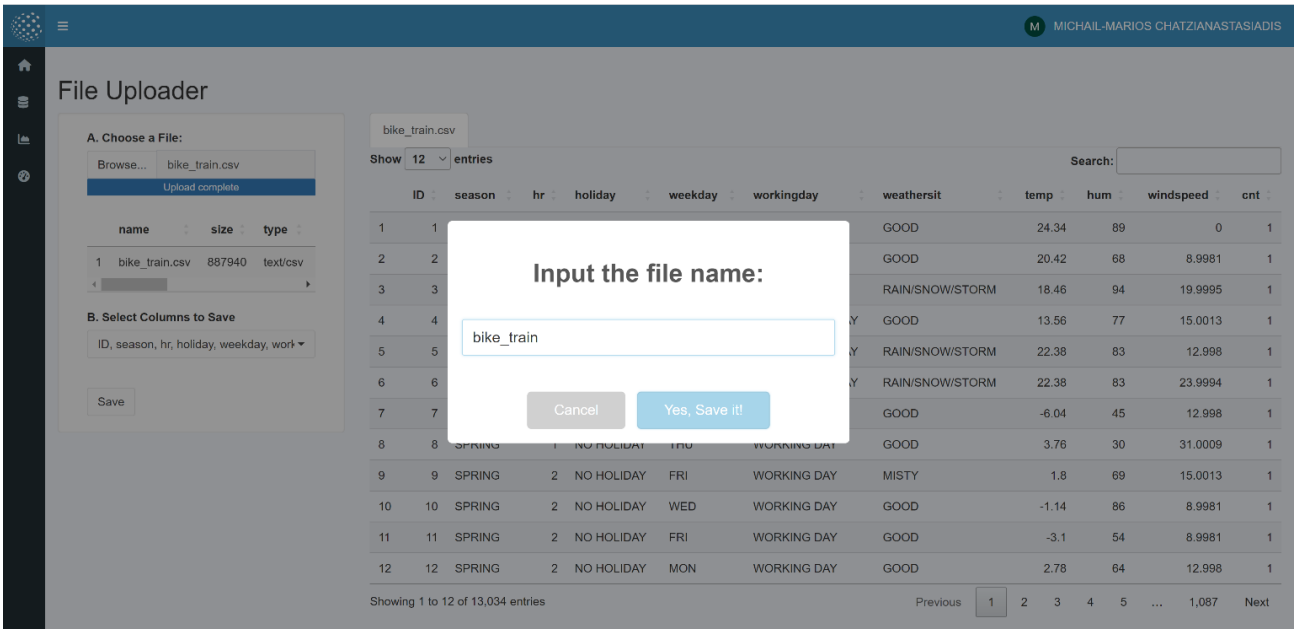
Εικόνα 14: Επιλογή Στηλών Αρχείου προς Εμφάνιση και Αποθήκευση

Με το κουμπί **Save** (Εικόνα 15), τα επιλεγμένα πεδία του αρχείου αποθηκεύονται ως Dataset στη βάση δεδομένων της εφαρμογής.



Εικόνα 15: Κουμπί Αποθήκευσης Dataset

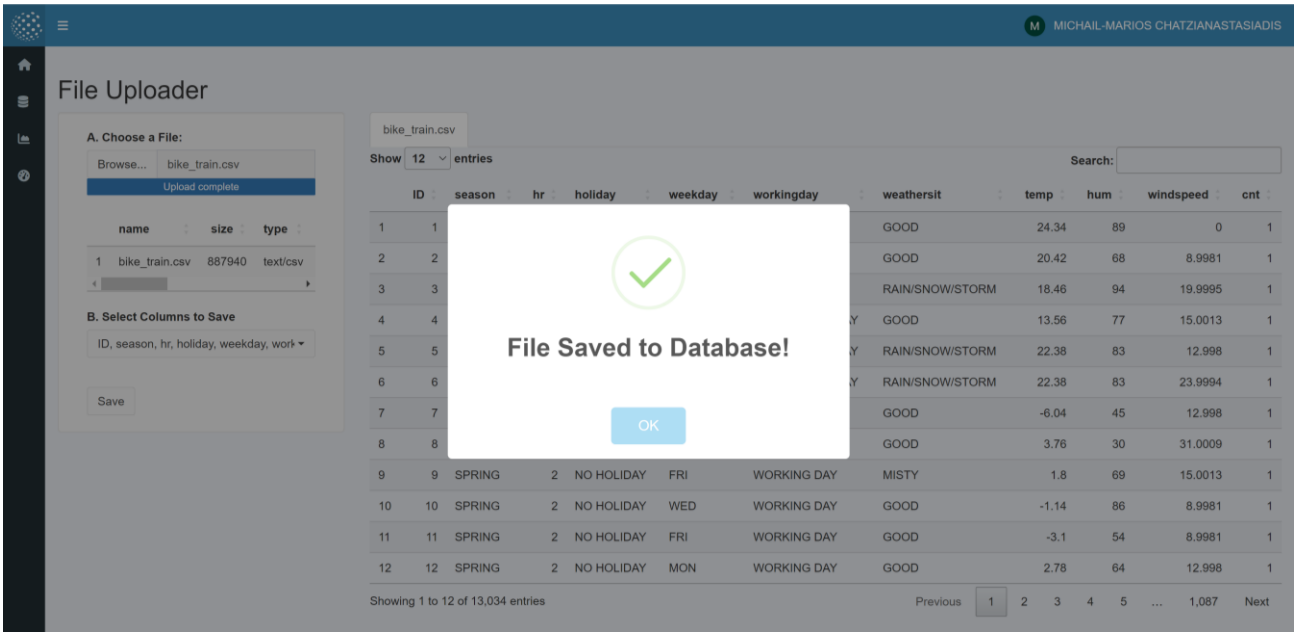
Μόλις ο χρήστης «πατήσει» το κουμπί **Save**, τότε μέσω Input ShinyAlert[9], θα ζητηθεί η εισαγωγή του ονόματος για την αποθήκευση του Dataset στη βάση δεδομένων, σύμφωνα με την Εικόνα 16.



Εικόνα 16: Input ShinyAlert - Εισαγωγή Ονόματος Αρχείου

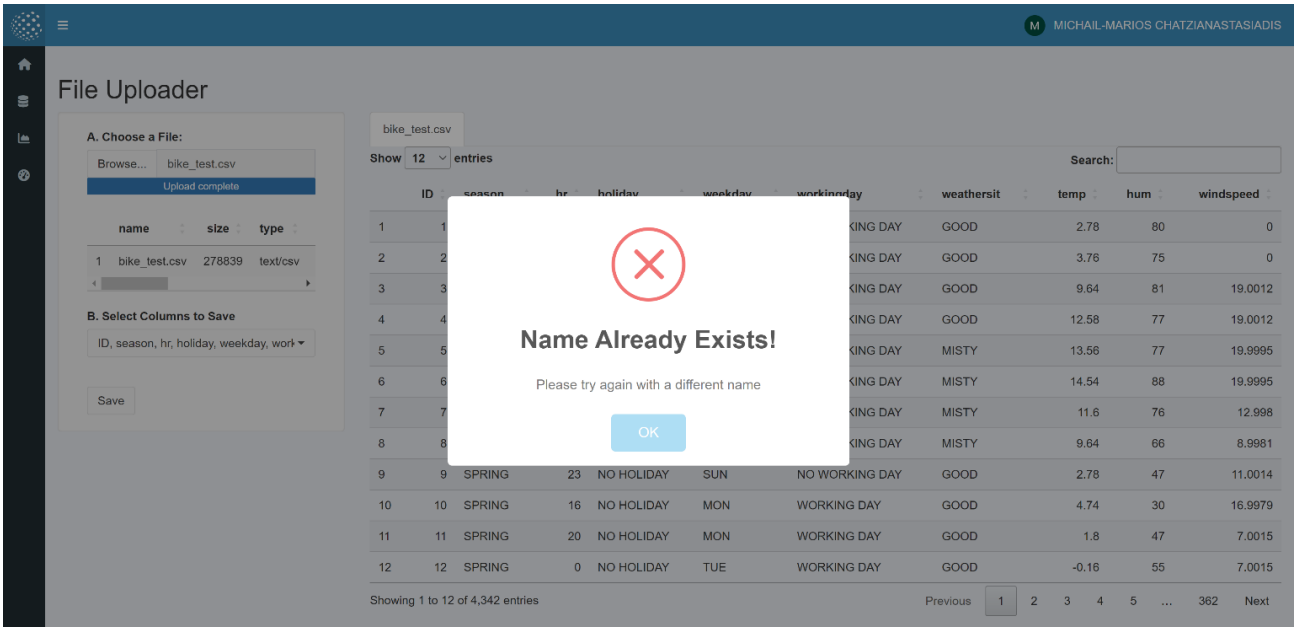
[9] https://github.com/daattali/shinyalert

Η ολοκλήρωση της διαδικασίας αποθήκευσης γίνεται με το «πάτημα» του κουμπιού **"Yes, Save it!"**, όπου εφόσον το όνομα του αρχείου είναι έγκυρο, η αποθήκευση ολοκληρώνεται επιτυχώς με την εμφάνιση του Success ShinyAlert (Εικόνα 17).



Εικόνα 17: Success ShinyAlert - Επιτυχής Αποθήκευση με Έγκυρο Όνομα Αρχείου

Σε αντίθετη περίπτωση, δηλαδή όταν υπάρχει ήδη αποθηκευμένο αρχείο με το ίδιο όνομα, θα εμφανιστεί Error ShinyAlert, που θα ενημερώνει τον χρήστη να δοκιμάσει ένα διαφορετικό όνομα για το Dataset που θέλει να αποθηκεύσει (Εικόνα 18). Η νέα προσπάθεια εισαγωγής έγκυρου ονόματος, θα γίνει μέσω ενός Input ShinyAlert (όπως της Εικόνα 16), που θα εμφανιστεί με το «πάτημα» του **OK**.



Εικόνα 18: Error ShinyAlert - Το Όνομα του Αρχείου Υπάρχει στη Βάση Δεδομένων

4.6.1.1 R Κώδικας για τον Έλεγχο Εγκυρότητας του Ονόματος Αποθήκευσης

Κατά το «πάτημα» του **"Yes, Save it!"** (Εικόνα 17), εκτελείται η συνάρτηση ελέγχου εγκυρότητας του ονόματος. Ειδικότερα, ελέγχεται αν υπάρχει ήδη αποθηκευμένο Dataset με το ίδιο όνομα:

```
fileNameExistsinDB <- function (sub, file_name) {

 dbcon <- DBI::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password
 )

 request0 <- DBI::dbGetQuery(dbcon, "show tables")
 print(request0)

 if (!("uploadedfiles" %in% request0$Tables_in_mbads_thesis_db)) {

  return(TRUE)
  DBI::dbDisconnect(dbcon)

 }

 else{

  file_name <- paste0(file_name, ".csv")
  print(file_name)

  query <-
   paste0(
    "SELECT COUNT(*) FROM uploadedfiles WHERE file_name = '",
    file_name,
    "' AND sub = '",
    sub,
    "';"
   )

  print(query)

  request <- DBI::dbGetQuery(dbcon, query)
  print(request)


  if (request == 0) {
   DBI::dbDisconnect(dbcon)
   return(TRUE)


  } else {
   DBI::dbDisconnect(dbcon)
   return(FALSE)

  }

 }

}
```

4.6.1.2 R Κώδικας για την Αποθήκευση του Dataset στη Βάση Δεδομένων

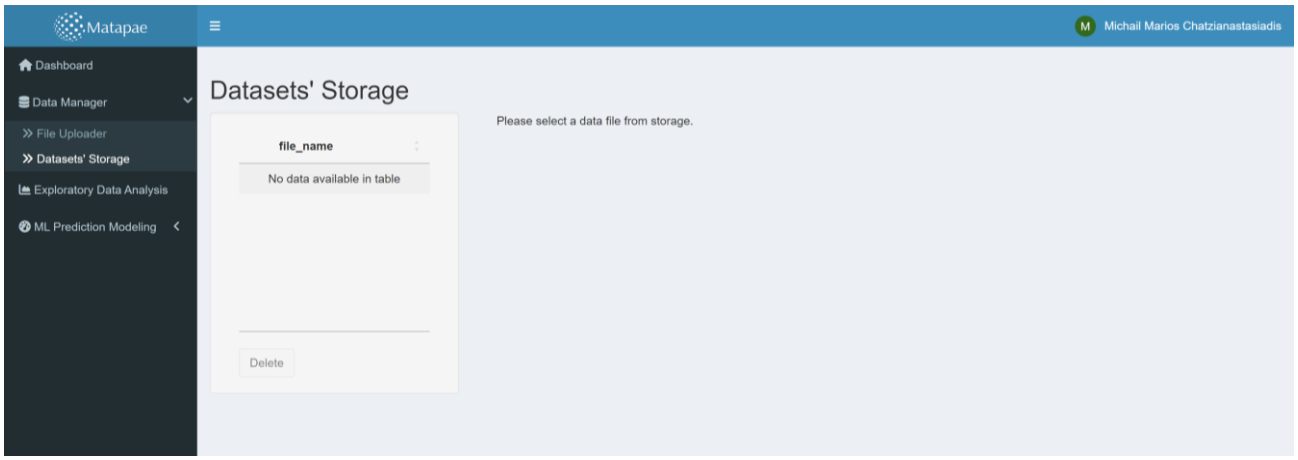
Στις περιπτώσεις που ο έλεγχος εγκυρότητας του ονόματος είναι επιτυχής (4.6.1.1), με την εκτέλεση της συνάρτησης που ακολουθεί, ολοκληρώνεται η διαδικασία αποθήκευσης του Dataset στη βάση δεδομένων της εφαρμογής:

```
uploadedFiletoDB <- function (sub, file_name) {

 print ("**************uploadedFiletoDB***************")

 dbcon <- RMySQL::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password
 )

 file_name <- paste0(file_name,".csv")
 print(file_name)

 insert_data = data.frame(sub, file_name)
 print(insert_data)

 RMySQL::dbWriteTable(
                dbcon,
                "uploadedfiles",
                insert_data,
                field.types = c( id = "INTEGER AUTO_INCREMENT PRIMARY KEY",
                 sub = "VARCHAR(255)",
                 file_name = "TEXT",
                 uploaded_file = "LONGBLOB"),
         append = TRUE,
                row.names = FALSE)

 uploaded_file_query    <-    paste0("C:\\ProgramData\\MySQL\\MySQL    Server
8.0\\Uploads\\", file_name)

 query <- dbplyr::build_sql(
                      "UPDATE uploadedfiles SET uploaded_file = LOAD_FILE(",
                      uploaded_file_query,
                      ") WHERE sub = ",
                      sub,
                      " AND file_name = ",
                      file_name,
                      ";",
                      con = dbcon)

 rs <- DBI::dbSendQuery(dbcon, query)

 DBI::dbClearResult(rs)

 RMySQL::dbDisconnect(dbcon)

}
```

4.6.2 Datasets' Storage

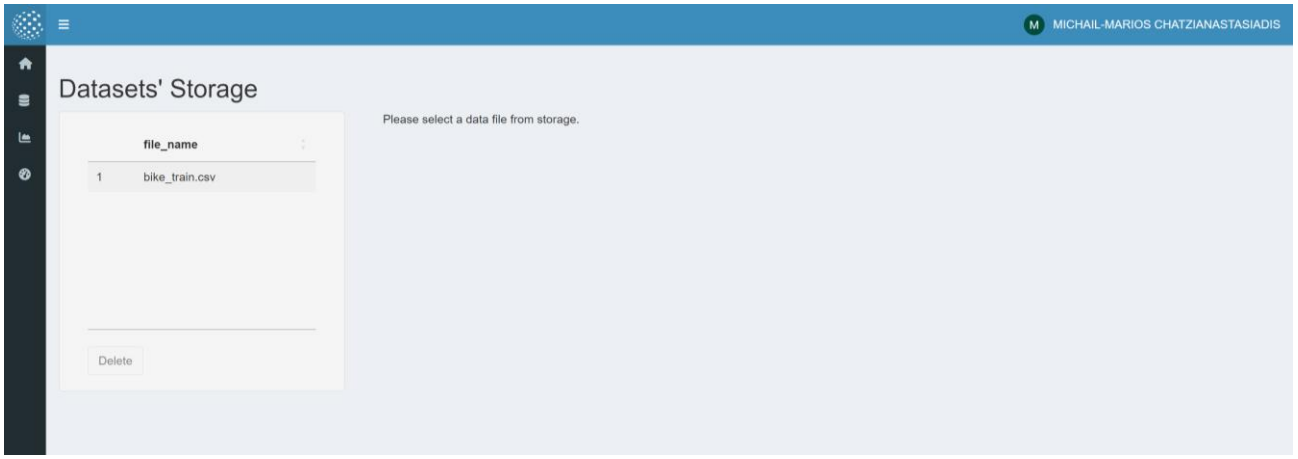
Η δεύτερη επιλογή της ενότητας **Data Manager** που ονομάζεται **Datasets' Storage**, δίνει στον χρήστη τη δυνατότητα πρόσβασης στη λίστα των Datasets που έχει ήδη αποθηκεύσει στον «προσωπικό» του χώρο αποθήκευσης. Επίσης, μπορεί να δει το περιεχόμενο τους ή να διαγράψει οποιοδήποτε Dataset δεν του είναι πλέον χρήσιμο.

Όταν ο «προσωπικός» χώρος αποθήκευσης του χρήστη είναι «άδειος», δηλαδή δεν υπάρχει κάποιο αποθηκευμένο Dataset, τότε εμφανίζεται το ενημερωτικό μήνυμα "**No data available in table**" (Εικόνα 19).



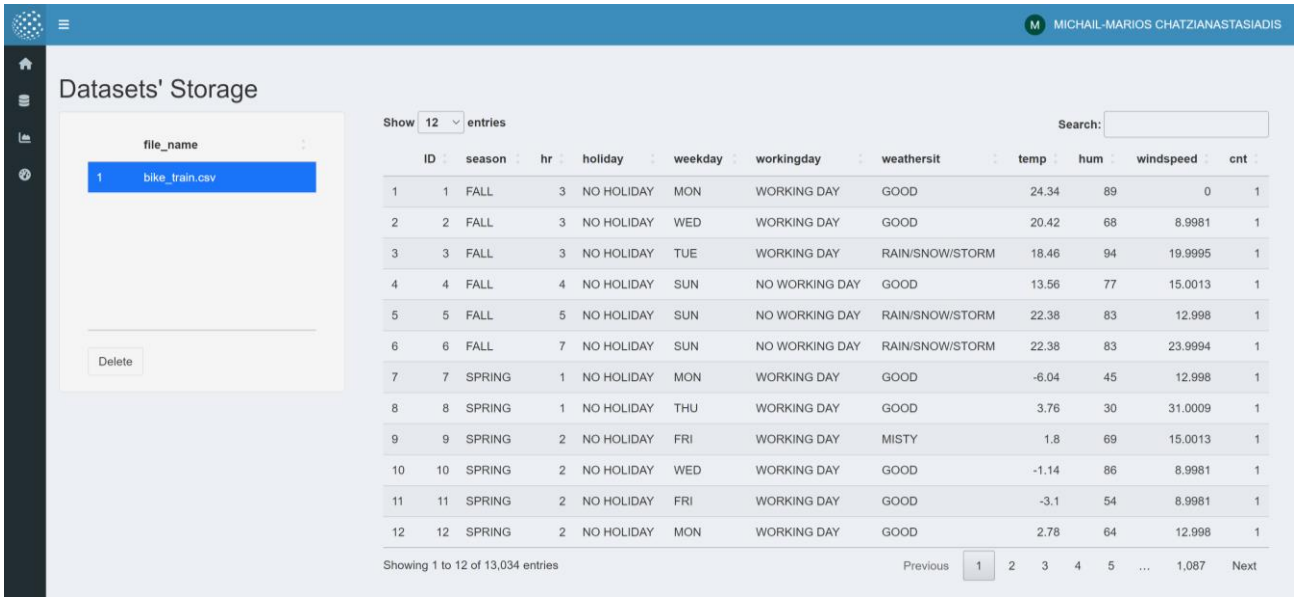
Εικόνα 19: Κενή Λίστα Αποθηκευμένων Datasets

Η λίστα των αποθηκευμένων Datasets του **Datasets' Storage** ενημερώνεται αυτόματα κάθε φορά που ο χρήστης αποθηκεύει επιτυχώς κάποιο νέο Dataset στη βάση δεδομένων της εφαρμογής, με τη χρήση του εργαλείου **File Uploader**, ή όταν διαγράφει ένα ήδη αποθηκευμένο Dataset (Εικόνα 20).



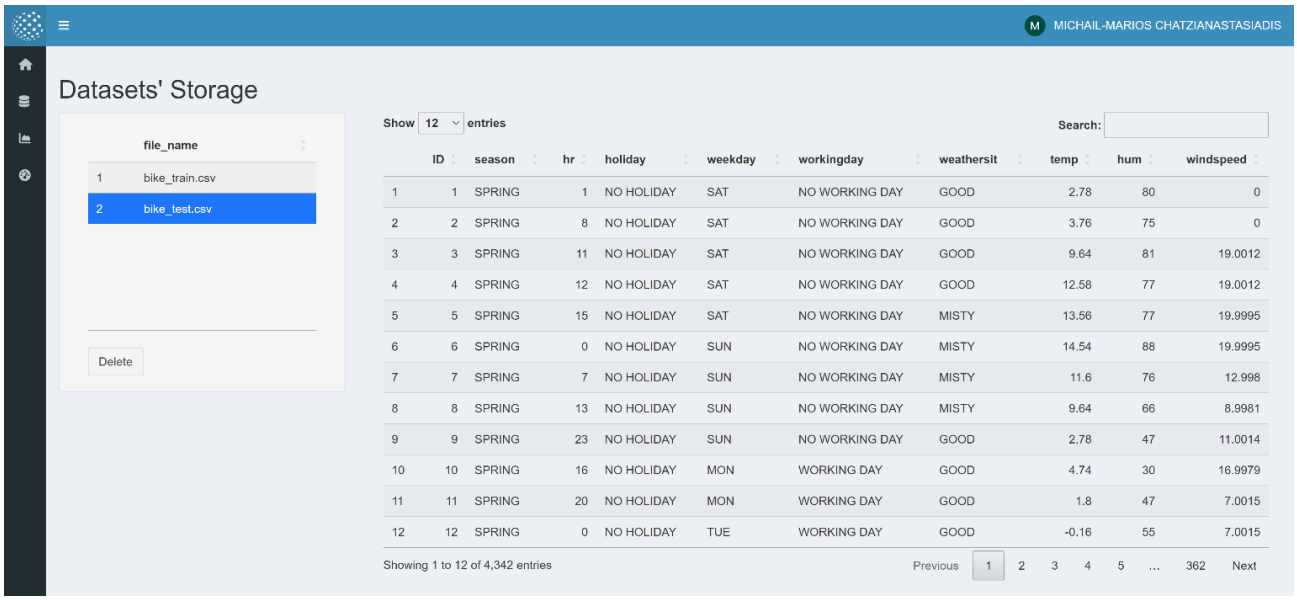
Εικόνα 20: Λίστα με Ένα Αποθηκευμένο Dataset ("bike_train.csv")

Σύμφωνα με το παράδειγμα της Εικόνα 21, κατά την επιλογή του αποθηκευμένου Dataset με το όνομα “bike_train.csv” από τη λίστα των αποθηκευμένων Datasets του Sidebar, στα αριστερά της εφαρμογής, ο χρήστης βλέπει το περιεχόμενο του, στην προεπισκόπηση που είναι διαθέσιμη στο δεξί τμήμα της εφαρμογής.



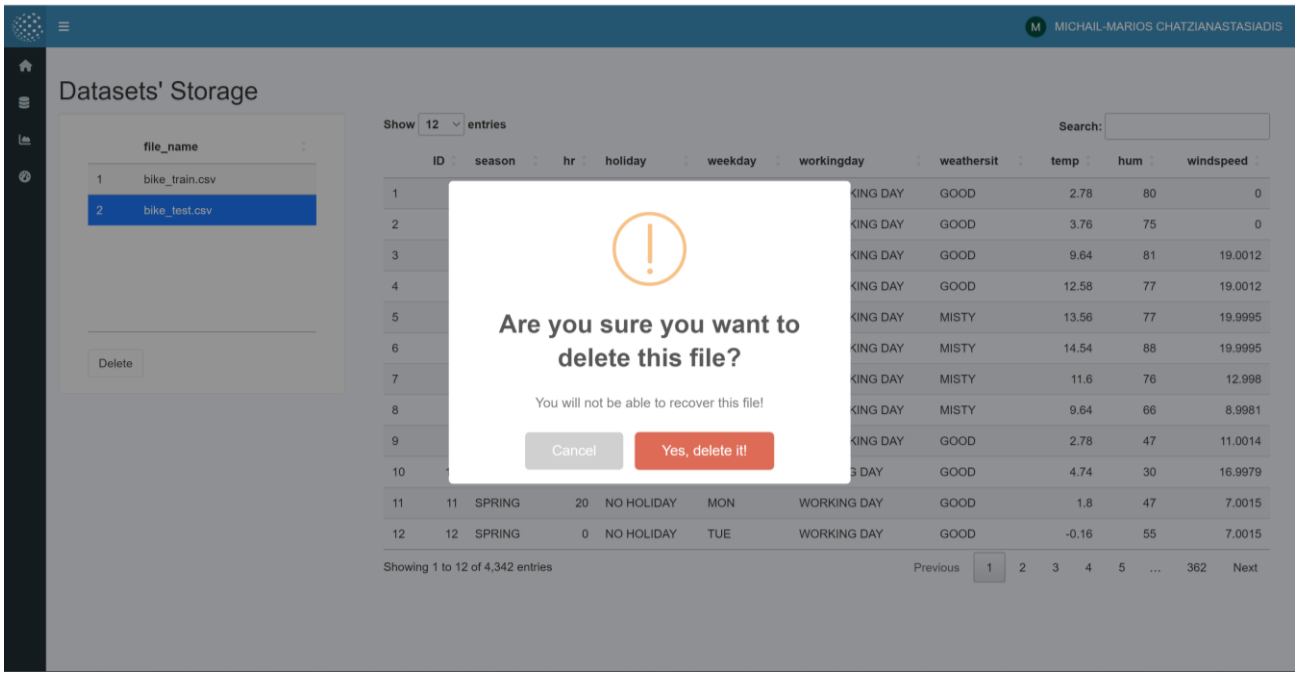
Εικόνα 21: Προεπισκόπηση Αποθηκευμένου Dataset ("bike_train.csv")

Μετά από την εισαγωγή κάθε Dataset στη βάση δεδομένων, η λίστα ανανεώνεται και υπάρχει η δυνατότητα της προεπισκόπησης του, όπως φαίνεται στην Εικόνα 22. Επισημαίνεται ότι, ο χρήστης μπορεί να μεταβεί από την προεπισκόπηση του επιλεγμένου Dataset σε ενός άλλου, επιλέγοντας το από τη λίστα. Η προεπισκόπηση κάθε Dataset μπορεί να «κλείσει», εφόσον ο χρήστης «πατήσει» επάνω στο όνομα του επιλεγμένου από τη λίστα Dataset.



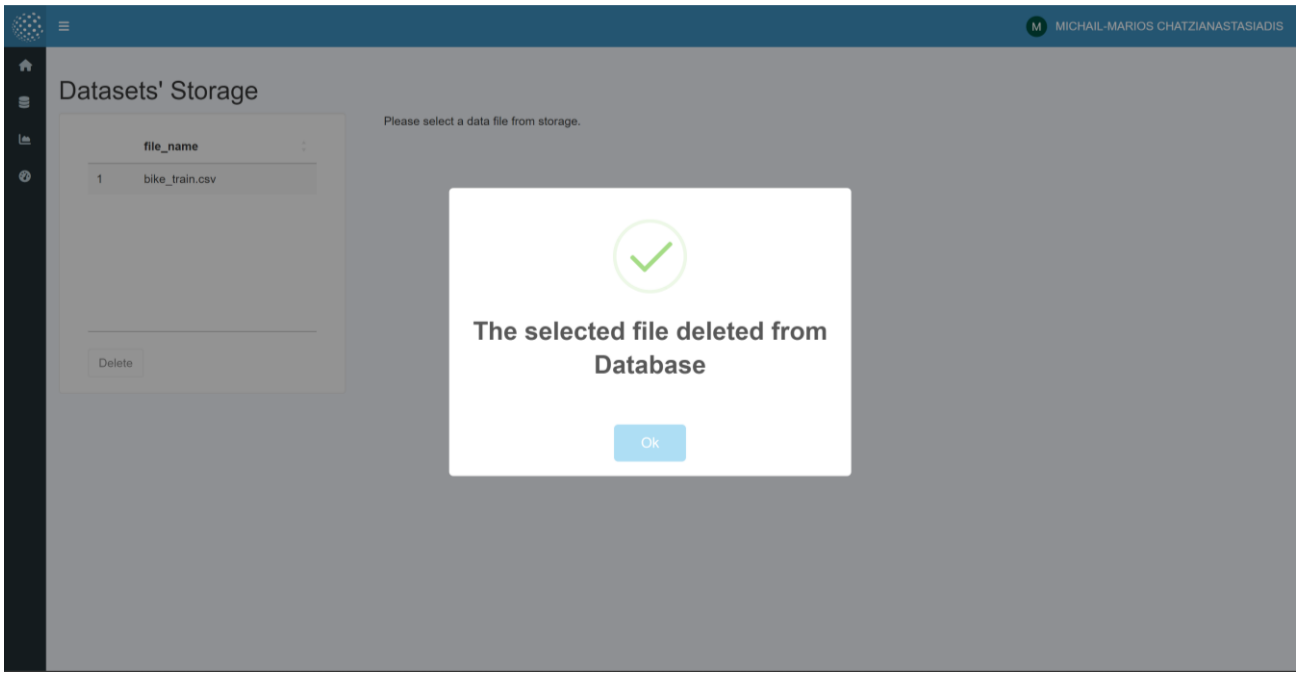
Εικόνα 22: Προεπισκόπηση Αποθηκευμένου Dataset ("bike_test.csv")

Όσο ένα Dataset είναι επιλεγμένο για προεπισκόπηση, ενεργοποιείται η δυνατότητα διαγραφής του από τη βάση δεδομένων, «πατώντας» το κουμπί **Delete**. Ο χρήστης ενημερώνεται, μέσω Warning ShinyAlert, ότι εάν προχωρήσει στην ενέργεια διαγραφής του Dataset, δεν θα υπάρχει η δυνατότητα επαναφοράς του. Η διαγραφή πραγματοποιείται, εφόσον ο χρήστης «πατήσει» το κουμπί **"Yes, delete it!"**, δίνοντας την έγκριση του (Εικόνα 23).



Εικόνα 23: Warning ShinyAlert - Επιβεβαίωση Διαγραφής Αρχείου

Τέλος, ο χρήστης ενημερώνεται μέσω Success ShinyAlert για την επιτυχή διαγραφή του Dataset από τη βάση δεδομένων, καθώς ανανεώνεται η λίστα του **Datasets' Storage** (Εικόνα 24).



Εικόνα 24: Success ShinyAlert - Ενημέρωση Επιτυχούς Διαγραφής

4.6.2.1 R Κώδικας για την Ανάκτηση της Λίστας των Αποθηκευμένων Datasets

Σε κάθε μετάβαση στη σελίδα **Datasets' Storage** ή στις περιπτώσεις που υπάρξει κάποια αλλαγή στα αποθηκευμένα Datasets, όπως μετά τη διαδικασία διαγραφής από τη βάση δεδομένων, πραγματοποιείται η αυτόματη ενημέρωση της λίστας των διαθέσιμων Datasets ανά χρήστη μέσω της κλήσης της συνάρτησης getSavedFiles(…):

```
getSavedFiles <-function(sub) {

 dbcon <- DBI::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password
 )

 query <-
  paste0(
   "SELECT file_name FROM uploadedfiles WHERE sub = '",
   sub,
   "';"
  )

 request <- DBI::dbGetQuery(dbcon, query)

 DBI::dbDisconnect(dbcon)

 return(request)
}
```

4.6.2.2 R Κώδικας για τη Διαγραφή Αποθηκευμένου Dataset

Η διαγραφή του εκάστοτε επιλεγμένου Dataset από τη βάση δεδομένων της εφαρμογής, πραγματοποιείται με την εκτέλεση της συνάρτησης deleteFilefromDB(…), έπειτα από την τελική έγκριση του χρήστη (**Delete** > **"Yes, delete it"**):

```
deleteFilefromDB <- function(sub,file_name) {

 dbcon <- DBI::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password
 )

 query <-
  paste0(
   "DELETE FROM uploadedfiles WHERE file_name = '",
   file_name,
   "' AND sub = '",
   sub,
   "';"
  )

 request <- DBI::dbGetQuery(dbcon, query)

 DBI::dbDisconnect(dbcon)


}
```

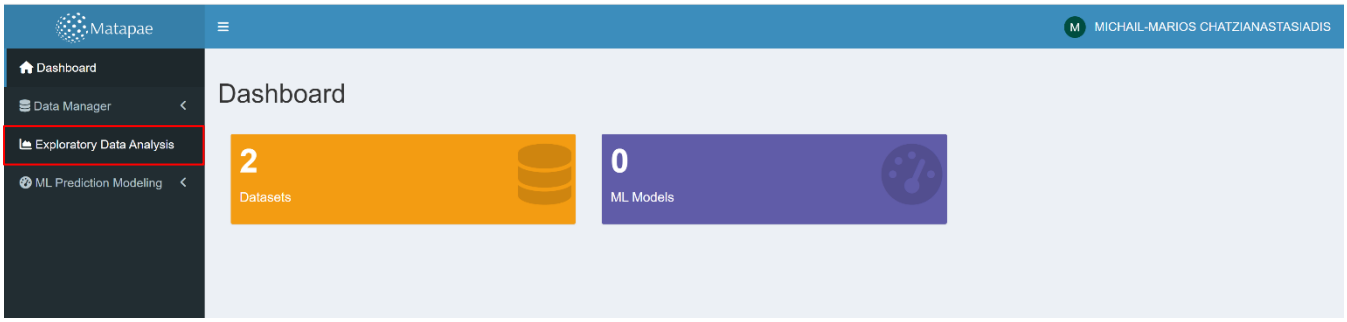
4.6.2.3 R Κώδικας για την Ανάκτηση Αποθηκευμένου Dataset

Με τη συνάρτηση getSelectedBlobFile(…) ανακτάται από τον «προσωπικό» χώρο αποθήκευσης του χρήστη, προς επεξεργασία, κάθε επιλεγμένο Dataset. Είτε για την προεπισκόπηση περιεχομένου του, στην ενότητα **Datasets' Storage**, είτε για τις οπτικοποιήσεις στην ενότητα **Exploratory Data Analysis**, είτε για τη συμμετοχή του στις διεργασίες **ML Prediction Modeling**.

```
getSelectedBlobFile <-function (sub, file_name){

 dbcon <- DBI::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password
 )

 query <- paste0("SELECT uploaded_file FROM uploadedfiles WHERE sub = '",
                 sub, "' AND file_name = '",file_name,"';")

 request <- DBI::dbGetQuery(dbcon, query)

 data <- read_csv(paste0(request, collapse = "\r\n"))

 DBI::dbDisconnect(dbcon)

 return(data)

}
```

4.7 Exploratory Data Analysis (EDA)

Ένα σύνολο δεδομένων συνήθως περιλαμβάνει ένα δείγμα αντιπροσωπευτικών μετρήσεων με συγκεκριμένα χαρακτηριστικά. Αυτές οι μετρήσεις και οι αντίστοιχες κατηγοριοποιήσεις τους αντιπροσωπεύουν τη δειγματική κατανομή της κάθε μεταβλητής, η οποία κατ' επέκταση αντιπροσωπεύει περίπου την πληθυσμιακή κατανομή των μεταβλητών. Κύριος στόχος της διερευνητικής ανάλυσης των δεδομένων (EDA) είναι να απεικονίσει και να συνοψίσει την κατανομή του δείγματος, επιτρέποντας στον αναλυτή να δημιουργήσει υποθέσεις σχετικά με την κατανομή του πληθυσμού και να τις ελέγξει, ώστε να γνωρίσει καλύτερα τα δεδομένα που μελετά [3]. Η εφαρμογή "Matapae", μέσω τεχνικών οπτικοποίησης που προσφέρει το πακέτο {DataExplorer[10]}, δίνει αυτή τη δυνατότητα στον αναλυτή, επιλέγοντας τη σχετική ενότητα που ονομάζεται **Exploratory Data Analysis** από το Sidebar Menu (Εικόνα 25).



Εικόνα 25: Επιλογή Exploratory Data Analysis από το Sidebar Menu

[10] https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html

4.7.1 Dataset Dimension

Η επιλογή **Dimension** προσφέρει τη διαγραμματική απεικόνιση της δομής ενός αποθηκευμένου Dataset (Εικόνα 26). Το Dataset, που έχει επιλεγεί ως παράδειγμα για τους σκοπούς του εγχειριδίου χρήσης είναι το “bike_train.csv”, περιλαμβάνει εγγραφές που αφορούν ενοικιάσεις ποδηλάτων. Ειδικότερα, η μεταβλητή **cnt** αναφέρεται στο σύνολο των ενοικιάσεων βάσει συγκεκριμένων γνωρισμάτων που περιγράφονται από τις παρακάτω μεταβλητές:

- **ID**: Μοναδικός αριθμός μέτρησης
- **season**: Εποχή του χρόνου
- **hr:** Ώρα της ημέρας
- **holiday:** Περίοδος (ή όχι) διακοπών
- **weekday:** Ημέρα της εβδομάδας
- **workingday:** Εργάσιμη (ή μη) ημέρα
- **weathersit:** Χαρακτηρισμός καιρικών συνθηκών
- **temp:** Βαθμοί Θερμοκρασίας (°C)
- **hum:** Ποσοστό ομίχλης (%)
- **windspeed:** Ταχύτητα ανέμου (Km/h)

Εικόνα 26: Dataset Dimension

4.7.2 Basic Information

Σύμφωνα με την Εικόνα 27, μέσω του Tab που ανοίγει με την επιλογής **Basic Info**, ο χρήστης πληροφορείται για την εκτιμώμενη μνήμη (σε Kbytes) που απαιτείται για τις ανάγκες της ανάλυσης του επιλεγμένου Dataset (Memory Usage). Ακόμα, δίνεται η δυνατότητα κατανόησης του είδους των μεταβλητών που περιλαμβάνει το Dataset, μέσω της διαγραμματικής απεικόνισης της ποσοστιαίας κατανομής των αριθμητικών μεταβλητών (Numerical Variables/Columns) σε:

- **Διακριτές Μεταβλητές (Discrete Variables/Columns):** Πρόκειται για μεταβλητές με ένα συγκεκριμένο σύνολο αριθμητικών τιμών που μπορούν να καταμετρηθούν ή να απαριθμηθούν.
    - Οι διακριτές μεταβλητές μπορούν επίσης να χαρακτηριστούν ως κατηγορικές μεταβλητές (Categorical Variables/Columns), όταν περιέχουν πεπερασμένο αριθμό τιμών, όπως είναι για παράδειγμα οι ημέρες της εβδομάδας.
- **Συνεχείς Μεταβλητές (Continuous Variables/Columns):** Οι παρατηρήσεις τους μπορούν να αποτελούνται από οποιαδήποτε τιμή σε όλο το εύρος των πραγματικών αριθμών. Η διαφορά μεταξύ δύο δυνατών τιμών τους μπορεί να είναι απειροελάχιστη. Παραδείγματα συνεχών μεταβλητών είναι: ο χρόνος, η θερμοκρασία, το ύψος, το βάρος κ.α.

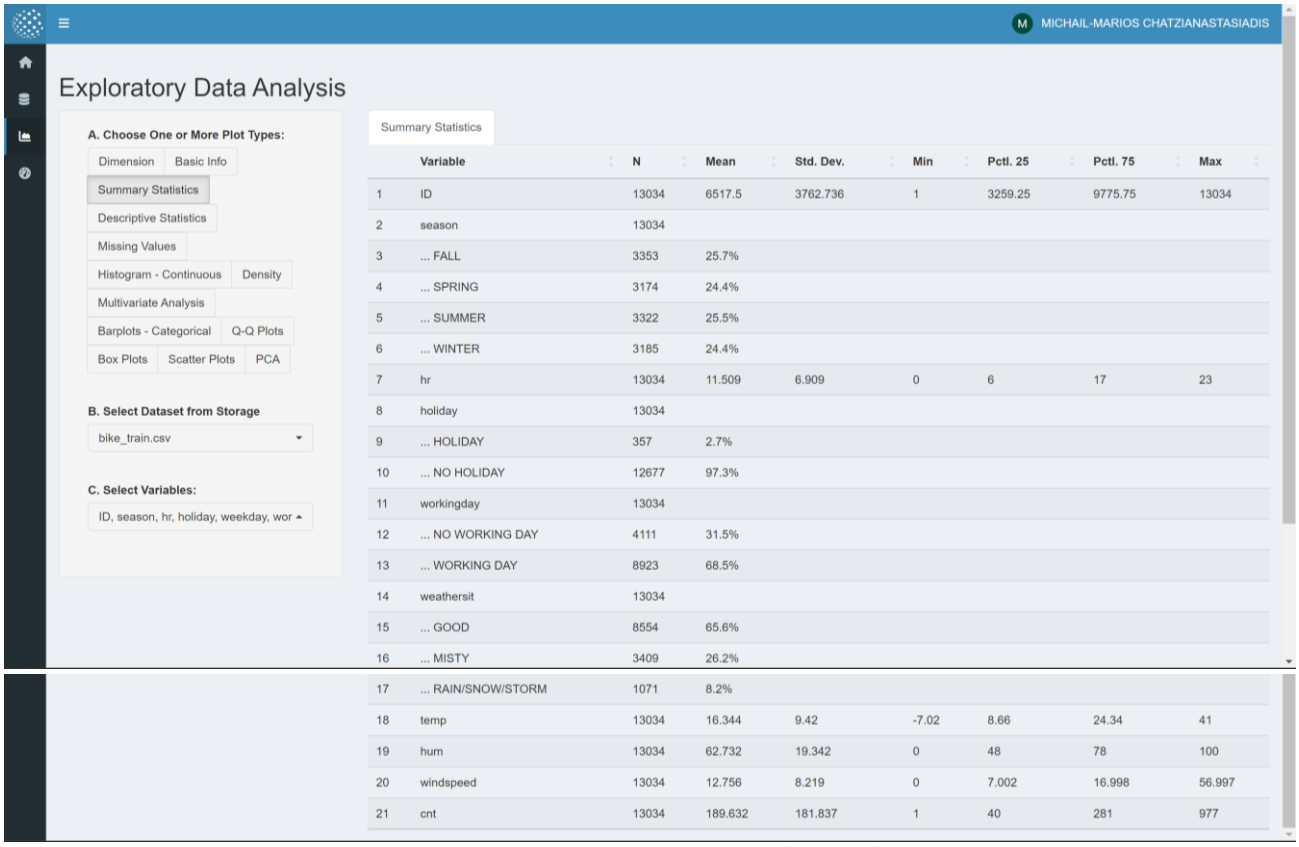
Επιπλέον, η μετρική **All Missing Columns**, περιγράφει το ποσοστό των στηλών του Dataset όπου λείπουν όλες οι τιμές από τις παρατηρήσεις τους. Η ένδειξη **Complete Rows** παρουσιάζει το ποσοστό των εγγραφών που δεν περιέχουν ελλείπεις τιμές, δηλαδή όσων εγγραφών (σειρών) οι παρατηρήσεις τους είναι πλήρως συμπληρωμένες. Τέλος, η ένδειξη **Missing Observations** αναφέρεται στο ποσοστό των κενών παρατηρήσεων σε ολόκληρο το Dataset.

Εικόνα 27: Basic Info

4.7.3 Summary Statistics

Με την επιλογή του Tab **Summary Statistics** (Εικόνα 28), παρέχονται συνοπτικά στατιστικά στοιχεία για τις μεταβλητές, όπως:

- **N:** Αριθμός των παρατηρήσεων, συνολικά και ανά κατηγορία στις περιπτώσεις των Categorical Variables.
- **Mean:** Μέσος όρος των παρατηρήσεων στις περιπτώσεις των Continuous Variable, ενώ για κάθε μια από τις κατηγορίες των Categorical Variables, δίνεται το ποσοστό της σχετικής συχνότητας.
- **Std. Dev.:** Η μετρική της τυπικής απόκλισης χρησιμοποιείται για να υπολογιστεί το ποσό της διασποράς ενός συνόλου τιμών. Μια χαμηλή τυπική απόκλιση υποδηλώνει ότι τα σημεία των παρατηρήσεων του δείγματος είναι κοντά στο μέσο όρο, ενώ μια υψηλή υποδεικνύει ότι τα σημεία των παρατηρήσεων απλώνονται σε ένα μεγαλύτερο εύρος τιμών.
- **Min/Max:** Μεγαλύτερη και μικρότερη τιμή των παρατηρήσεων κάθε μεταβλητής.
- **Percentile (Pctl.) 25%:** Υποδηλώνει ότι το 25% των παρατηρήσεων βρίσκεται κάτω από αυτή τη τιμή.
- **Percentile (Pctl.) 75%:** Υποδηλώνει ότι το 25% των παρατηρήσεων βρίσκεται πάνω από αυτή τη τιμή.

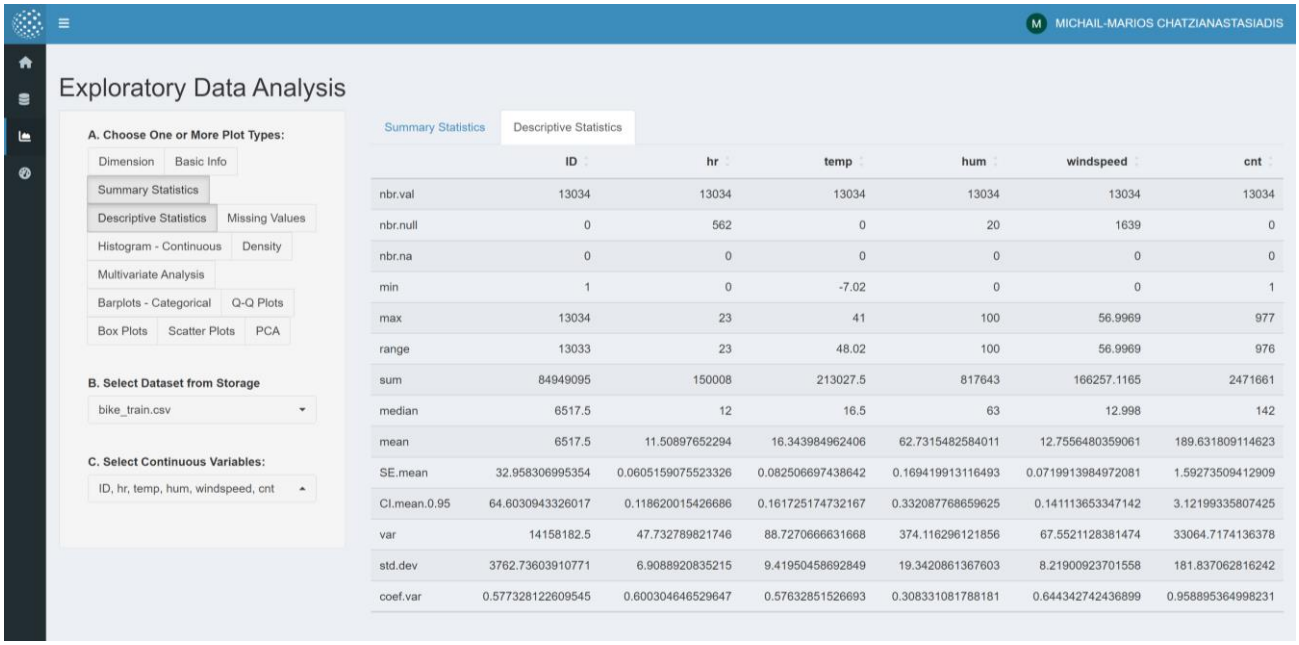


	Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
1	ID	13034	6517.5	3762.736	1	3259.25	9775.75	13034
2	season	13034						
3	... FALL	3353	25.7%					
4	... SPRING	3174	24.4%					
5	... SUMMER	3322	25.5%					
6	... WINTER	3185	24.4%					
7	hr	13034	11.509	6.909	0	6	17	23
8	holiday	13034						
9	... HOLIDAY	357	2.7%					
10	... NO HOLIDAY	12677	97.3%					
11	workingday	13034						
12	... NO WORKING DAY	4111	31.5%					
13	... WORKING DAY	8923	68.5%					
14	weathersit	13034						
15	... GOOD	8554	65.6%					
16	... MISTY	3409	26.2%					
17	... RAIN/SNOW/STORM	1071	8.2%					
18	temp	13034	16.344	9.42	-7.02	8.66	24.34	41
19	hum	13034	62.732	19.342	0	48	78	100
20	windspeed	13034	12.756	8.219	0	7.002	16.998	56.997
21	cnt	13034	189.632	181.837	1	40	281	977

Εικόνα 28: Summary Statistics

4.7.4 Descriptive Statistics

Στο Tab **Descriptive Statistics**, παρέχονται επιπλέον στατιστικοί δείκτες για την κατανόηση των δεδομένων, όπως εμφανίζονται στην Εικόνα 30:



	ID	hr	temp	hum	windspeed	cnt
nbr.val	13034	13034	13034	13034	13034	13034
nbr.null	0	562	0	20	1639	0
nbr.na	0	0	0	0	0	0
min	1	0	-7.02	0	0	1
max	13034	23	41	100	56.9969	977
range	13033	23	48.02	100	56.9969	976
sum	84949095	150008	213027.5	817643	166257.1165	2471661
median	6517.5	12	16.5	63	12.998	142
mean	6517.5	11.50897652294	16.343984962406	62.7315482584011	12.7556480359061	189.631809114623
SE.mean	32.958306995354	0.0605159075523326	0.082506697438642	0.169419913116493	0.0719913984972081	1.59273509412909
CI.mean.0.95	64.6030943326017	0.118620015426686	0.161725174732167	0.332087768659625	0.141113653347142	3.12199335807425
var	14158182.5	47.732789821746	88.7270666631668	374.116296121856	67.5521128381474	33064.7174136378
std.dev	3762.73603910771	6.9088920835215	9.41950458692849	19.3420861367603	8.21900923701558	181.837062816242
coef.var	0.577328122609545	0.600304646529647	0.57632851526693	0.308331081788181	0.644342742436899	0.958895364998231

Εικόνα 29: Descriptive Statistics

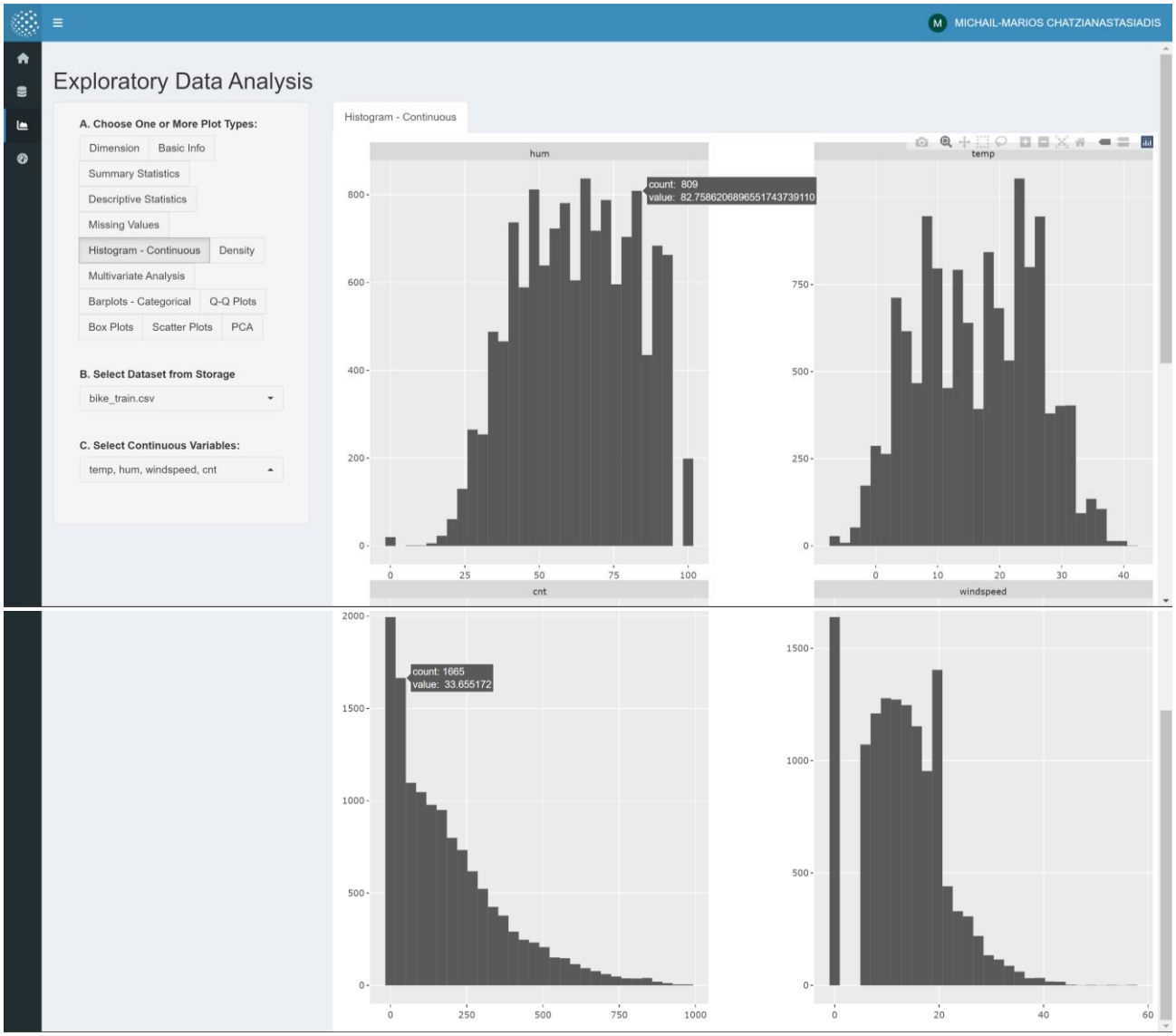
4.7.5 Missing Values

Στο Tab **Missing Values** (Εικόνα 30), ο χρήστης μπορεί να λάβει την πληροφορία για το ποσοστό των παρατηρήσεων που δεν έχουν τιμή ανά Feature (μεταβλητή). Οι κενές τιμές μπορεί να έχουν τη σημασία τους για ένα Feature. Για παράδειγμα, σε ένα Discrete Feature, πιθανώς να χρειάζεται η ομαδοποίηση των κενών τιμών σε μια ξεχωριστή ομάδα. Σε ένα Continuous Feature, ίσως χρειαστεί η συμπλήρωση των κενών τιμών με μια τιμή βάσει της υπάρχουσας γνώσης που προκύπτει από το ίδιο το Dataset. Ειδικότερα, μια πιθανή προσέγγιση συμπλήρωσης των κενών είναι υπολογίζοντας είτε τον μέσο όρο της προηγούμενης και της επόμενης παρατήρησης είτε τον μέσο όρο του συνόλου των παρατηρήσεων του εκάστοτε Feature.

Εικόνα 30: Missing Values

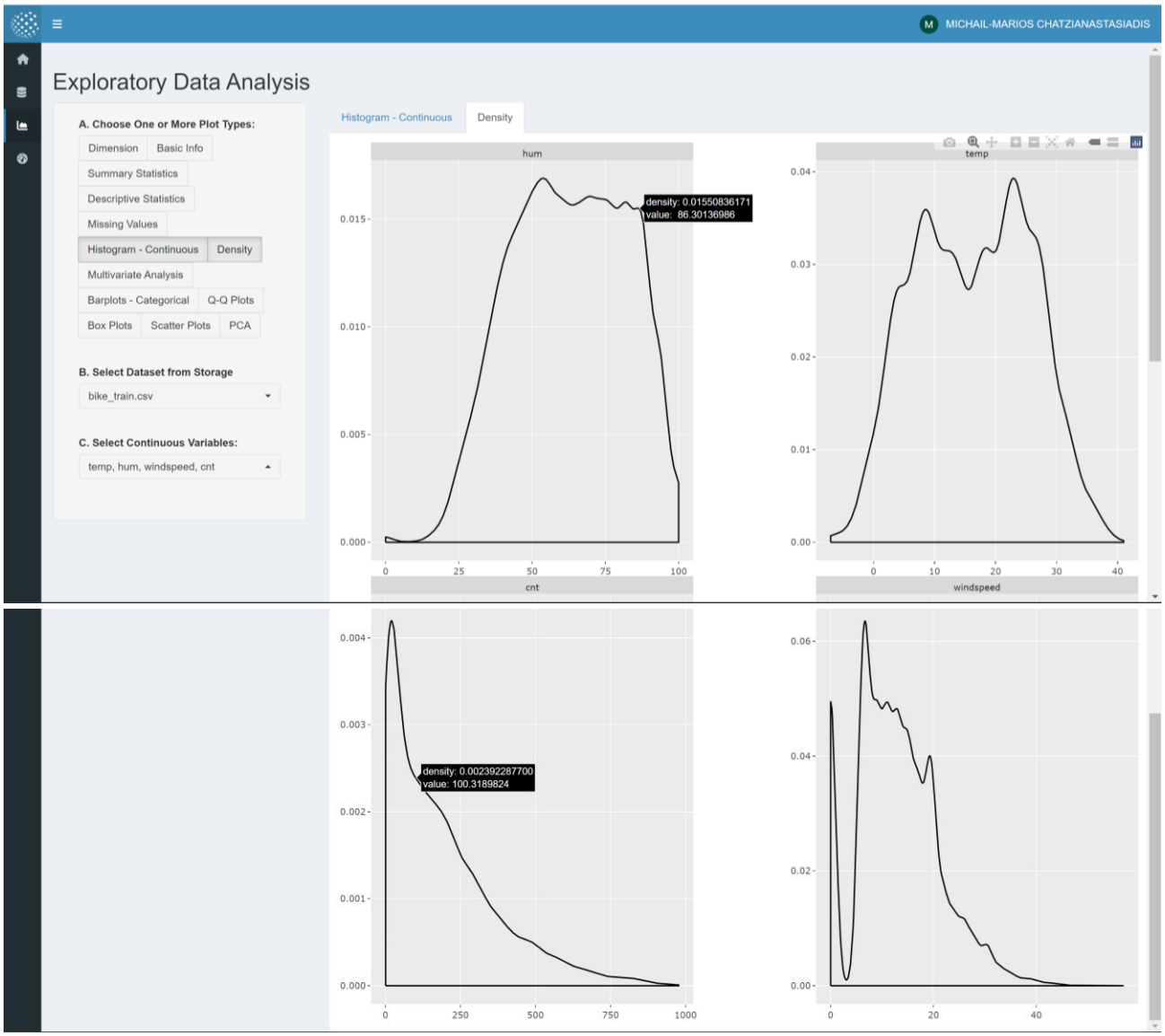
4.7.6 Histograms & Density Plots για τις Continuous Μεταβλητές

Το Tab **Histogram – Continuous** της Εικόνα 31, δίνει τη δυνατότητα οπτικοποίησης της κατανομής συχνοτήτων που παρουσιάζουν οι συνεχείς μεταβλητές του επιλεγμένου Dataset.



Εικόνα 31: Histograms για συνεχείς μεταβλητές

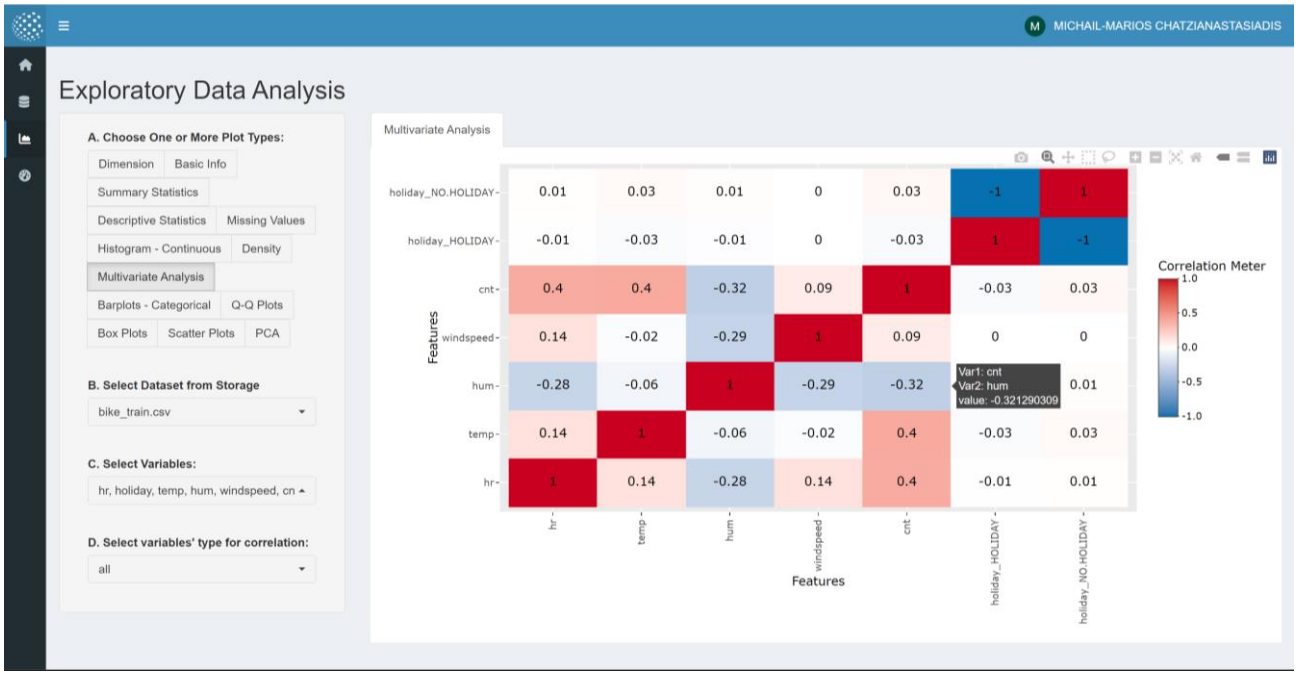
Τα διαγράμματα πυκνοτήτων (**Density** Plots), της Εικόνα 32, είναι μια συνεχής και ομαλοποιημένη έκδοση των Ιστογραμμάτων (Histograms). Δημιουργούνται μέσω της εκτίμησης πυκνότητας πυρήνα (Kernel Density Estimation – KDE).



Εικόνα 32: Density Plots

4.7.7 Multivariate Analysis

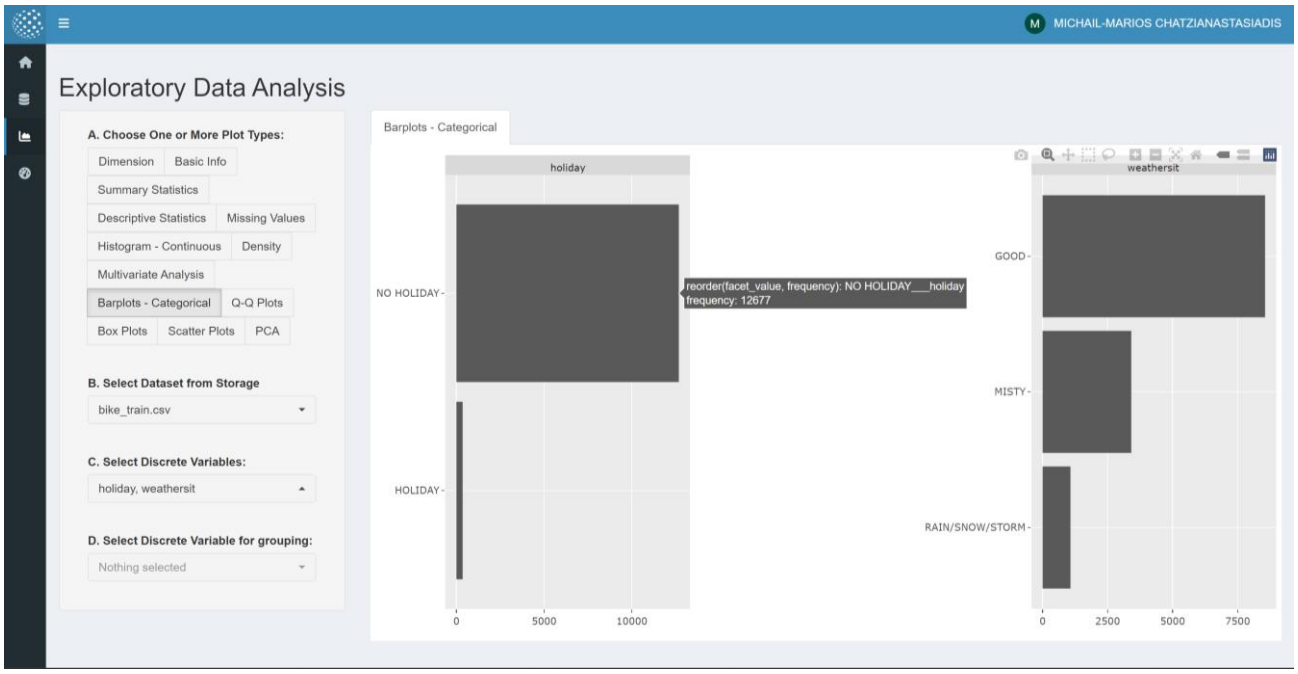
Το Tab **Multivariate Analysis** παρέχει τη δυνατότητα για οπτικοποιήσεις συσχετίσεων (Correlation Heatmap). Σύμφωνα με την Εικόνα 33, η μεταβλητή της υγρασίας (hum) με τη μεταβλητή των ενοικιάσεων ποδηλάτων παρουσιάζουν αρνητική συσχέτιση μεταξύ τους (-0.32).



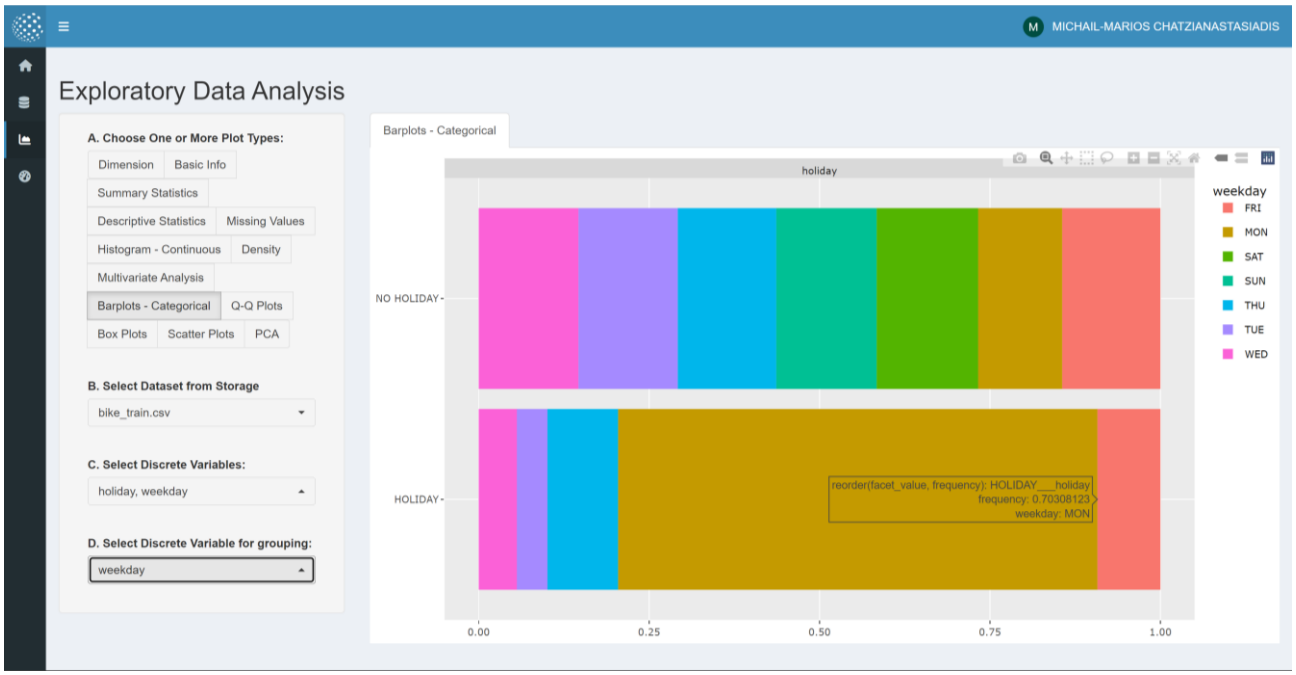
Εικόνα 33: Multivariate Analysis

4.7.8 Bar Plots για τις Categorical Μεταβλητές

Η επιλογή **Barplots – Categorical** προσφέρει οπτικοποιήσεις της κατανομής συχνοτήτων των διακριτών μεταβλητών, είτε μεμονωμένα βάσει των κατηγοριών της ίδιας της μεταβλητής που επιλέγεται στο πεδίο C (Εικόνα 34), είτε διαχωρισμένες βάσει των κατηγοριών κάποιας άλλης διακριτής μεταβλητής (Discrete Variable) που επιλέγεται στο πεδίο D (Εικόνα 35).



Εικόνα 34: Simple Categorical Bar Plots

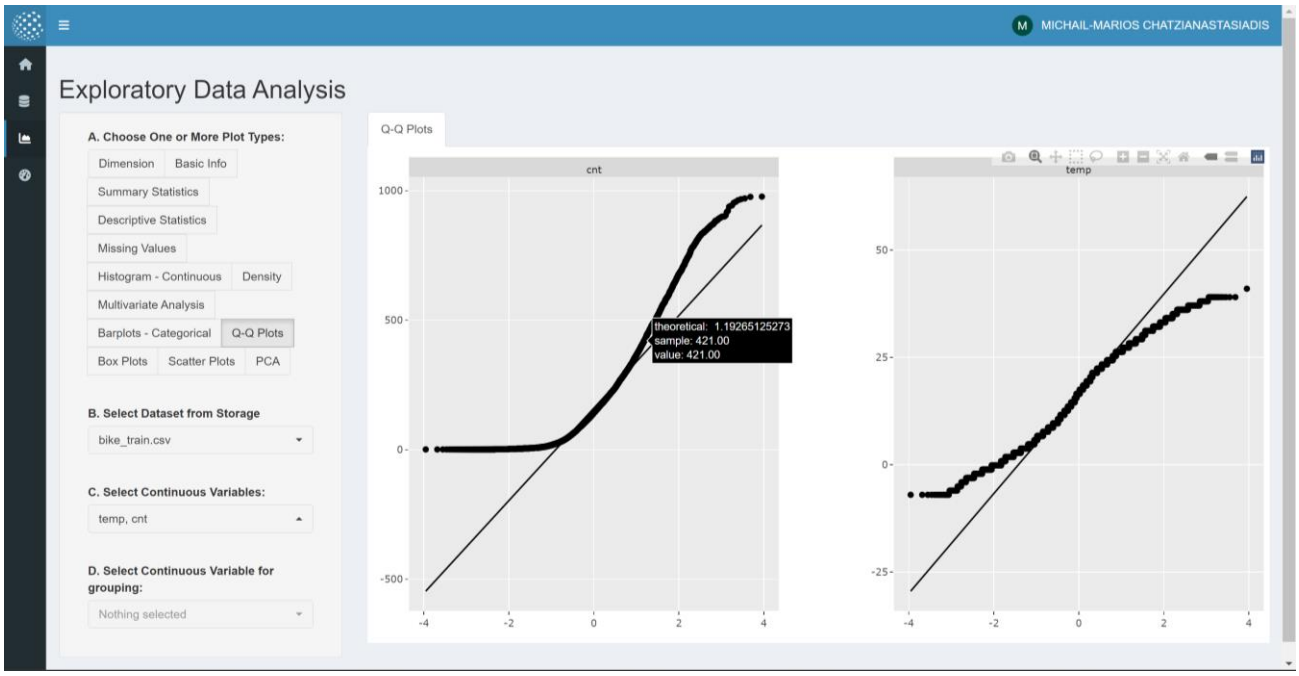


Εικόνα 35: Categorical Bar Plots Using Grouping Field

4.7.9 Quantile-Quantile Plots – Κανονική Κατανομή (Normal Distribution)

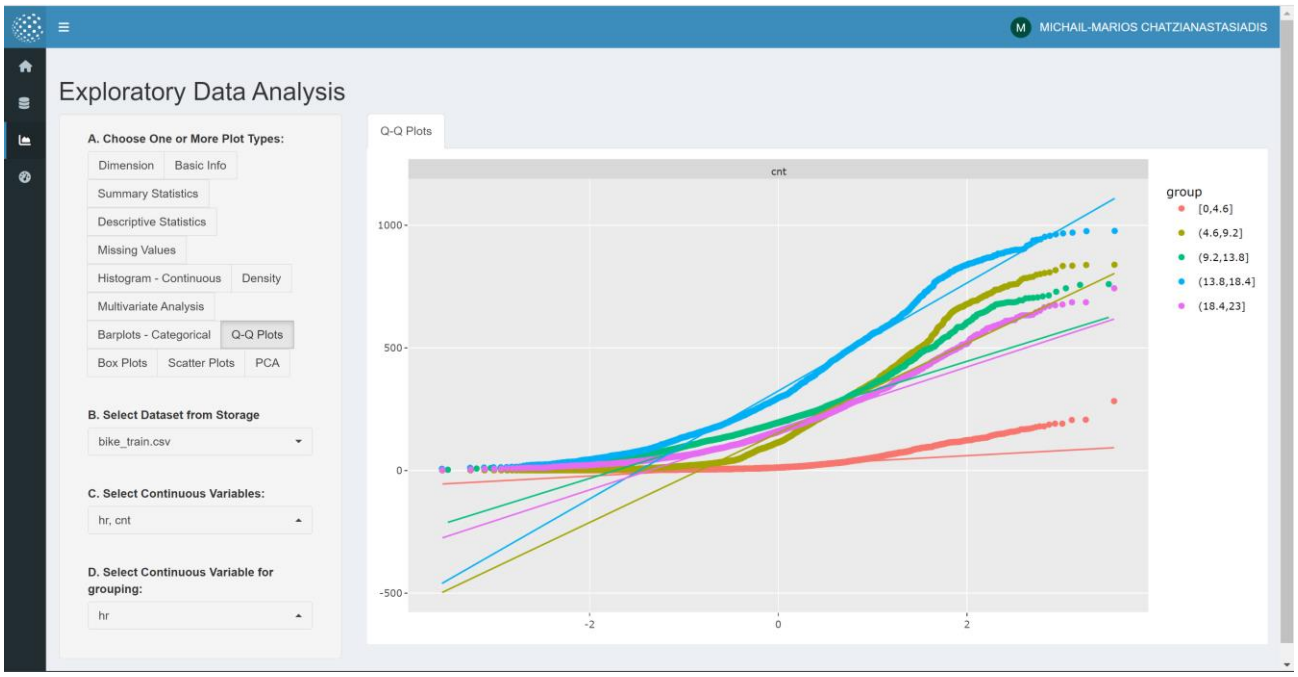
Ο έλεγχος απόκλισης των δεδομένων από μια συγκεκριμένη πιθανότητα θεωρητικής κατανομής, γίνεται με τις γραφικές παραστάσεις των δειγματικών ποσοστημορίων (**Q-Q Plots**), ως προς τα θεωρητικά ποσοστημόρια της εκάστοτε κατανομής. Στην τρέχουσα υλοποίηση της εφαρμογής "Matapae" (Εικόνα 36), ελέγχεται εάν τα δεδομένα ακολουθούν κανονική κατανομή.

Από την ανάλυση των διαγραμμάτων, δίνεται η δυνατότητα στον χρήστη να καταλάβει εάν υπάρχει η ανάγκη μαθηματικού μετασχηματισμού των δεδομένων. Ένα παράδειγμα αποτελεί ο λογαριθμικός μετασχηματισμός (Log Transformation), όπου συνήθως προτείνεται για τα δεδομένα που πρόκειται να χρησιμοποιηθούν στη δημιουργία μοντέλων γραμμικής παλινδρόμησης (Linear Regression). Πιο επεξηγηματικά, στις συνεχείς μεταβλητές (Continuous Variables), όσο τα σημεία των δεδομένων «απομακρύνονται» από τη γραμμή που αναπαριστά τα θεωρητικά ποσοστημόρια της κανονικής κατανομής, τόσο «χειρότερη» είναι η προσαρμογή τους στην κατανομή. Συμπεραίνεται λοιπόν, ότι τα δεδομένα δεν προέρχονται από κανονική κατανομή. Σε αυτή την περίπτωση, για να επιτευχθούν καλύτερα και πιο έγκυρα αποτελέσματα σε οποιαδήποτε στατιστική ανάλυση ή δημιουργία μοντέλων μηχανικής μάθησης, θα πρέπει πρώτα να εφαρμοστεί κάποιος μαθηματικός μετασχηματισμός (π.χ. Log Transformation), έτσι ώστε τα δεδομένα να τείνουν τελικά όσο το δυνατόν περισσότερο στην κανονική κατανομή [38].



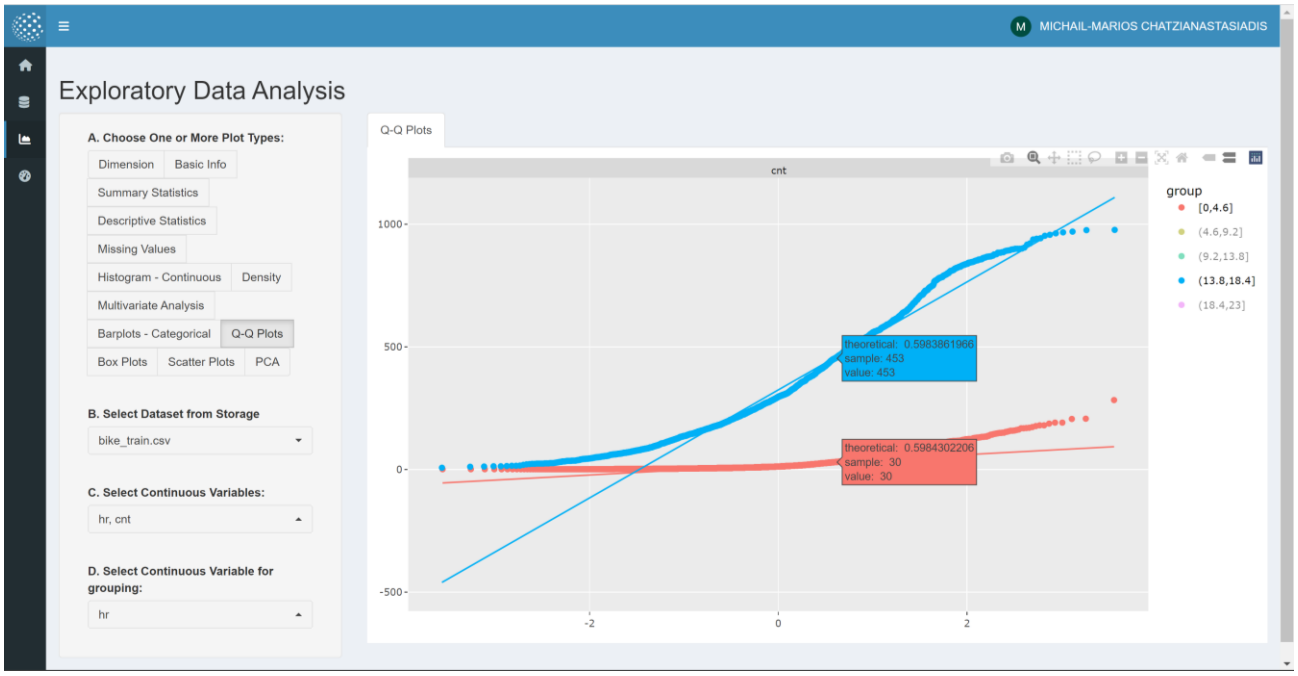
Εικόνα 36: Simple Q-Q Plots

Εάν κρίνεται απαραίτητο, δίνεται η δυνατότητα σύγκρισης με ομαδοποίηση τιμών από κάποιο άλλο Feature (Εικόνα 37).



Εικόνα 37: Q-Q Plots with Grouping

Τέλος, προσφέρονται οι δυνατότητες επιλογής για απόκρυψη ή εμφάνιση των επιμέρους Q-Q Plots από το υπόμνημα των Groups, καθώς επίσης και ταυτόχρονης σύγκρισης των σημείων του γραφήματος βάσει του άξονα X (Εικόνα 38).

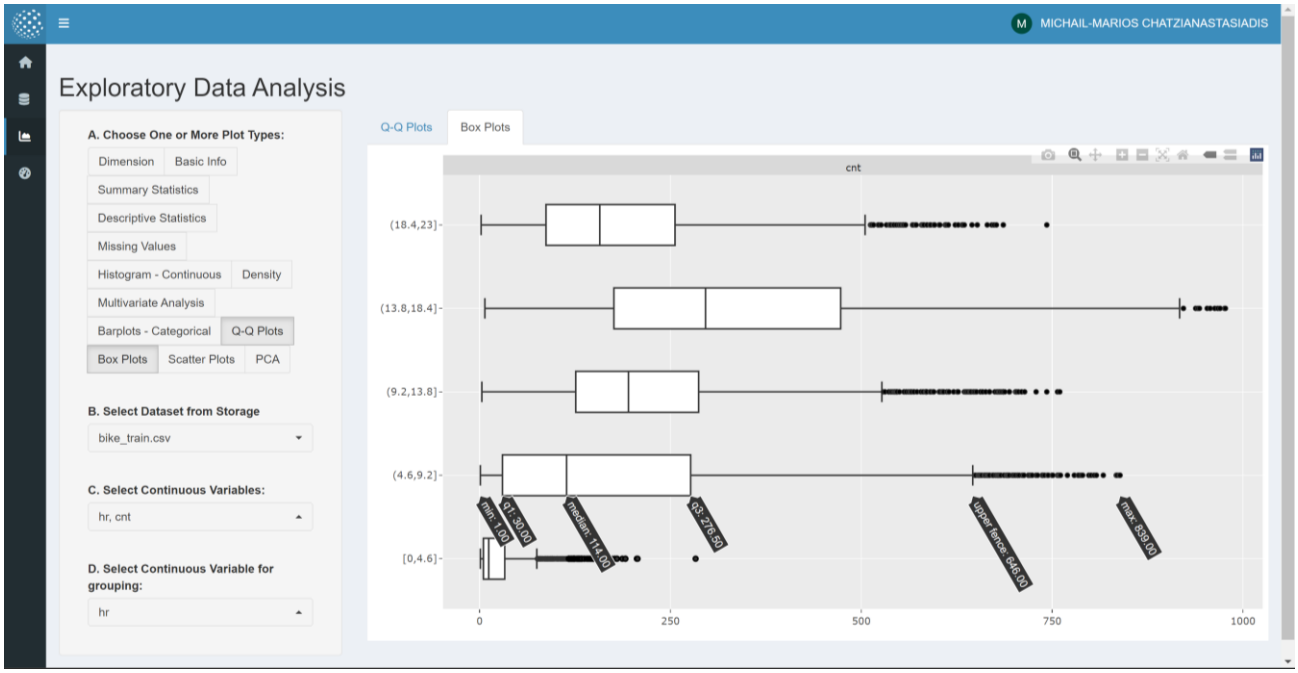


Εικόνα 38: Q-Q Plot με Δυνατότητες Επιλογής και Σύγκρισης

4.7.10 Box Plots

Μια ακόμα δυνατότητα απεικόνισης των δεδομένων είναι τα **Box Plots** (Εικόνα 39). Πρόκειται για αναπαραστάσεις που χρησιμεύουν στον γραφικό έλεγχο της κατανομής των Continuous Variables σε σχέση με κάποιο άλλο Feature. Το Hovering του χρήστη επάνω στα επιμέρους Box Plots προσφέρει άμεσα τις παρακάτω πληροφορίες για τα δεδομένα:

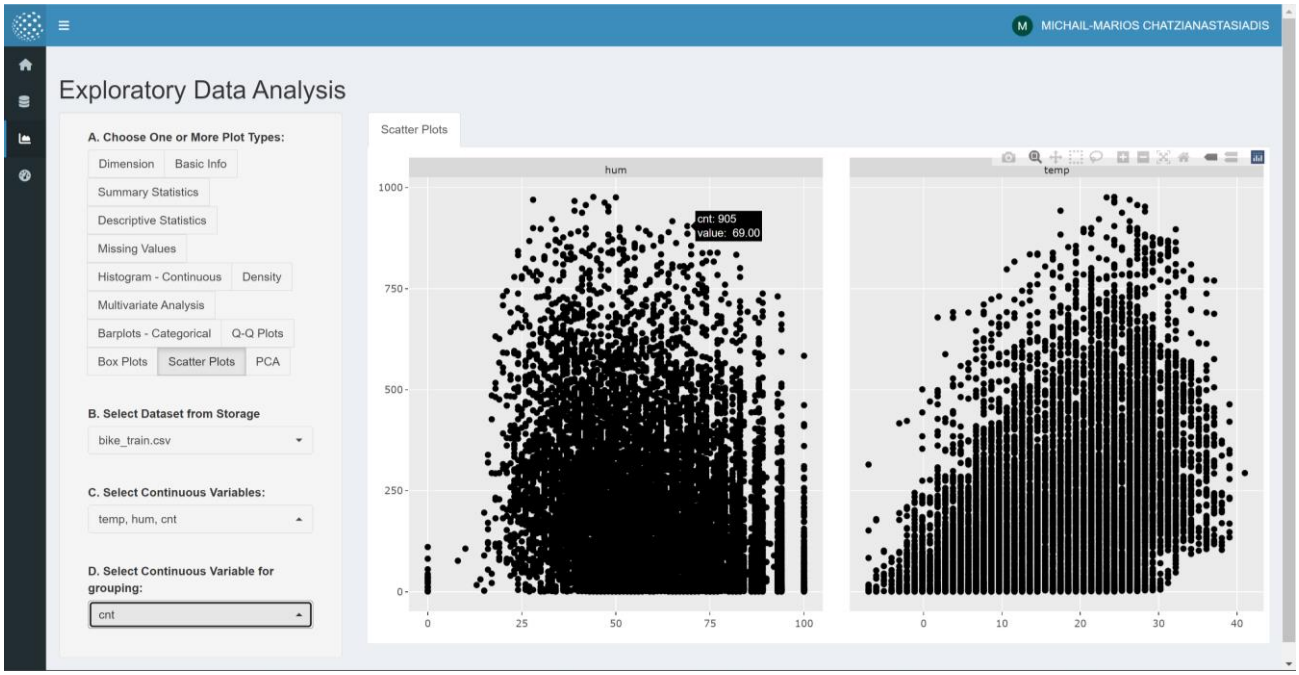
- Ελάχιστη τιμή (**Min**)
- Μέγιστη τιμή (**Max**)
- 1ο τεταρτημόριο (**Q1** – το 25% των παρατηρήσεων)
- 3ο τεταρτημόριο (**Q2** – το 75% των παρατηρήσεων)
- Διάμεσος (**Median**)
- Άνω φράγμα (**Upper Fence**): Οι παρατηρήσεις μετά από εκείνο το σημείο χαρακτηρίζονται ως Outliers.



Εικόνα 39: Box Plots with Grouping

4.7.11 Scatter Plots

Τα **Scatter Plots** είναι μια επιπλέον επιλογή γραφικής αναπαράστασης για τις Continuous Variables με δυνατότητα ομαδοποίησης βάσει ενός επιλεγμένου Feature (Εικόνα 40).

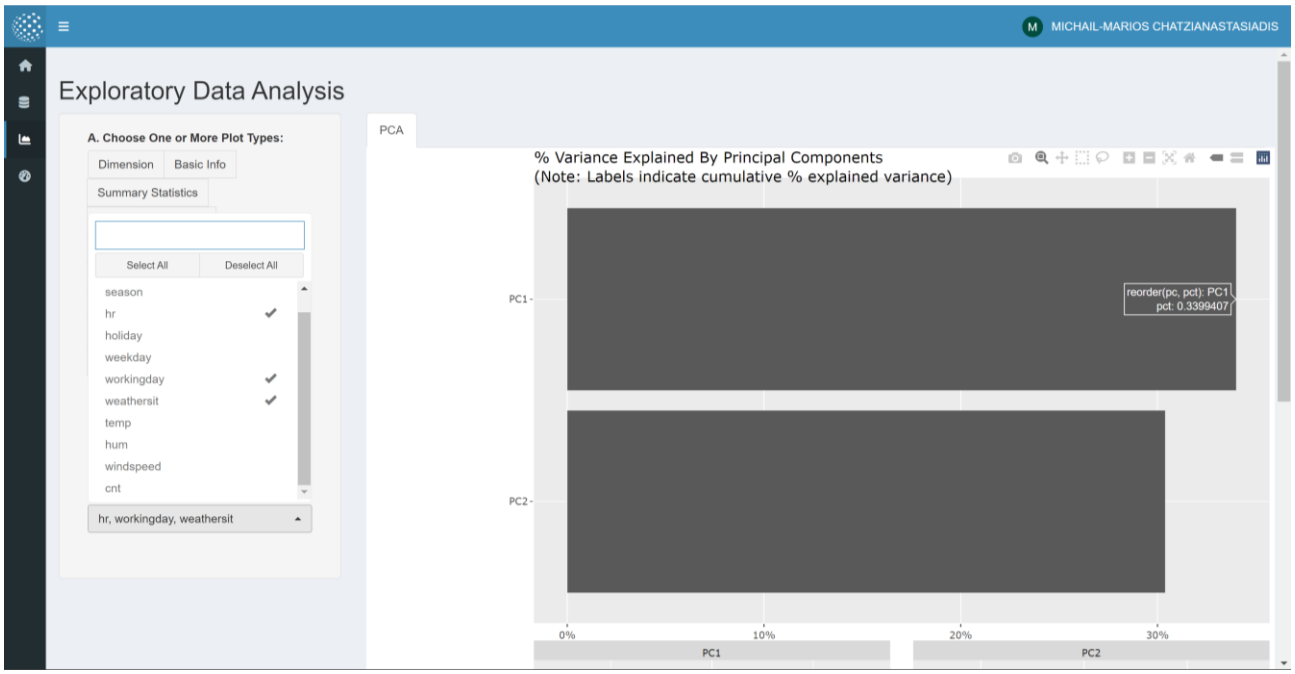


Εικόνα 40: Scatter Plots with Grouping

4.7.12 Principal Component Analysis (PCA)

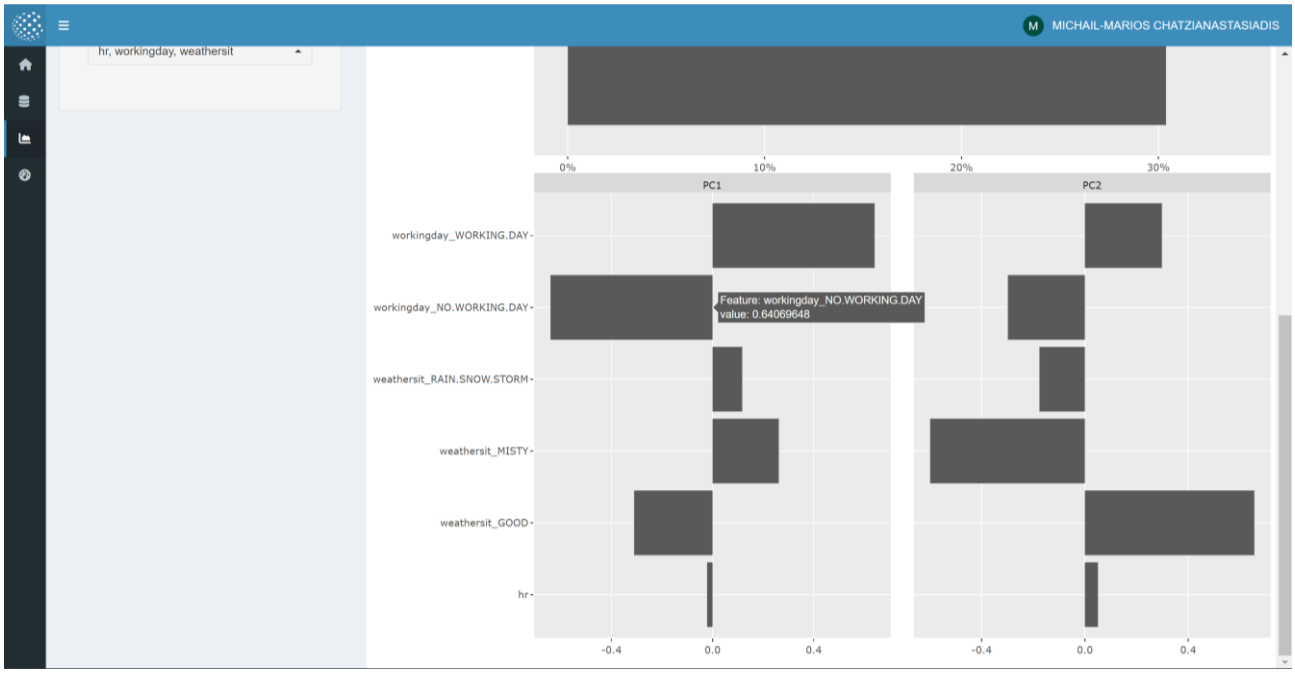
Οι χρήστες της εφαρμογής έχουν τη δυνατότητα να αναλύσουν καλύτερα, να καταλάβουν και να εξάγουν συμπεράσματα για τα Datasets μεγάλων διαστάσεων μέσω της Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis – **PCA**). Πρόκειται για μια στατιστική μέθοδο που μετατρέπει ένα σύνολο παρατηρήσεων, πιθανώς συσχετισμένων μεταβλητών, σε ένα σύνολο τιμών γραμμικά ασύνδετων μεταβλητών που ονομάζονται κύριες συνιστώσες (PC). Είναι ένα εργαλείο οπτικοποίησης και μείωσης των διαστάσεων. Επισημαίνεται ότι η δημιουργία PCs και η αποθήκευση τους ως νέα Features στον χώρο αποθήκευσης (Storage) ώστε να μπορούν να χρησιμοποιηθούν για τη δημιουργία μοντέλων μηχανικής μάθησης, δεν έχει υλοποιηθεί στην τρέχουσα έκδοση της εφαρμογής “Matapae”. Από την PCA προκύπτουν δύο τύποι γραφημάτων.

Ο πρώτο τύπος γραφήματος, της Εικόνα 42, παρουσιάζει το ποσοστό της διακύμανσης που ερμηνεύουν (περιέχουν) οι ασυσχέτιστες μεταβλητές των κύριων συνιστωσών. Το ποσοστό αυτό θα πρέπει να είναι όσο γίνεται μεγαλύτερο, ώστε να ερμηνεύεται το μεγαλύτερο μέρος της διακύμανσης των αρχικών δεδομένων.



Εικόνα 41: Principal Component Analysis – Πρώτος Τύπος Γραφήματος

Στο δεύτερο τύπο γραφημάτων, της Εικόνα 42, επεξηγούνται οι κύριες συνιστώσες μια προς μια. Ο χρήστης κάνοντας Hover επάνω στο γράφημα, μπορεί να διακρίνει τον βαθμό επίδρασης της κάθε μεταβλητής σε κάθε κύρια συνιστώσα μέσω της ένδειξης Value.



Εικόνα 42: Principal Component Analysis – Δεύτερος Τύπος Γραφημάτων

4.7.13 R Κώδικας για τη διαχείριση των EDA Tabs

Καθώς ο χρήστης επιλέγει τους διαφορετικούς τύπους οπτικοποιήσεων από το πεδίο A του Sidebar της ενότητας Exploratory Data Analysis, δημιουργούνται στο UI περιβάλλον τα Tabs με τα αντίστοιχα γραφήματα του πακέτου {DataExplorer} της R. Η διαδικασία αυτή γίνεται μέσω της συνάρτησης create_tabs(…):

```
output$plot_tabs <- renderUI({

      req(input$select_dataset)
      req(input$plot_types)

  create_tabs <- function(x) {

    shiny::tabPanel(x, {

        if (x == "Dimension") {

         renderUI({

             req(input$selected_vars_EDA)

             radialNetwork(
              DataExplorer::plot_str(
               values$selected_file_from_DB_to_plot[input$selected_vars_EDA],
               type = "radial" )}, width = "auto")

         })

        } else if (x == "Basic Info") {

          renderPlot({

              req(input$selected_vars_EDA)

              DataExplorer::plot_intro(
               values$selected_file_from_DB_to_plot[input$selected_vars_EDA])

          }, height = 590)

        } else if (x == "Summary Statistics") {

           renderDataTable({

               req(input$selected_vars_EDA)

               st(values$selected_file_from_DB_to_plot
                  [input$selected_vars_EDA], out="return")

           }, options = list(scrollX = TRUE, pageLength = -1, dom = 't'))

        } else if (x == "Descriptive Statistics") {

           renderDataTable({

              req(input$selected_vars_EDA_continuous)

              stat.desc(values$selected_file_from_DB_to_plot
                        [input$selected_vars_EDA_continuous])
```

```
    }, options = list(scrollX = TRUE, pageLength = -1, dom = 't'))

} else if (x == "Missing Values") {

  renderPlot({

      req(input$selected_vars_EDA)

      DataExplorer::plot_missing(
       values$selected_file_from_DB_to_plot[input$selected_vars_EDA]
      )

  }, height = 590)

} else if (x == "Histogram - Continuous") {

  renderPlotly({

      req(input$selected_vars_EDA_continuous)

      Hist <-
      DataExplorer::plot_histogram(
                  values$selected_file_from_DB_to_plot
                  [input$selected_vars_EDA_continuous],
                  nrow = 1L,
                  ncol = 2L)

      subplot(Hist, nrows = length(Hist), margin = 0.02)
      %>% layout(height = (length(Hist) * 590))

  })

} else if (x == "Density") {

  renderPlotly({

   req(input$selected_vars_EDA_continuous)

   Density <-
   DataExplorer::plot_density(
          values$selected_file_from_DB_to_plot
          [input$selected_vars_EDA_continuous],
          nrow = 1L,
          ncol = 2L)

   subplot(Density, nrows = length(Density), margin = 0.02)
   %>% layout(height = (length(Density) * 590))


  })

} else if (x == "Multivariate Analysis") {

  renderPlotly({

      req(input$selected_vars_EDA)

      DataExplorer::plot_correlation(
          values$selected_file_from_DB_to_plot
             [input$selected_vars_EDA],
          type = input$select_corr_calc_type,
          cor_args = list(
```

```
                    "use" = "pairwise.complete.obs"))
                    %>% ggplotly(height = 590)

  })

} else if (x == "Barplots - Categorical") {

 renderPlotly({

     req(input$selected_vars_EDA_discrete)

     Barplot <-
      DataExplorer::plot_bar(
             values$selected_file_from_DB_to_plot
             [input$selected_vars_EDA_discrete],
         by = input$selected_vars_EDA_grouped_discrete,
         nrow = 1L,
         ncol = 2L)

      subplot(Barplot, nrows = length(Barplot), margin = 0.02)
      %>% layout(height = (length(Barplot) * 590))

  })

} else if (x == "Q-Q Plots") {

 renderPlotly({

     req(input$selected_vars_EDA_continuous)

     QQ <-
      DataExplorer::plot_qq(
             values$selected_file_from_DB_to_plot
             [input$selected_vars_EDA_continuous],
         by = input$selected_vars_EDA_grouped_continuous,
         nrow = 1L,
         ncol = 2L)

      subplot(QQ, nrows = length(QQ), margin = 0.02)
      %>% layout(height = (length(QQ) * 590))

  })

} else if (x == "Box Plots") {

  renderPlotly({

      req(input$selected_vars_EDA_continuous)

      Box <-
       DataExplorer::plot_boxplot(
          values$selected_file_from_DB_to_plot
             [input$selected_vars_EDA_continuous],
          by = input$selected_vars_EDA_grouped_continuous,
          nrow = 1L,
          ncol = 2L)

       subplot(Box, nrows = length(Box), margin = 0.02)
       %>% layout(height = (length(Box) * 590))


   })
```

```
      } else if (x == "Scatter Plots") {


        renderPlotly({

            req(input$selected_vars_EDA_continuous)
            req(input$selected_vars_EDA_grouped_continuous)

            Scatter <-
             DataExplorer::plot_scatterplot(
                      values$selected_file_from_DB_to_plot
                      [input$selected_vars_EDA_continuous],
                      by = input$selected_vars_EDA_grouped_continuous,
                      nrow = 1L,
                      ncol = 2L)

            subplot(Scatter, nrows = length(Scatter), margin = 0.02)
            %>% layout(height = (length(Scatter) * 590))

        })

      } else if (x == "PCA") {

        renderPlotly({

            req(input$selected_vars_EDA)

            PCA <-
             DataExplorer::plot_prcomp(na.omit(
                         values$selected_file_from_DB_to_plot
                         [input$selected_vars_EDA]),
                      nrow = 1L,
                      ncol = 2L)

            subplot(PCA, nrows = length(PCA), margin = 0.02)
            %>% layout(height = (length(PCA) * 590))

        })

      }

  })

 }

 myTabs <- lapply(input$plot_types, create_tabs)

 do.call(tabsetPanel, c(id = "plots_tabsetPanel", myTabs))

})
```

4.8 ML Prediction Modeling

Στην τελευταία ενότητα του μενού **ML Prediction Modeling** (Εικόνα 43: Μενού επιλογών ML Prediction Modeling με ήδη αποθηκευμένα DatasetsΕικόνα 43), υπάρχουν δύο επιμέρους επιλογές:

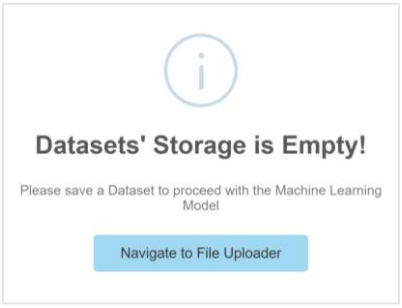
- **Supervised ML:** Όπου ο χρήστης έχει τη δυνατότητα να δημιουργήσει και να εκπαιδεύσει Supervised Machine Learning Models σε User-friendly περιβάλλον με καθορισμένα βήματα.
- **ML Models' Storage:** Πρόκειται για το περιβάλλον μέσω του οποίου ο χρήστης μπορεί να διαχειριστεί τα αποθηκευμένα μοντέλα μηχανικής μάθησης και να έχει πρόσβαση στις διαθέσιμες λεπτομέρειες των αποτελεσμάτων της εκπαίδευσης τους. Επιπλέον, ο χρήστης μπορεί να «τρέξει» τη διαδικασία προβλέψεων σε νέα δεδομένα. Στην περίπτωση που δεν έχει αποθηκευτεί τουλάχιστον ένα εκπαιδευμένο μοντέλο στη βάση δεδομένων της εφαρμογής, η επιλογή του μενού **ML Models' Storage** δεν είναι διαθέσιμη (Εικόνα 44).

Εικόνα 43: Μενού επιλογών ML Prediction Modeling με ήδη αποθηκευμένα Datasets

Εικόνα 44: Μενού επιλογών ML Prediction Modeling χωρίς αποθηκευμένα Datasets

4.8.1 Supervised ML

Στην περίπτωση που ο χρήστης προχωρήσει με την επιλογή **Supervised ML**, χωρίς να υπάρχει ήδη αποθηκευμένο Dataset στην βάση δεδομένων της εφαρμογής, εμφανίζεται σχετικό μήνυμα, μέσω Warning ShinyAlert, που τον ενημερώνει ότι για την ενέργεια δημιουργίας μοντέλου μηχανικής μάθησης, απαιτείται ένα τουλάχιστον αποθηκευμένο Dataset. Για το σκοπό αυτό, το κουμπί "**Navigate to File Uploader**", ανακατευθύνει τον χρήστη στο εργαλείο "**File Uploader**" (Εικόνα 45).



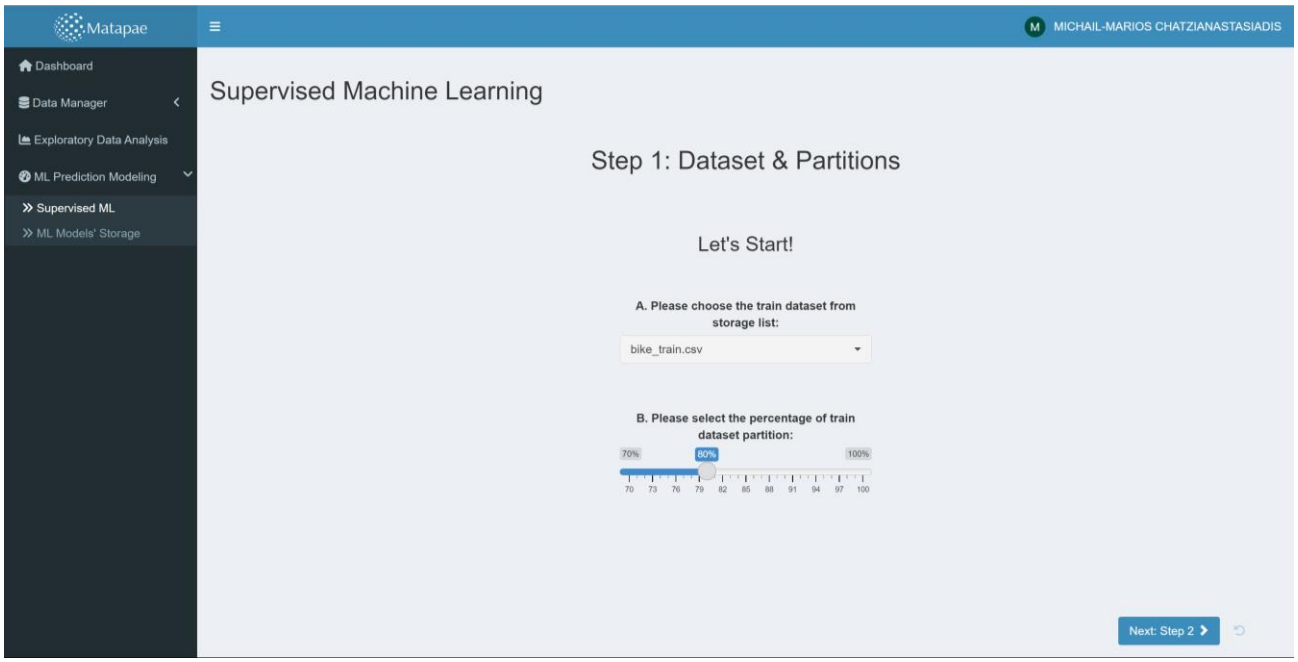
Εικόνα 45: Warning ShinyAlert - Δεν Υπάρχει Αποθηκευμένο Αρχείο

Μόλις ολοκληρωθεί η διαδικασία αποθήκευσης του πρώτου Dataset και ο χρήστης επιστρέψει στην επιλογή **Supervised ML**, ξεκινάει η διαδικασία παραμετροποίησης για την εκπαίδευσης του μοντέλου εποπτευόμενης μηχανικής μάθησης. Πρόκειται για μια διαδικασία τεσσάρων βημάτων που είναι τα εξής:

- Step 1 – Dataset and Partitions
- Step 2 – Prediction Type: Regression or Classification
- Step 3 – Dependent and Independent Variables
- Step 4 – Cross Validation and Hyperparameter Tuning

4.8.1.1 Step 1 – Dataset and Partitions

Σύμφωνα με το στιγμιότυπο της Εικόνα 46, στο πρώτο βήμα της παραμετροποίησης για την εκπαίδευση ενός μοντέλου εποπτευόμενης μηχανικής μάθησης (Supervised ML Model), ο χρήστης έχει τη δυνατότητα να επιλέξει από το Dropdown Menu (πεδίο A) το Dataset βάσει του οποίου θα εκπαιδευτεί το μοντέλο. Επισημαίνεται ότι το πεδίο A έχει προεπιλεγμένο το πρώτο κατά σειρά αποθηκευμένο στη βάση δεδομένων Dataset. Επιπλέον, στο πεδίο B, έχει τη δυνατότητα να ορίσει το ποσοστό του τυχαίου διαχωρισμού των παρατηρήσεων σε Training και Validation Datasets. Ειδικότερα, δίνεται η επιλογή του ποσοστού διαχωρισμού για το Training Dataset, στο εύρος των τιμών από 70% έως 100% των παρατηρήσεων με προεπιλεγμένη την τιμή 80%. Το υπόλοιπο ποσοστό συνεπάγεται με αυτό του Validation Dataset. Με το κουμπί **Next: Step 2**, ο χρήστης μπορεί να προχωρήσει στο επόμενο βήμα της διαδικασίας.



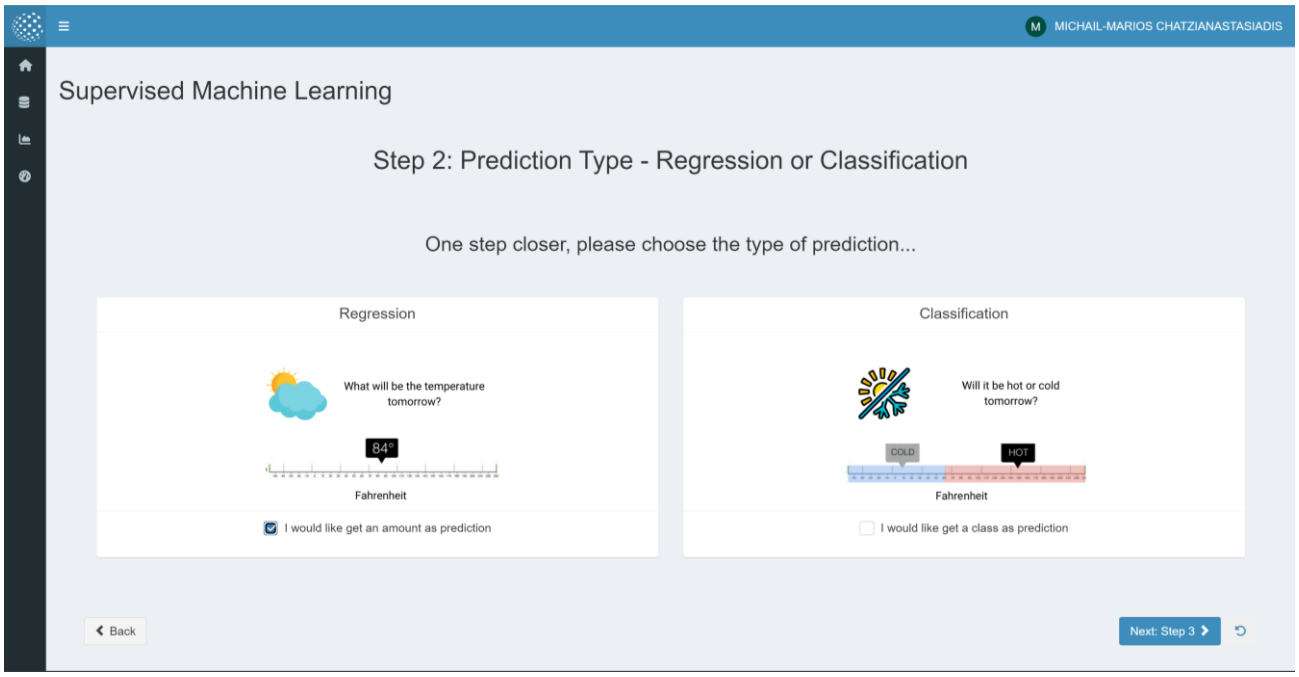
Εικόνα 46: Step 1 - Dataset & Partitions

4.8.1.2 Step 2 – Prediction Type: Regression or Classification

Στο δεύτερο βήμα της παραμετροποίησης (Εικόνα 47), η εφαρμογή καθοδηγεί τον χρήστη με δύο απλά εικονογραφημένα παραδείγματα, να επιλέξει τον επιθυμητό τύπο του μοντέλου πρόβλεψης που θα εκπαιδευτεί. Η επιλογή αυτή εξαρτάται τόσο από τα διαθέσιμα δεδομένα, όσο και από το είδος του προβλήματος που καλείται να επιλύσει.

Ειδικότερα, ο χρήστης θα επιλέξει από το περιβάλλον της εφαρμογής το Checkbox με την ένδειξη **“I would like get an amount as prediction”**, στην περίπτωση που καλείται να επιλύσει ένα πρόβλημα παλινδρόμησης (Regression), ενώ θα επιλέξει το Checkbox **“I would like get a class as prediction”,** για την επίλυση ενός προβλήματος κατηγοριοποίησης (Classification). Οι διαφορές τους, επεξηγούνται στον χρήστη μέσω ενδεικτικών ερωτημάτων και των αναμενόμενων τιμών πρόβλεψης τους. Πιο συγκεκριμένα, ένα Regression πρόβλημα εκφράζεται, για παράδειγμα, με έναν τύπο ερώτησης όπως: «Ποια θα είναι η θερμοκρασία αύριο;» και η πρόβλεψη αναμένεται να είναι μια συγκεκριμένη αριθμητική τιμή, όπως η θερμοκρασία (π.χ. 84 °F). Σε αντίθετη περίπτωση, ένα Classification πρόβλημα εκφράζεται με μια ερώτηση τύπου: «Θα κάνει ζέστη ή κρύο αύριο;» και η πρόβλεψη θα προκύψει από την πιθανότητα η αυριανή θερμοκρασία να ανήκει στην ομάδα των «ζεστών» ή «κρύων» θερμοκρασιών. Στο συγκεκριμένο παράδειγμα, ο χρήστης δηλώνει ότι επιθυμεί να λαμβάνει αριθμητικές αριθμητικές τιμές πρόβλεψης από το μοντέλο που θα εκπαιδευτεί, επομένως η διαδικασία θα προχωρήσει με την παραμετροποίηση του αλγορίθμου XGBoost για Regression προβλήματα.

Όταν ο χρήστης επιλέξει ένα από τα δύο Checkboxes, θα ενεργοποιηθεί το κουμπί **Next: Step 3**. Με κάθε αλλαγή της επιλογής του χρήστη, αποεπιλέγεται αυτόματα η προηγούμενη επιλογή αφού μόνο ένα Checkbox μπορεί να είναι επιλεγμένο κάθε φορά.



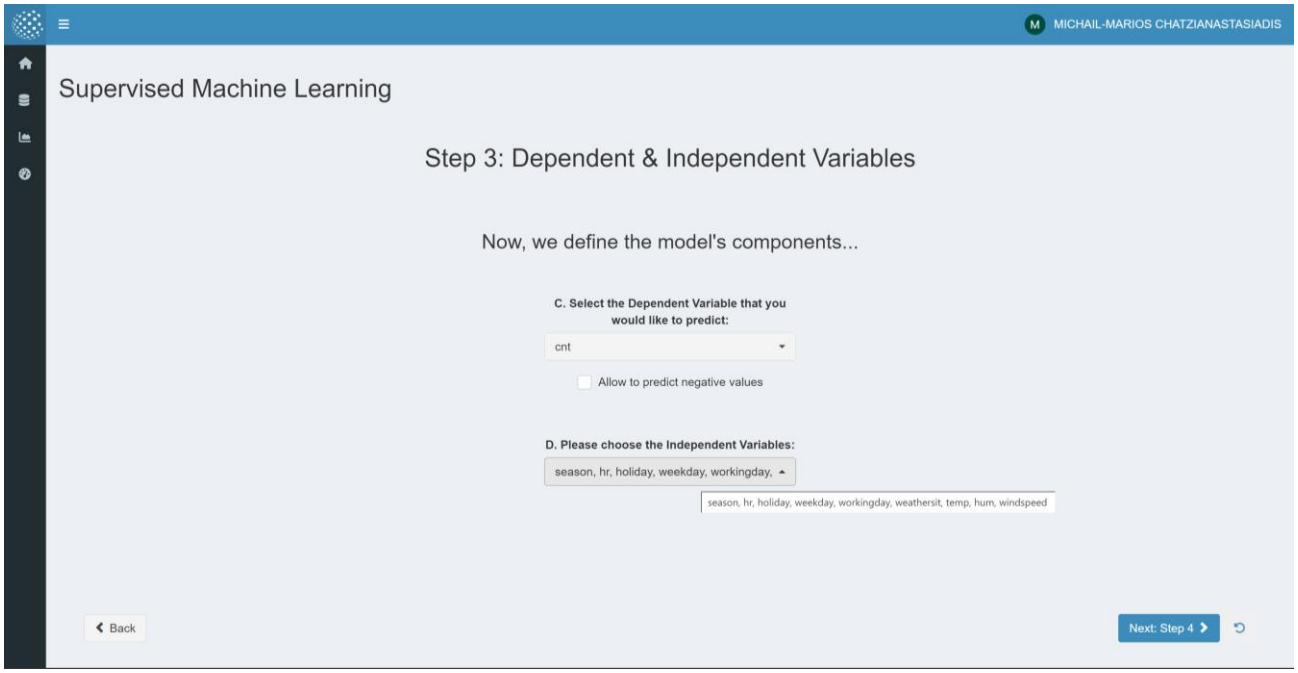
Εικόνα 47: Step 2 - Prediction Type: Regression or Classification Choice

4.8.1.3 Step 3 – Dependent and Independent Variables

Στο τρίτο βήμα της διαδικασίας (Εικόνα 48), γίνεται η επιλογή της εξαρτημένης μεταβλητής (Dependent Variable), καθώς και των ανεξάρτητων μεταβλητών (Features ή Independent Variables) βάσει των οποίων θα εκπαιδευτεί το μοντέλο εποπτευόμενης μηχανικής μάθησης, στα πεδία C και D αντίστοιχα.

Στο πεδίο C, ο χρήστης έχει τη δυνατότητα να επιλέξει μόνο μία μεταβλητή, ενώ στο πεδίο D μπορεί να γίνει επιλογή περισσότερων της μίας μεταβλητής. Σημειώνεται ότι η μεταβλητή που επιλέγεται ως εξαρτημένη, δεν μπορεί ταυτόχρονα να επιλεχθεί και ως ανεξάρτητη, γι' αυτό εξαιρείται αυτομάτως από την λίστα των ανεξάρτητων μεταβλητών του πεδίου D. Στην περίπτωση που στο Dataset υπάρχει το πεδίο ID, αφαιρείται και από τις δύο λίστες μεταβλητών, καθώς δεν μπορεί να προσδώσει χρήσιμη για το μοντέλο πληροφορία.

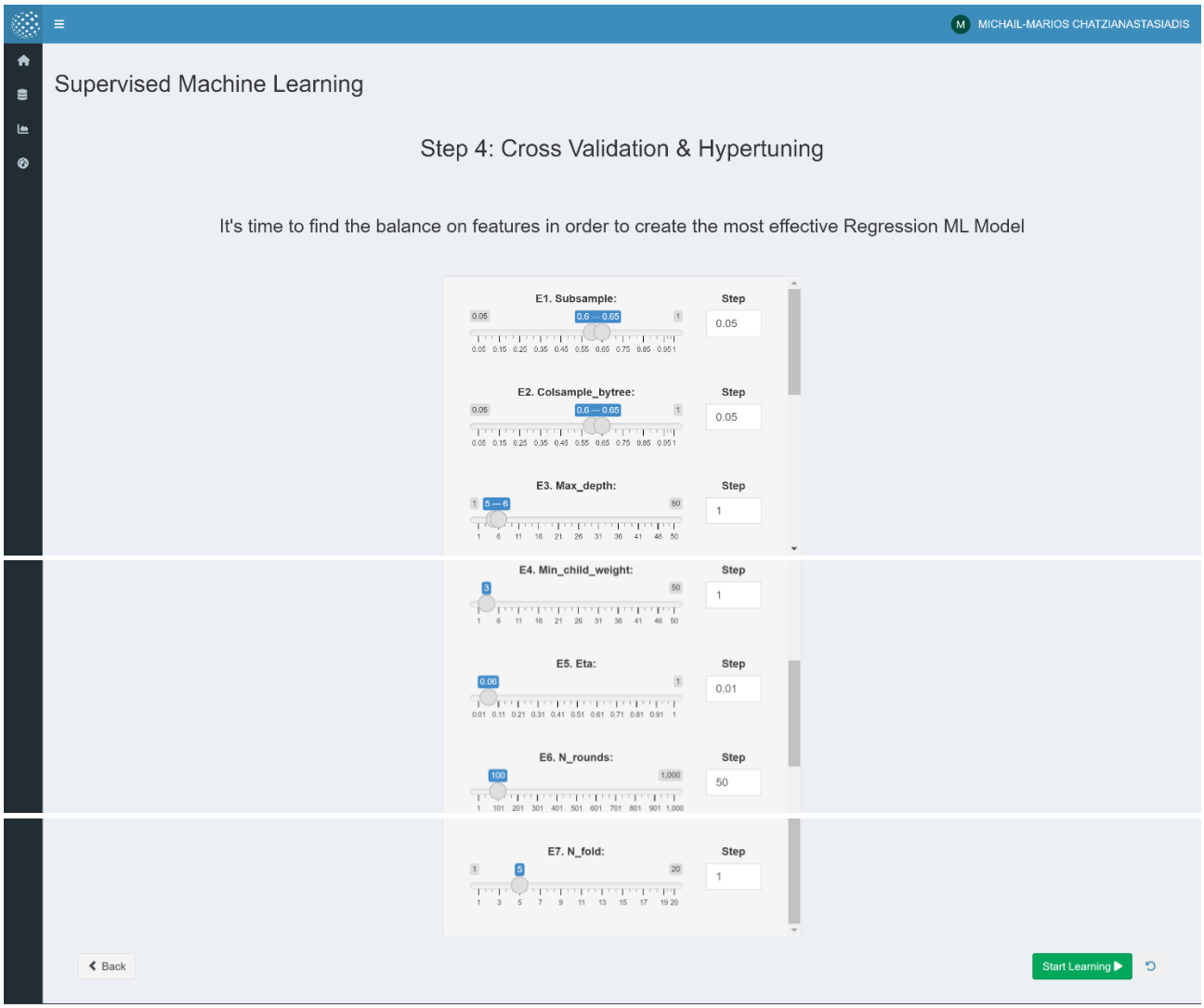
Επιπλέον, ο χρήστης έχει τη δυνατότητα να ορίσει εάν το μοντέλο θα προβλέψει αρνητικές τιμές, και αυτό επιτυγχάνεται με το Checkbox **"Allow to predict negative values"**. Με το κουμπί **Next: Step 4**, γίνεται η μετάβαση του χρήστη στο τέταρτο και τελευταίο βήμα της παραμετροποίησης.



Εικόνα 48: Step 3 - Dependent & Independent Variables

4.8.1.4 Step 4 – Cross Validation and Hyperparameter Tuning

Στο τελευταίο βήμα πριν από την έναρξη της εκπαίδευσης του μοντέλου (Εικόνα 49), καθορίζονται οι τιμές των παραμέτρων του αλγόριθμου εποπτευόμενης μηχανικής μάθησης eXtreme Gradient Boosting (XGBoost). Η εφαρμογή της μεθόδου **Hyperparameter Tuning**, η οποία υλοποιείται με τη συνάρτηση expand.grid(…)[11] του {Base} πακέτου της R, συμβάλλει στη βέλτιστη παραμετροποίηση του αλγορίθμου. Για το σκοπό αυτό, ζητείται από τον χρήστη να ρυθμίσει για κάθε μια από τις παραμέτρους, το εύρος τιμών και το βήμα ελέγχου, βάσει των οποίων θα δημιουργηθούν όλοι οι δυνατοί συνδυασμοί παραμετροποίησης. Στο τέλος της εκπαίδευσης, έπειτα από μια επαναληπτική διαδικασία δοκιμών εκπαίδευσης (Trial and Error), θα προκύψει το αποδοτικότερο μοντέλο πρόβλεψης από τον βέλτιστο συνδυασμό παραμέτρων, εξασφαλίζοντας τη μέγιστη δυνατή ελαχιστοποίηση των σφαλμάτων πρόβλεψης.

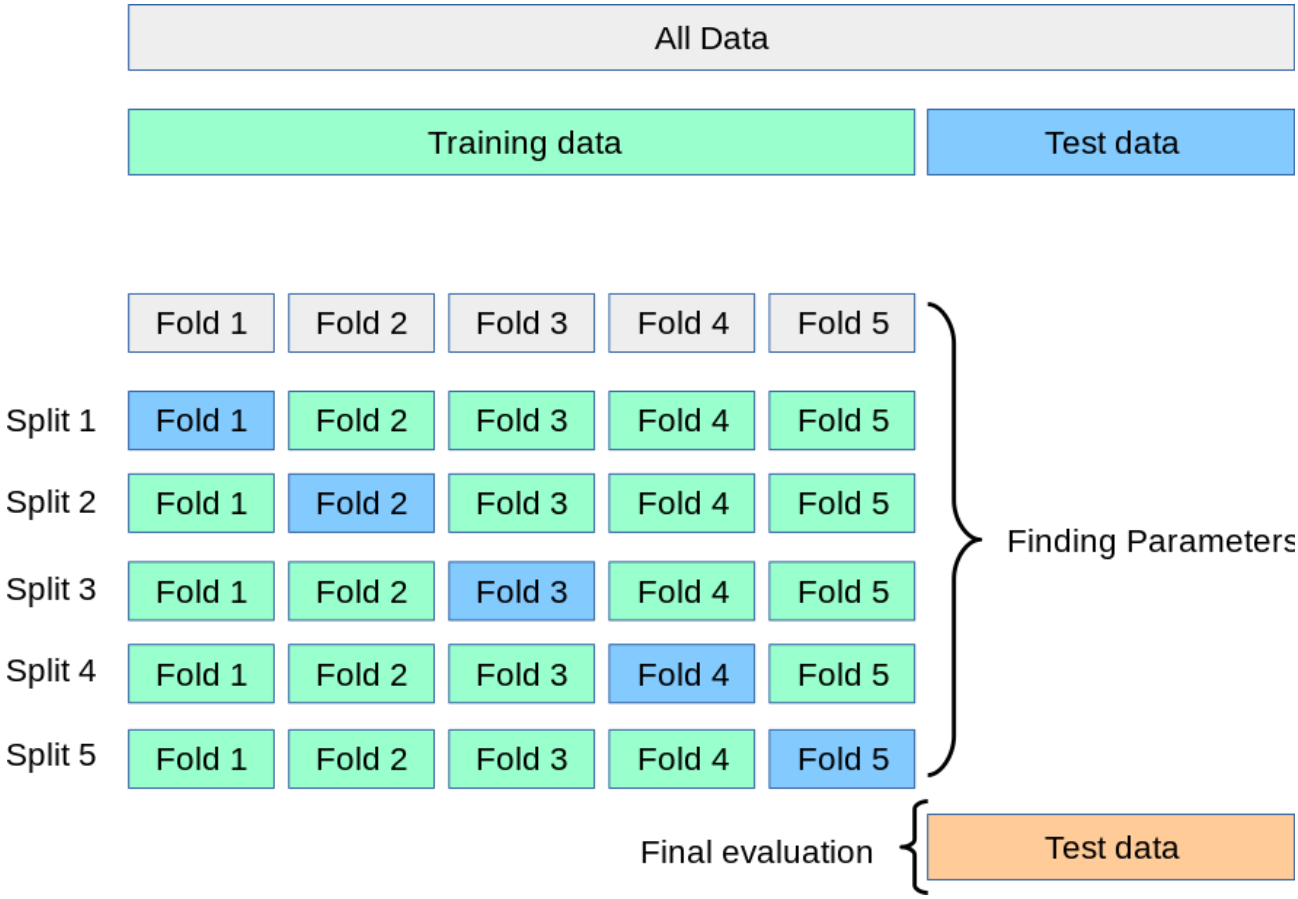


Εικόνα 49: Step 4 - Ρύθμιση παραμέτρων εκπαίδευσης αλγορίθμου XGBoost

[11] https://stat.ethz.ch/R-manual/R-devel/library/base/html/expand.grid.html

Ωστόσο ακόμα και μετά τη διαδικασία αξιολόγησης διαφορετικών συνδυασμών παραμετροποίησης (Hyperparameter Tuning), εξακολουθεί να υφίσταται ο κίνδυνος υπερβολικής προσαρμογής (Overfitting) του Test Dataset, με το οποίο δοκιμάζεται η αποδοτικότητα του μοντέλου πρόβλεψης. Αυτό συμβαίνει επειδή οι παράμετροι μπορούν να προσαρμόζονται μέχρι το μοντέλο εκτίμησης να αποδώσει τα μέγιστα, χωρίς όμως η απόδοση αυτή να μπορεί να γενικευτεί.

Η διαδικασία του **Cross-Validation**[12] **(CV)** αποτελεί τη λύση αυτού του προβλήματος, χρησιμοποιώντας το αρχικό Dataset το οποίο διαχωρίζεται σε k ίσα Data Subsets, όπως απεικονίζεται στην Εικόνα 50.



Εικόνα 50: k-Fold Cross-Validation

Πιο συγκριμένα, η διαδικασία Cross-Validation, που περιγράφεται παρακάτω, ακολουθείται για κάθε k-Fold ως εξής:

- Τα k-1 Subsets κάθε τυχαίου διαχωρισμού (Split) χρησιμοποιούνται ως Train Dataset για την εκμάθηση των μοντέλων.
- Ένα Subset από κάθε τυχαίο διαχωρισμό (Split) «φυλάσσεται» και χρησιμοποιείται ως Test Dataset για τις επιμέρους αξιολογήσεις των μοντέλων που προκύπτουν κατά τη διαδικασία. Η απόδοση του μοντέλου σε κάθε Split αξιολογείται βάσει του σχετικού μέσου τετραγωνικού σφάλματος (Root Mean Squared Error – RMSE).

[12] https://scikit-learn.org/stable/modules/cross_validation.html

- Η διαδικασία αυτή γίνεται επαναληπτικά με σκοπό την εύρεση εκείνου του συνδυασμού παραμέτρων που θα συνθέσει ένα αποδοτικό μοντέλο χωρίς Overfitting.
- Η τελική μέτρηση της απόδοσης (RMSE) των k-Folds κατά τη διαδικασία Cross-Validation είναι ο μέσος όρος των επιμέρους μετρήσεων της επαναληπτικής διαδικασίας επικύρωσης κάθε τυχαίου διαχωρισμού των δεδομένων (Split).

Αυτή η προσέγγιση μπορεί να χρειάζεται αρκετούς υπολογιστικούς πόρους, αλλά δεν «ξοδεύει» πολλά δεδομένα, όπως θα γινόταν σε μια συντηρητική προσέγγιση αυθαίρετου διαχωρισμού σε Train, Validation και Test Datasets. Αυτό το γεγονός είναι σημαντικό κυρίως σε διαδικασίες εκμάθησης με μικρά δείγματα δεδομένων. Στην εφαρμογή "Matapae" η διαδικασία Cross-Validation εφαρμόζεται με τη χρήση της συνάρτησης xgb.cv(…)[13] του {XGBoost} πακέτου της R.

Η έναρξη της εκπαίδευσης του μοντέλου γίνεται με το κουμπί **Start Learning**, μετά από τη κατάλληλη ρύθμιση των παραμέτρων[14] στα αντίστοιχα πεδία της εφαρμογής (Εικόνα 49):

- **E1 – Subsample:** Εκφράζει την αναλογία των παρατηρήσεων εκπαίδευσης που θα επιλεγούν τυχαία σε κάθε επανάληψη. Για παράδειγμα, εάν η ρύθμιση της παραμέτρου είναι 0.5, σημαίνει ότι ο αλγόριθμος XGBoost θα κάνει τυχαία δειγματοληψία στα μισά από τα Training Data πριν από την ανάπτυξη των δέντρων ώστε να αποτραπεί η υπερπροσαρμογή (Overfitting). Οι χαμηλές τιμές δημιουργούν ένα πιο συντηρητικό μοντέλο και αποτρέπουν το Overfitting του ML μοντέλου στο σύνολο των δεδομένων εκπαίδευσης. Αντιθέτως, με μικρές τιμές ρύθμισης πιθανώς να προκύψει υποπροσαρμογή (Underfitting). Το εύρος των τιμών είναι (0, 1]. Η ελάχιστη τιμή ρύθμισης που επιτρέπεται από την εφαρμογή είναι 0.05.
- **E2 – Colsample_bytree:** Δηλώνει την αναλογία του υποδείγματος των μεταβλητών που θα επιλεγούν τυχαία κατά την κατασκευή του δέντρου και πραγματοποιείται μια φορά για κάθε δέντρο που κατασκευάζεται. Η εξ ορισμού (by default) τιμή της παραμέτρου είναι 1 και σημαίνει ότι θα χρησιμοποιηθούν όλες οι μεταβλητές. Το εύρος των τιμών είναι (0, 1]. Η ελάχιστη τιμή ρύθμισης που επιτρέπεται από την εφαρμογή είναι 0.05.
- **E3 – Max_depth:** Αποτελεί το μέγιστο βάθος του δέντρου απόφασης. Δέχεται αριθμητικές τιμές στο εύρος [0, ∞]. Για την εύρεση της βέλτιστης τιμής, συνηθίζεται η ρύθμιση της παραμέτρου μεταξύ των τιμών 2 έως 10. Οι υψηλές τιμές μπορούν να οδηγήσουν σε Overfitting, λόγω της παροχής μεγάλης ποσότητας πληροφορίας από τα δεδομένα εκπαίδευσης στο μοντέλο, ενώ οι χαμηλές τιμές της παραμέτρου παρέχουν λιγότερες λεπτομέρειες στο μοντέλο με κίνδυνο να παρουσιαστεί πρόβλημα στην απόδοση του (Underfitting). Η ρύθμιση με την τιμή 0, υποδηλώνει ότι δεν καθορίζεται συγκεκριμένο όριο στο βάθος του δέντρου που θα προκύψει.
- **E4 – Min_child_weight:** Εκφράζει το ελάχιστο άθροισμα βάρους κάθε παρατήρησης που απαιτείται σε έναν κόμβο-παιδί. Εάν το βήμα διχοτόμησης του δέντρου έχει ως αποτέλεσμα έναν κόμβο-φύλλο με άθροισμα βάρους μικρότερο από το καθορισμένο βάρος (Min_child_weight), τότε η διαδικασία θα προχωρήσει με τη δημιουργία επιπλέον

[13] https://search.r-project.org/CRAN/refmans/xgboost/html/xgb.cv.html
[14] https://xgboost.readthedocs.io/en/latest/parameter.html

διχοτομήσεων. Όσο μεγαλύτερο είναι το Min_child_weight, τόσο πιο συντηρητικός θα είναι ο αλγόριθμος. Το εύρος των τιμών που λαμβάνει είναι [0, ∞].

- **E5 – Eta (Learning_rate):** Δηλώνει τον βαθμό συρρίκνωσης του μεγέθους του βήματος ώστε να αποφευχθεί το Overfitting. Έπειτα από κάθε Boosting Step, λαμβάνονται άμεσα τα βάρη των νέων Features, τα οποία μειώνονται από τον καθορισμό της παραμέτρου Eta έτσι ώστε να γίνει η διαδικασία Boosting πιο συντηρητική. Το εύρος των τιμών είναι [0, 1].
- **E6 – N_rounds:** Δηλώνει τον μέγιστο αριθμό Boosting επαναλήψεων. Δηλαδή τον μέγιστο αριθμό δέντρων που μπορεί να απαρτίζουν το μοντέλο. Ένας μεγάλος αριθμός δέντρων μπορεί να επιφέρει Overfitting. Οι υψηλές τιμές του N_rounds, ωθούν το Eta σε χαμηλές τιμές, ενώ οι μικρές τιμές N_rounds, απαιτούν αυξημένες τιμές της παραμέτρου Eta. Το εύρος των τιμών που επιτρέπονται για ρύθμιση στην εφαρμογή είναι [1, 1000] με βήμα μεγαλύτερο του 1.
- **E7 – N_fold:** Πρόκειται για μια παράμετρο της διαδικασίας k-Fold Cross-Validation και δηλώνει τον αριθμό των υποσυνόλων επικύρωσης που θέλουμε να δημιουργήσουμε. Το εύρος των τιμών που επιτρέπονται για ρύθμιση στην εφαρμογή είναι [1, 20] με βήμα μεγαλύτερο του 1.

Αναφορικά με τα δεδομένα της εκπαίδευσης, μετατρέπονται σε δομή DMatrix που παρέχεται από το {XGBoost} πακέτου της R μέσω της συνάρτησης xgb.DMatrix(…)[15]. Πρόκειται για μια εσωτερική δομή δεδομένων που χρησιμοποιείται από τον αλγόριθμο XGBoost, η οποία βελτιστοποιεί τόσο την απόδοση της μνήμης όσο και την ταχύτητα της εκπαίδευσης.

[15] https://search.r-project.org/CRAN/refmans/xgboost/html/xgb.DMatrix.html

4.8.1.4.1 R Κώδικας για την Εκπαίδευση XGBoost Regression Models

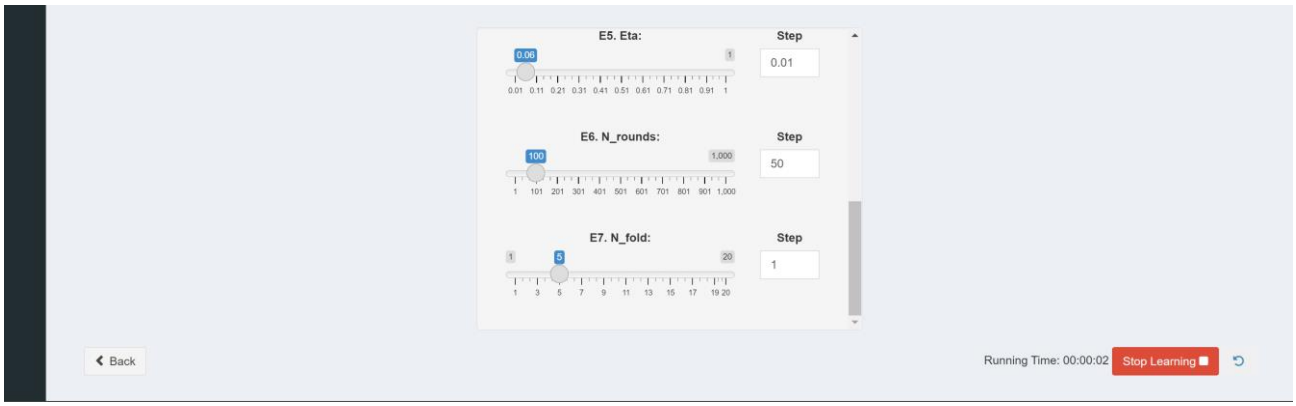
Κατά την εντολή έναρξης της εκπαίδευσης (**Start Learning**), καλείται η συνάρτηση xgb_gs_cv_regression(…) για τις περιπτώσεις που αφορούν Regression Models.

```
xgb_gs_cv_regression <- function (train_data_x,
                  train_label_y,
                  subsample_choice,
                  colsample_bytree_choice,
                  max_depth_choice,
                  min_child_weight_choice,
                  eta_choice,
                  n_rounds_choice,
                  n_fold_choice) {

 searchGridSubCol <- expand.grid(subsample = subsample_choice,
                  colsample_bytree = colsample_bytree_choice,
                  max_depth = max_depth_choice,
                  min_child_weight = min_child_weight_choice,
                  eta = eta_choice,
                  n_rounds = n_rounds_choice,
                  n_fold = n_fold_choice #cv parameter)

 rmseErrorsHyperparameters <- apply (searchGridSubCol,
                                     1,
                    function(parameterList){

    #Extract Parameters to test
    currentSubsampleRate <- parameterList[["subsample"]]
    currentColsampleRate <- parameterList[["colsample_bytree"]]
    currentDepth <- parameterList[["max_depth"]]
    currentEta <- parameterList[["eta"]]
    currentMinChildWeight <- parameterList[["min_child_weight"]]
    currentNRounds <- parameterList[["n_rounds"]]
    currentNFold <- parameterList[["n_fold"]]


    xgboostModelCV <-
         xgboost::xgb.cv (objective = "reg:squarederror", #xgb
               data = xgboost::xgb.DMatrix(data = train_data_x,
                                           label = train_label_y),
            booster = "gbtree", #xgb parameter
            showsd = TRUE, #xgb parameter
            verbose = TRUE, #xgb print statistics
            print_every_n = 10, #k-folds cv param
            early_stopping_rounds = 10, #k-folds
            eval_metric = "rmse", #xgb parameter
            "nrounds" = currentNRounds, #k-folds
            "nfold" = currentNFold, #k-folds cv
            "max_depth" = currentDepth,
            "eta" = currentEta,
            "subsample" = currentSubsampleRate,
                        "colsample_bytree" = currentColsampleRate,
                       "min_child_weight" = currentMinChildWeight)

    xgb_cv_xvalidationScores <- xgboostModelCV$evaluation_log

    #best score
    test_rmse <- tail(xgb_cv_xvalidationScores$test_rmse_mean, 1)
    train_rmse <- tail(xgb_cv_xvalidationScores$train_rmse_mean,1)
```

```
    gs_results_output <- c(test_rmse,
                           train_rmse,
                           currentSubsampleRate,
                currentColsampleRate,
                           currentDepth,
                           currentEta,
                currentMinChildWeight,
                           currentNRounds,
                           currentNFold)

    return(gs_results_output)

   }

 )

 gs_results_varnames <- c("TestRMSE",
                          "TrainRMSE",
                          "SubSampRate",
                          "ColSampRate",
                          "Depth",
                          "eta",
                          "MinChildWeight",
                          "nrounds",
                          "nfold")

 t_rmseErrorsHyperparameters <- as.data.frame(t(rmseErrorsHyperparameters))
 names(t_rmseErrorsHyperparameters) <- gs_results_varnames

 t_rmse_min <-
t_rmseErrorsHyperparameters[which.min(t_rmseErrorsHyperparameters$TestRMSE),]

 xgb_model_train <-
         xgboost::xgboost(objective = "reg:squarederror", #xgb param
             data = xgboost::xgb.DMatrix(data = train_data_x,
                                                 label = train_label_y),
             booster = "gbtree", #xgb parameter
             showsd = TRUE, #xgb parameter
             verbose = TRUE, #xgb print statistics
             print_every_n = 10, #k-folds cv parameter
             early_stopping_rounds = 10, #k-folds cv
             eval_metric = "rmse", #xgb parameter
             "nrounds" = t_rmse_min$nrounds, #k-folds
             "nfold" = t_rmse_min$nfold, #k-folds cv
             "max_depth" = t_rmse_min$Depth,
             "eta" = t_rmse_min$eta,
             "subsample" = t_rmse_min$SubSampRate,
                        "colsample_bytree" = t_rmse_min$ColSampRate,
             "min_child_weight" = t_rmse_min$MinChildWeight)

 return(xgb_model_train)

}
```

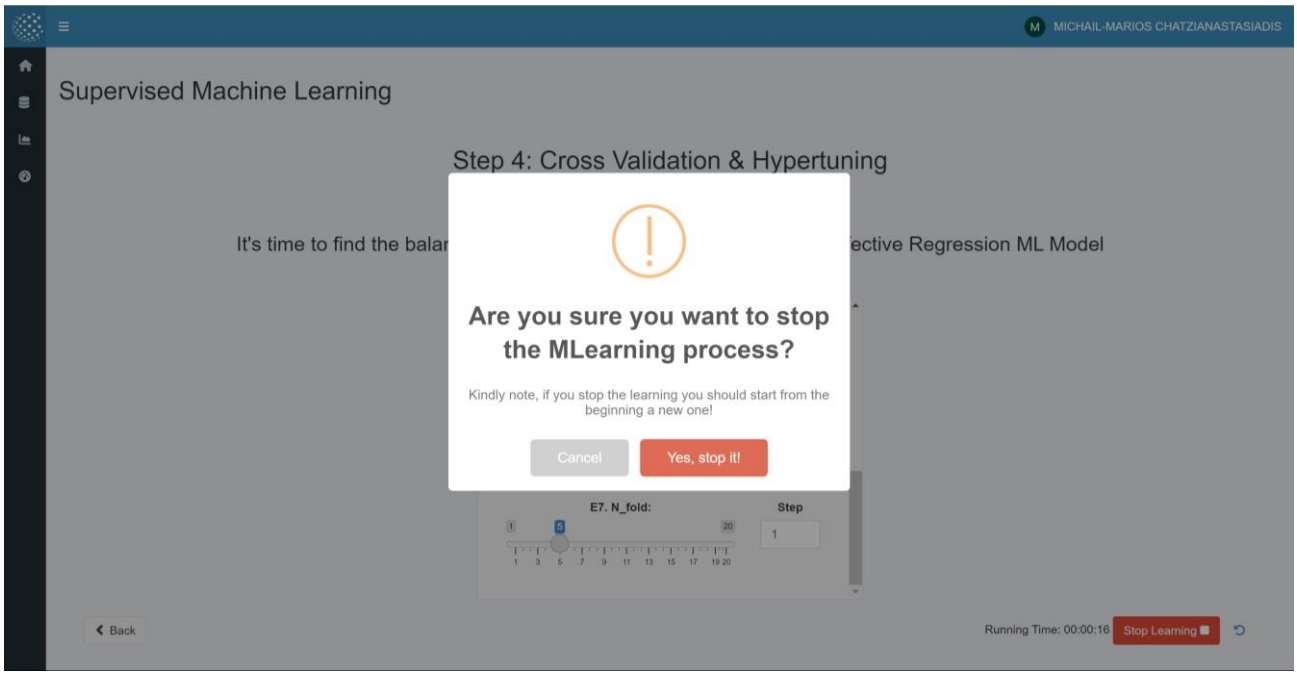
4.8.2 Learning Process and Results

Καθώς έχει ήδη ολοκληρωθεί η διαδικασία παραμετροποίησης του αλγορίθμου XGBoost, με το «πάτημα» του **Start Learning** (Εικόνα 49) σηματοδοτείται η έναρξη της εκπαίδευσης του προβλεπτικού μοντέλου μηχανικής μάθησης. Καθ' όλη τη διάρκεια της εκπαίδευσης, ένα χρονόμετρο κάτω δεξιά της οθόνης ενημερώνει τον χρήστη για τον τρέχων χρόνο εκτέλεσης (**Running Time**). Επίσης, δίνεται η δυνατότητα διακοπής της διαδικασίας με το κουμπί **Stop Learning** (Εικόνα 51).



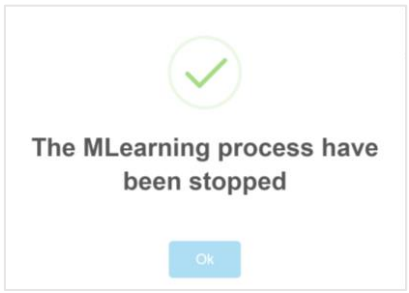
Εικόνα 51: Running Time/Stop Learning

Εάν ο χρήστης «πατήσει» το κουμπί **Stop Learning**, ένα Warning ShinyAlert τον ενημερώνει για την ενέργεια διακοπής που πρόκειται να πραγματοποιηθεί, ζητώντας την επιβεβαίωση του με το κουμπί **"Yes, stop it!"**. Στην περίπτωση που ο χρήστης δεν επιθυμεί να διακόψει την εκπαίδευση του μοντέλου πρόβλεψης, τότε μπορεί να «πατήσει» το κουμπί **"Cancel"** (Εικόνα 52).



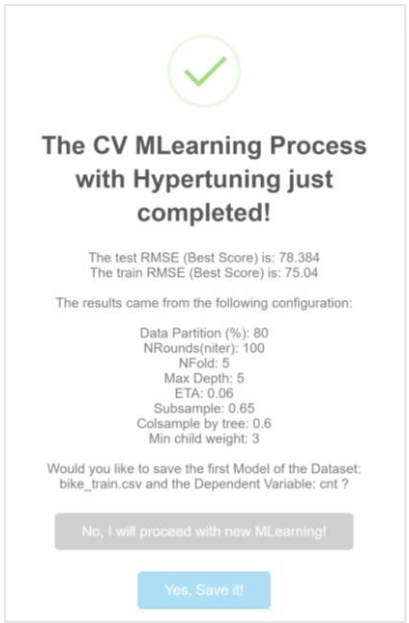
Εικόνα 52: Warning ShinyAlert - Διακοπή Διαδικασίας Εκπαίδευσης Μοντέλου

Το Success ShinyAlert, της Εικόνα 53, ενημερώνει τον χρήστη για την επιτυχή διακοπή της εκπαίδευσης του μοντέλου, αμέσως μετά την έγκριση που έδωσε για την ενέργεια αυτή με το κουμπί **"Yes, stop it!"** (Εικόνα 52).



Εικόνα 53: Success ShinyAlert - Επιτυχής Διακοπή της Διαδικασίας Εκπαίδευσης

Στην περίπτωση που δεν υπάρχει αποθηκευμένο ML Supervised Regression Model για τη μεταβλητή **"cnt"** του **"bike_train.csv"**, το Success ShinyAlert, της Εικόνα 54, ενημερώνει για την ολοκλήρωση της διαδικασίας εκπαίδευσης με το λεκτικό **"The CV MLearning Process with Hypertuning just completed"**. Επιπλέον, αναφέρονται τα αποτελέσματα της εκπαίδευσης και οι παράμετροι που χρησιμοποιήθηκαν για να προκύψει το βέλτιστο μοντέλο από τη διαδικασία Hyperparameter Tuning, από τα οποία θα αξιολογηθεί η αποδοτικότητα του μοντέλου βάσει των δεικτών του σχετικού μέσου τετραγωνικού σφάλματος (Root Mean Squared Error – RMSE).



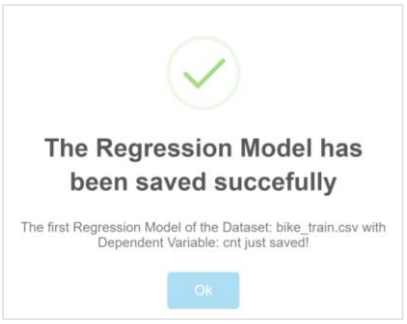
Εικόνα 54: Success ShinyAlert - Ολοκλήρωση Εκπαίδευσης

Πιο αναλυτικά, σε αυτό το σημείο ο χρήστης πρέπει να αποφασίσει εάν θα προχωρήσει με την αποθήκευση του παραγόμενου μοντέλου λαμβάνοντας υπόψιν τα εξής:

- **Train RMSE:** Δείχνει την μέση απόκλιση της τιμής πρόβλεψης από τις πραγματικές τιμές ενοικιάσεων ποδηλάτων, το οποίο προέκυψε από τις παρατηρήσεις του Train Dataset που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου.

- **Test RMSE:** Εκφράζει την απόδοση του μοντέλου στις παρατηρήσεις του Validation Dataset που δεν συμμετείχαν στη διαδικασία εκπαίδευσης, δηλαδή την απόδοση σε άγνωστες για το μοντέλο παρατηρήσεις. Η τιμή του Test RMSE αναμένεται να είναι μεγαλύτερη από την τιμή του Train RMSE.

Για την ολοκλήρωση της διαδικασίας με αποθήκευση, ο χρήστης πρέπει να επιλέξει το κουμπί **"Yes, Save it!"**, διαφορετικά με το κουμπί **"No, I will proceed with new MLearning!"** μπορεί να προβεί σε νέα προσπάθεια εκπαίδευσης με διαφορετική παραμετροποίηση. Η επιτυχής αποθήκευση του πρώτου Regression Model, επιβεβαιώνεται με την εμφάνιση του Success ShinyAlert της Εικόνα 55.



Εικόνα 55: Success ShinyAlert - Επιτυχής Αποθήκευση ML Supervised Regression Model

Όταν υπάρχει ένα ήδη αποθηκευμένο ML Supervised Regression Model για τη μεταβλητή **"cnt"** του **"bike_train.csv"** και ο χρήστης επιθυμεί να προβεί σε μία νέα προσπάθεια πρόβλεψης, τότε:

A. Εάν το **Test RMSE (New)** του ML Supervised Regression Model της μεταβλητής πρόβλεψης είναι **μεγαλύτερο ή ίσο** από το **Test RMSE (Old)** του ήδη αποθηκευμένου μοντέλου στη βάση δεδομένων, τότε εμφανίζεται Warning ShinyAlert που ενημερώνει ότι δεν προέκυψε καλύτερο μοντέλο πρόβλεψης με το μήνυμα **"The Regression Model hasn't been improved"** (Εικόνα 56). Ο χρήστης μπορεί να αποθηκεύσει το ML μοντέλο πατώντας το κουμπί **"Yes, save the new model"** (λαμβάνοντας έπειτα την ενημέρωση επιτυχούς αποθήκευσης μέσω Success ShinyAlert) ή να αγνοήσει αυτή την ενέργεια πατώντα το κουμπί **"Ignore, keep the old one"**.

B. Εάν το **Test RMSE (New)** του ML Regression Model της μεταβλητής πρόβλεψης είναι **μικρότερο** από το **Test RMSE (Old)** του ήδη αποθηκευμένου μοντέλου στη βάση δεδομένων, τότε εμφανίζεται Success ShinyAlert που ενημερώνει για την επιτυχή αποθήκευση του νέου βελτιωμένου ML Regression Model με το μήνυμα **"New Best Score!"** (Εικόνα 57).

Εικόνα 56: Warning ShinyAlert - Διαχείριση μη Βελτιστοποιημένου ML Regression Model

Εικόνα 57: Success ShinyAlert - Ενημέρωση Επιτυχούς Αποθήκευσης Νέου Βελτιστοποιημένου ML Regression Model

Έπειτα από κάθε επιτυχημένη αποθήκευση μοντέλου, η εφαρμογή μεταβαίνει στην ενότητα **ML Models' Storage** για τον έλεγχο των αποτελεσμάτων και την πρόβλεψη σε νέα δεδομένα. Επισημαίνεται ότι στη βάση δεδομένων, επιτρέπεται η αποθήκευση ενός μόνο μοντέλου ανά μεταβλητή και είδος πρόβλεψης.

4.8.2.1 R Κώδικας για την Αποθήκευση ML Model στη Βάση Δεδομένων

Για την αποθήκευση των ML Models στη βάση δεδομένων καλείται η συνάρτηση saveModeltoDB(…):

```
saveModeltoDB <- function(sub,
             modeltype_id,
             dataset_id,
             dependent_var,
             independent_vars,
             data_partition,
             test_rmse,
             test_original_partition,
             test_prediction) {

 print ("**************saveModeltoDB***************")

 dbcon <- RMySQL::dbConnect(
  RMySQL::MySQL(),
  dbname = options()$mysql$databaseName,
  host = options()$mysql$host,
  port = options()$mysql$port,
  user = options()$mysql$user,
  password = options()$mysql$password
 )

 independent_vars <- toString(independent_vars)

 test_original_partition <- toString(test_original_partition)

 test_prediction <- toString(test_prediction)

 insert_data = data.frame(sub,
              modeltype_id,
              dataset_id,
              dependent_var,
              independent_vars,
              data_partition,
              test_rmse,
              test_original_partition,
              test_prediction)

 print(insert_data)

 RMySQL::dbWriteTable(dbcon, "savedmodels",
            insert_data,
            field.types = c(sub = "VARCHAR(255)",
                    modeltype_id = "INTEGER",
                    dataset_id = "INTEGER",
                    dependent_var = "VARCHAR(255)",
                    independent_vars = "TEXT",
                    data_partition = "INTEGER",
                    test_rmse = "DOUBLE",
                    test_original_partition = "LONGTEXT",
                    test_prediction = "LONGTEXT",
                    uploaded_model = "LONGBLOB"),
            append = TRUE,
            #overwrite = TRUE,
            row.names = FALSE)
```

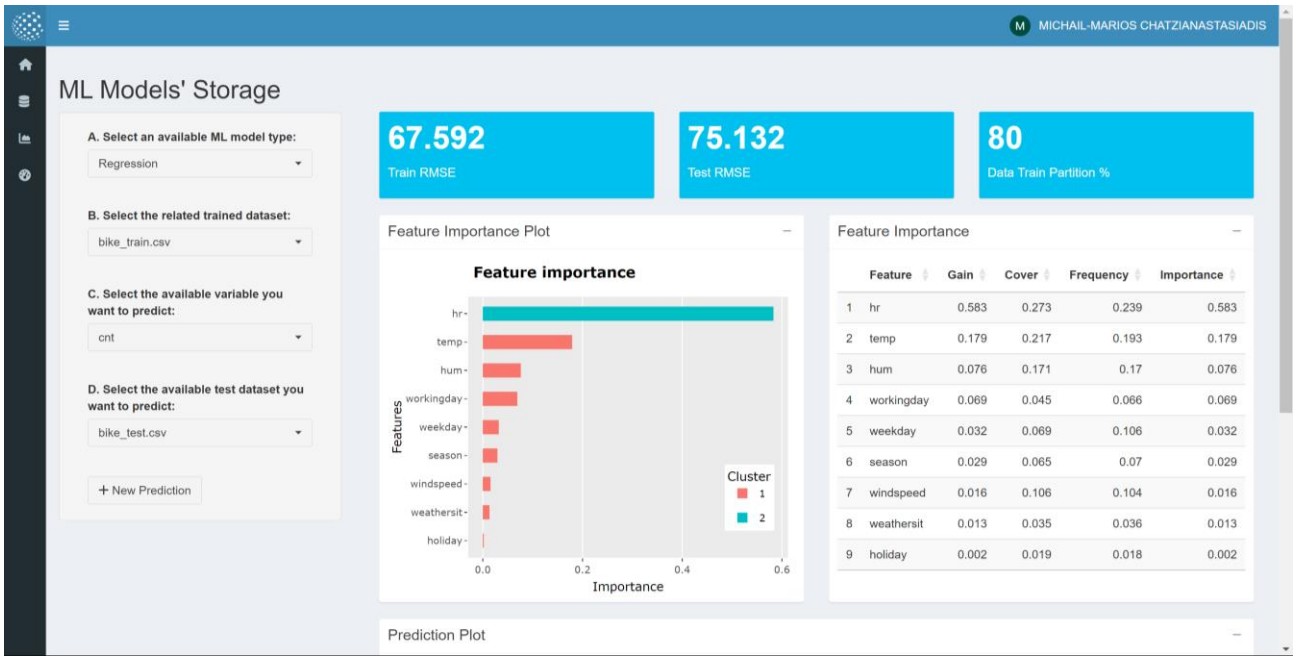
```
 uploaded_file_query <-
      paste0("C:\\ProgramData\\MySQL\\MySQL Server 8.0\\Uploads\\xgb.model")

 query <- dbplyr::build_sql("UPDATE savedmodels
                SET dependent_var = ",dependent_var,",
                independent_vars = ",independent_vars,",
                data_partition = ",data_partition,",
                test_rmse = ",test_rmse,",
                test_original_partition = ",test_original_partition,",
                test_prediction = ",test_prediction,",
                uploaded_model = LOAD_FILE(",uploaded_file_query,")
                WHERE sub = ",sub,
                " AND dataset_id =",dataset_id,
                " AND modeltype_id = ",modeltype_id,
                " AND dependent_var = ",dependent_var,
                " AND independent_vars = ",independent_vars,
                                ";",
                con = dbcon)

 rs <- DBI::dbSendQuery(dbcon, query)

 DBI::dbClearResult(rs)

 RMySQL::dbDisconnect(dbcon)

 shinyjs::show(selector = "a[data-value='ML_models_storage']")

}
```

4.8.3 ML Models' Storage

Στην ενότητα **ML Models' Storage** (Εικόνα 58), ο χρήστης μπορεί να αναζητήσει και να ανακτήσει τα διαθέσιμα ML Supervised Models, που είναι αποθηκευμένα στη βάση δεδομένων, από το Sidebar στα αριστερά της οθόνης, ακολουθώντας τα παρακάτω βήματα:

1. Στο πεδίο A, θα πρέπει να επιλέξει τον τύπο του μοντέλου, δηλαδή εάν το μοντέλο που επιθυμεί να ανακτήσει είναι Regression ή Classification.
2. Στο πεδίο B, το οποίο ενημερώνεται αυτόματα βάσει της επιλογής του πεδίου A, ανακτώνται από τη βάση δεδομένων όλα τα διαθέσιμα Datasets του επιλεγμένου τύπου εκπαίδευσης, για τα οποία υπάρχει τουλάχιστον ένα εκπαιδευμένο μοντέλο. Ο χρήστης επιλέγει το Dataset που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου που αναζητά.
3. Αντίστοιχα στο πεδίο C «φορτώνουν» όλες οι διαθέσιμες μεταβλητές πρόβλεψης για τις οποίες υπάρχουν αντιστοιχισμένα μοντέλα στη βάση δεδομένων για το επιλεγμένο Dataset.

Εφόσον έχουν συμπληρωθεί τα πεδία A, B, C, το επιθυμητό ML Supervised Model ανακτάται από τη βάση δεδομένων της εφαρμογής, ενώ στοιχεία της εκπαίδευσης του, όπως τα **Train RMSE**, **Test RMSE**, **Data Train Partition %**, **Feature Importance Plot**, **Feature Importance** και **Prediction Plot**, παρουσιάζονται λεπτομερώς στο δεξί μέρος της οθόνης.



Εικόνα 58: ML Models' Storage - Model Details and Feature Importance

4.8.3.1 Feature Importance

Η ανάλυση Feature Importance[16] υποδεικνύει τη συνεισφορά κάθε μεταβλητής ή αλλιώς γνωρίσματος (Variable/Feature) στην πρόβλεψη ενός μοντέλου μηχανικής μάθησης, η οποία παρέχεται από το πακέτο {XGBoost} μέσω των συναρτήσεων xgb.importance(…)[17] και xgb.ggplot.importance(…)[18] για την αντίστοιχη οπτικοποίηση. Η βαθμολόγηση του κάθε γνωρίσματος (Feature) προκύπτει από τον υπολογισμό τριών επιμέρους σκορ σημαντικότητας:

1. **Gain:** Δηλώνει τη σχετική συνεισφορά ενός Feature στο μοντέλο και υπολογίζεται λαμβάνοντας υπόψιν τη συνεισφορά του εκάστοτε Feature σε κάθε δέντρο του μοντέλου. Μια υψηλή βαθμολογία υποδηλώνει ότι το Feature είναι περισσότερο σημαντικό στην πρόβλεψη του μοντέλου.
2. **Cover:** Εκφράζει τις σχετικές παρατηρήσεις που συνδέονται με έναν παράγοντα πρόβλεψης. Για παράδειγμα, το Feature X χρησιμοποιείται για τον προσδιορισμό του τερματικού κόμβου για δεκαπέντε παρατηρήσεις στο δέντρο A και για δέκα παρατηρήσεις στον δέντρο B. Επομένως, οι παρατηρήσεις που σχετίζονται με το Feature X είναι συνολικά εικοσιπέντε και το Cover υπολογίζεται ως ο λόγος των εικοσιπέντε παρατηρήσεων προς το άθροισμα των παρατηρήσεων όλων των Features.
3. **Frequency:** Αναφέρεται στη σχετική συχνότητα εμφάνισης ενός Feature στο σύνολο των δέντρων απόφασης. Για παράδειγμα, εάν από το Feature X προκύψει μια διακλάδωση του δέντρου A και δύο στο δέντρο B, τότε οι συνολικές εμφανίσεις του θα είναι τρεις και η σχετική συχνότητα θα είναι ο λόγος τους με τις συνολικές εμφανίσεις όλων των Features.

Συνήθως, οι κατηγορικές μεταβλητές (Categorical Variables) και ιδιαίτερα αυτές που έχουν χαμηλό Cardinality, δηλαδή μικρό αριθμό μοναδικών τιμών, παρουσιάζουν χαμηλό Frequency καθώς χρησιμοποιούνται σπάνια σε κάθε δέντρο απόφασης. Αντιθέτως, οι συνεχείς μεταβλητές (Continuous Variables) παρουσιάζουν υψηλό Cardinality, γεγονός που αυξάνει τις πιθανότητες εμφάνισης τους στο μοντέλο (υψηλό Frequency). Επομένως, αποφεύγεται η χρήση του Frequency Score για την εξέταση της σημαντικότητας των Features.

Αναφορικά με το Gain, αποτελεί το πιο πολύτιμο σκορ αξιολόγησης των Features σε ένα μοντέλο XGBoost και χρησιμοποιείται ως μετρητής σφάλματος, κατατάσσοντας τα Features σε φθίνουσα σειρά από το πιο σημαντικό στο πιο ασήμαντο. Επιπλέον, τα Clusters, στη γραφική απεικόνιση του Feature Importance (Εικόνα 58), παρουσιάζουν τις ομάδες μεταβλητών με παρόμοιο βαθμό σημαντικότητας για το προβλεπτικό μοντέλο.

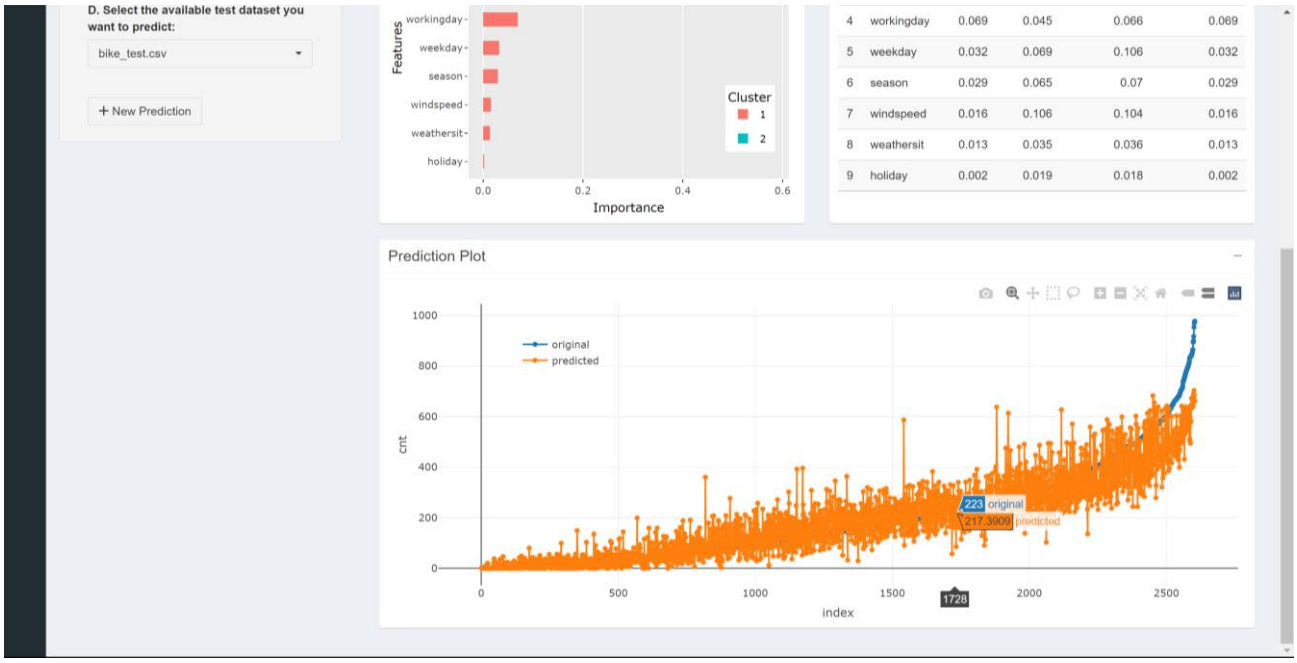
[16] https://www.r-bloggers.com/2019/10/explaining-predictions-boosted-trees-post-hoc-analysis-xgboost/
[17] https://www.rdocumentation.org/packages/xgboost/versions/0.6.4.1/topics/xgb.importance
[18] https://www.rdocumentation.org/packages/xgboost/versions/0.6.4.1/topics/xgb.ggplot.importance

4.8.3.2 Prediction Plot

Σε κάθε διαδικασία εκμάθησης, το αρχικό Dataset διαχωρίζεται σε Train και Validation Partitions. Η επιλογή γίνεται από τον χρήστη κατά το πρώτο βήμα της παραμετροποίησης (4.8.1.1). Η εκπαίδευση του επιλεγμένου μοντέλου έγινε με Data Train Partition 80% του αρχικού Dataset (Εικόνα 58). Το υπόλοιπο 20% των παρατηρήσεων, δηλαδή το Data Validation Partition, δεν συμμετείχε στην εκμάθηση, παρά μόνο στην επικύρωση των προβλέψεων του παραγόμενου μοντέλου.

Ειδικότερα, η αποδοτικότητα του μοντέλου πρόβλεψης αξιολογείται με τις παρατηρήσεις που δεν συμμετείχαν στη διαδικασία εκπαίδευσης. Ουσιαστικά, μετά την ολοκλήρωση της εκπαίδευση του μοντέλου με τις παρατηρήσεις του Data Train Partition, εισάγονται σε αυτό μόνο τα Features του Data Validation Partition, χωρίς τη μεταβλητή πρόβλεψης, «ζητώντας» από το μοντέλο να προβλέψει τις τιμές της, παρόλο που είναι γνωστές. Με τον τρόπο αυτό, θα κριθεί το εύρος της απόκλισης των προβλέψεων από τις πραγματικές τιμές. Σκοπός είναι να διαπιστωθεί εάν από την εκπαίδευση έχει δημιουργηθεί ένα μοντέλο, ικανό να κάνει έγκυρες προβλέψεις σε νέα δεδομένα. Πρόκειται για τη διαδικασία από την οποία προκύπτει το Test RMSE.

Έπειτα με το **Prediction Plot** (Εικόνα 59), ο χρήστης έχει τη δυνατότητα να εξετάσει μια προς μια την απόκλιση των τιμών πρόβλεψης (Predicted) από την πραγματική τους τιμή (Original), βάσει του μοναδικού αναγνωριστικού (ID) των παρατηρήσεων. Για παράδειγμα, η πρόβλεψη των ενοικιάσεων ποδηλάτων (cnt), για την παρατήρηση με μοναδικό αναγνωριστικό (ID) 1728 είναι 217.3909, ενώ η πραγματική της τιμή είναι 223 ενοικιάσεις. Επομένως, για τη συγκεκριμένη παρατήρηση η πρόβλεψη έγινε με μικρή απόκλιση, καθώς το ML Supervised Regression Model προέβλεψε μόνο 5.6 ενοικιάσεις ποδηλάτων λιγότερες από τις πραγματικές.



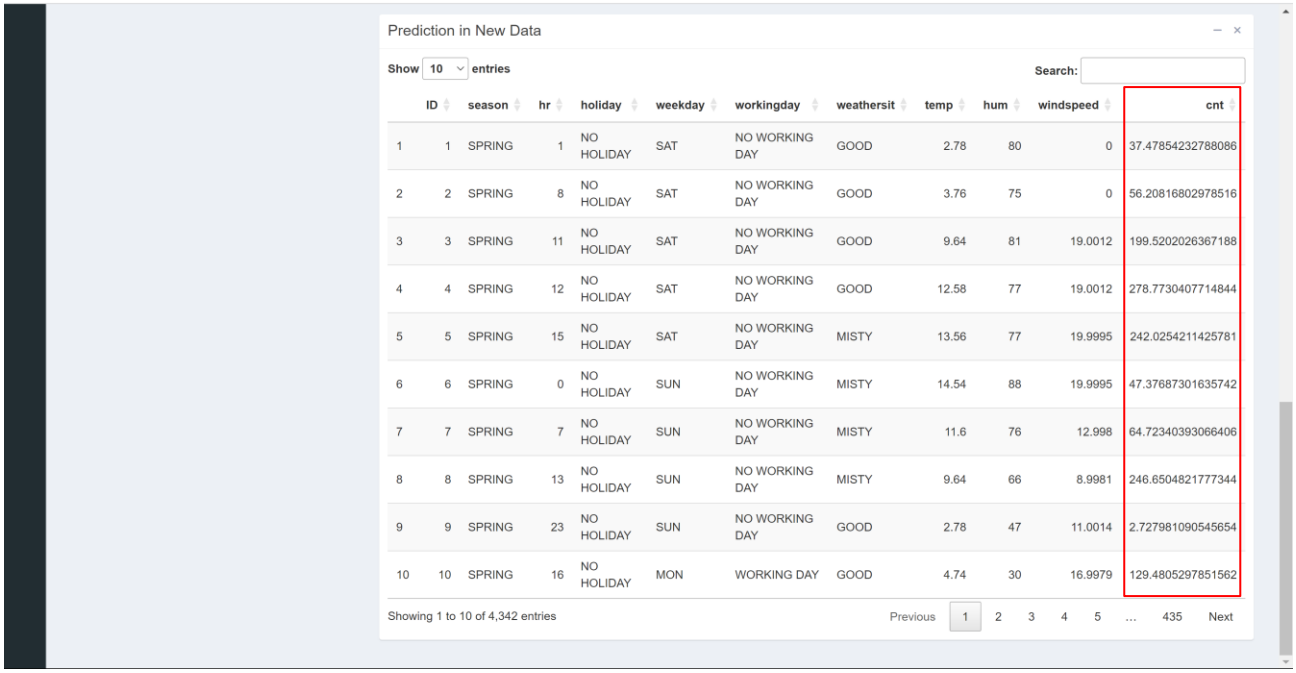
Εικόνα 59: ML Models' Storage - Prediction Plot

4.8.3.3 Prediction in New Data

Η πρόβλεψη σε νέα δεδομένα αποτελεί μια ακόμα σημαντική λειτουργικότητα που είναι διαθέσιμη στην ενότητα ML Models' Storage. Ο χρήστης έχει τη δυνατότητα να εισάγει στην εφαρμογή ένα νέο (Test) Dataset, το οποίο όμως θα πρέπει να περιλαμβάνει τιμές για τα ίδια Features με αυτά που έχουν ήδη χρησιμοποιηθεί στην εκπαίδευση του προβλεπτικού μοντέλου μηχανικής μάθησης. Το μοντέλο, για κάθε μια παρατήρηση του Dataset, θα επιστρέψει την τιμή της πρόβλεψης σε μια νέα στήλη. Η εισαγωγή και η αποθήκευση του Dataset θα πρέπει να γίνει από την ενότητα File Uploader (4.6.1).

Σύμφωνα με το παράδειγμα της Εικόνα 58, στο πεδίο D, έχει επιλεγεί το Dataset "bike_test.csv". Η επιλογή του Dataset στο πεδίο D δεν μπορεί να συμπίπτει με αυτή του πεδίου B που έχει χρησιμοποιηθεί στην εκπαίδευση του μοντέλου. Για τον λόγο αυτό, αφαιρείται αυτόματα ως επιλογή από το πεδίο D. Όταν ο χρήστης επιλέξει το επιθυμητό Dataset στο πεδίο D, το κουμπί **"New Prediction"** ενεργοποιείται. Στην περίπτωση που το επιλεγμένο Dataset δεν περιέχει τα Features που έχουν χρησιμοποιηθεί στην εκπαίδευση του μοντέλου, τότε το κουμπί παραμένει απενεργοποιημένο.

Μόλις ο χρήστης πατήσει το κουμπί **"New Prediction"** (Εικόνα 58), εμφανίζεται ένας πίνακας στο UI όπου η τελευταία στήλη (cnt) περιλαμβάνει τις προβλέψεις της μεταβλητής πρόβλεψης για κάθε μια δοθείσα σειρά Features, όπως παρουσιάζεται στην Εικόνα 60. Η πρόβλεψη πραγματοποιείται με τη χρήση του ML Supervised Regression Model που έχει ήδη εκπαιδευτεί και αποθηκευτεί στη βάση δεδομένων.



Εικόνα 60: ML Models' Storage - Predictions in New Dataset

4.8.3.4 R Κώδικας για την Ανάκτηση Αποθηκευμένου ML Model

Η ανάκτηση ενός .model αρχείου από τη βάση δεδομένων της εφαρμογής γίνεται με τη συνάρτηση getModelFile(…), ο κώδικας της οποίας παρατίθεται παρακάτω:

```
getModelFile <-function(sub,
                        dataset_id,
                        modeltype_id,
                        dependent_var){

        dbcon <- DBI::dbConnect(
         RMySQL::MySQL(),
         dbname = options()$mysql$databaseName,
         host = options()$mysql$host,
         port = options()$mysql$port,
         user = options()$mysql$user,
         password = options()$mysql$password
        )

        query <-
       paste0("SELECT uploaded_model FROM savedmodels WHERE sub = '",sub,
              "' AND dataset_id = '",dataset_id,
              "' AND modeltype_id = '",modeltype_id,
              "' AND dependent_var = '",dependent_var,
              "' into dumpfile 'C:/ProgramData/MySQL/MySQL
                   Server 8.0/Uploads/xgb.model';")

        request <- DBI::dbGetQuery(dbcon, query)

        DBI::dbDisconnect(dbcon)

        return(request)

}
```

Βιβλιογραφία

[1] C. A. R. A. R. Pinheiro and M. Patetta, Introduction to Statistical and Machine Learning Methods for Data Science, SAS Institute, 2021.

[2] J. B. Ramsey, "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 31, pp. 350-371, July 1969.

[3] L. Igual and S. Seguí, Introduction to Data Science, Springer Cham, 2017.

[4] M. L. Brodie, "What Is Data Science?," in *Applied Data Science*, Springer, 2019, pp. 101-130.

[5] M. Franke, "An Introduction to Data Analysis," 12 October 2021. [Online]. Available: https://michael-franke.github.io/intro-data-analysis/Chap-02-01-data-kinds-of-data.html.

[6] C. Shah, A Hands-On Introduction to Data Science, Cambridge University Press, 2020.

[7] S. Rajput, "Big Data 3 V's and 5 V's," 24 June 2020. [Online]. Available: https://medium.com/analytics-vidhya/big-data-3-vs-and-5-v-s-c1cae2a6d311.

[8] M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi and J. M. Elmirghani, "Big data analytics for wireless and wired network design: A survey," in *Computer Networks*, Elsevier, 2018, pp. 180-199.

[9] P. Ukhalkar, R. Phursule, D. Gadekar and N. Sable, "Business Intelligence and Analytics: Challenges and Opportunities," *International Journal of Advanced Science and Technology,* vol. 29, pp. 2669-2676, August 2020.

[10] N. Kyriakou, E. Loukis and M. M. Chatzianastasiadis, "Enterprise Systems, ICT Capabilities and Business Analytics Adoption – An Empirical Investigation," in *Information Systems*, Springer International Publishing, 2020, pp. 433-448.

[11] Z. Sun, H. Zou and K. D. Strang, "Big Data Analytics as a Service for Business Intelligence," in *Open and Big Data Management and Innovation*, Springer International Publishing, 2015, pp. 200-211.

[12] D. Nam, J. Lee and H. Lee, "Business analytics adoption process: An innovation diffusion perspective," *International Journal of Information Management,* vol. 49, pp. 411-423, December 2019.

[13] A. S. Aydiner, E. Tatoglu, E. Bayraktar, S. Zaim and D. Delen, "Business analytics and firm performance: The mediating role of business process performance," *Journal of Business Research,* vol. 96, pp. 228-237, March 2019.

[14] M. Attaran and S. Attaran, "Opportunities and Challenges of Implementing Predictive Analytics for Competitive Advantage," *International Journal of Business Intelligence Research.,* vol. 9, July 2018.

[15] W. Raghupathi and V. Raghupathi, "Contemporary Business Analytics: An Overview," *Challenges in Business Intelligence,* vol. 6, no. 8, p. 86, 4 August 2021.

[16] C. Cote, "4 TYPES OF DATA ANALYTICS TO IMPROVE DECISION-MAKING," 13 October 2021. [Online]. Available: https://online.hbs.edu/blog/post/types-of-data-analysis.

[17] T. H. Davenport and J. Kim, Keeping up with the quants: Your guide to understanding and using analytics, Harvard Business Review Press, 2013.

[18] H. Li, "Which machine learning algorithm should I use?," 9 December 2020. [Online]. Available: https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/.

[19] R. Khandelwal, "Supervised, Unsupervised, and Reinforcement Learning," 20 July 2022. [Online]. Available: https://arshren.medium.com/supervised-unsupervised-and-reinforcement-learning-245b59709f68.

[20] R. J. Hyndman and G. Athanasopoulos, "Time series patterns," in *Forecasting: principles and practice*, Melbourne, Australia, OTexts, 2021.

[21] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining*, ACM, 2016.

[22] P. Banerjee, "XGBoost + k-fold CV + Feature Importance," 8 December 2020. [Online]. Available: https://www.kaggle.com/code/prashant111/xgboost-k-fold-cv-feature-importance#8.-References-.

[23] P. Banerjee, "Bagging vs Boosting," 30 June 2020. [Online]. Available: https://www.kaggle.com/code/prashant111/bagging-vs-boosting/notebook.

[24] V. Morde, "XGBoost Algorithm: Long May She Reign!," 8 April 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

[25] A. Nagpal, "L1 and L2 Regularization Methods," 13 October 2017. [Online]. Available: https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c.

[26] K. Wakefield, "A guide to the types of machine learning algorithms and their applications," [Online]. Available: https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html.

[27] "What is the Apriori algorithm?," [Online]. Available: https://www.educative.io/answers/what-is-the-apriori-algorithm.

[28] A. S, "Association Rule Mining Including Apriori Algorithm," 2 January 2022. [Online]. Available: https://medium.com/analytics-vidhya/association-rule-mining-including-apriori-algorithm-8c0f9888e125.

[29] M. Banoula, "What Is Apriori Algorithm in Data Mining: Everything You Need to Know," 20 February 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/apriori-algorithm-in-data-mining.

[30] S. Kumar, "5 Anomaly Detection Algorithms every Data Scientist should know," 13 December 2021. [Online]. Available: https://towardsdatascience.com/5-anomaly-detection-algorithms-every-data-scientist-should-know-b36c3605ea16.

[31] K. Mandal, "How Anomaly Detection using Isolation Forest can play a role in Telecom?," 13 October 2020. [Online]. Available:

https://www.whatech.com/og/telecommunications/blog/668543-how-anomaly-detection-using-isolation-forest-can-play-a-role-in-telecom.

[32] X. He, K. Zhao and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems,* p. 106622, 5 January 2021.

[33] M.-A. Zöller and M. F. Huber, "Benchmark and Survey of Automated," *Journal of Artificial Intelligence Research,* vol. 70, pp. 409-474, December 2019.

[34] Q. Yao, M. Wang, Y. Chen, W. Dai, Y.-F. Li, W.-W. Tu, Q. Yang και Y. Yu, «Taking the Human out of Learning Applications: A Survey on Automated Machine Learning,» 8 November 2018.

[35] J. Reilly, "How Does Automated Machine Learning Work?," 10 June 2021. [Online]. Available: https://www.akkio.com/post/how-does-automated-machine-learning-work.

[36] «The Machine Learning Life Cycle Explained,» October 2022. [Ηλεκτρονικό]. Available: https://www.datacamp.com/blog/machine-learning-lifecycle-explained.

[37] «What is Shiny (R)?,» [Ηλεκτρονικό]. Available: https://www.dominodatalab.com/data-science-dictionary/shiny-in-r.

[38] K. S. Htoon, "Log Transformation: Purpose and Interpretation," 29 February 2020. [Online]. Available: https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9.