Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Ανάλυση και Πρόβλεψη Χρονικών Καθυστερήσεων σε Δρομολόγια Αεροπορικών Πτήσεων με Χρήση Τεχνικών Μηχανικής Μάθησης

Διπλωματικη Εργασια

του

ΠΑΠΑΓΙΑΝΝΟΠΟΥΛΟΥ ΕΥΑΓΓΕΛΟΥ

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Ανάλυση και Πρόβλεψη Χρονικών Καθυστερήσεων σε Δρομολόγια Αεροπορικών Πτήσεων με Χρήση Τεχνικών Μηχανικής Μάθησης

Διπλωματικη Εργασια

του

ΠΑΠΑΓΙΑΝΝΟΠΟΥΛΟΥ ΕΥΑΓΓΕΛΟΥ

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22 Δεκεμβρίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Θεοδώρα Βαρβαρίγου	Εμμανουήλ Βαρβαρίγος	Συμεών Παπαβασιλείου
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2022

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Copyright © – All rights reserved. Με την επιφύλαξη παντός δικαιώματος.
Ευάγγελος Παπαγιαννόπουλος, 2022.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα.

Το περιεχόμενο αυτής της εργασίας δεν απηχεί απαραίτητα τις απόψεις του Τμήματος, του Επιβλέποντα, ή της επιτροπής που την ενέκρινε.

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Ευάγγελος
Παπαγιαννόπουλος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

22 Δεκεμβρίου 2022

Περίληψη

Οι καθυστερήσεις πτήσεων είναι ένα πολύ σημαντικό θέμα για τις αεροπορικές μεταφορές σε όλο τον κόσμο λόγω των συσχετιζόμενων οικονομικών ζημιών που υφίσταται συνεχώς η αεροπορική βιομηχανία.

Αυτές οι καθυστερήσεις δεν προκαλούν μόνο ταλαιπωρία στις αεροπορικές εταιρείες, αλλά και στους επιβάτες. Με τον αυξημένο χρόνο ταξιδιού έρχεται και η αύξηση των εξόδων που σχετίζονται με το φαγητό και τη διαμονή και αυτό οδηγεί σε πρόσθετο άγχος μεταξύ των επιβατών. Οι αεροπορικές εταιρείες επίσης υποφέρουν από πρόσθετα κόστη όπως αυτά που σχετίζονται το πλήρωμα των αεροσκαφών, την αλλαγή αεροσκαφών και την αυξημένη κατανάλωση καυσίμου προσπαθώντας να μειωσουν τον συνολικό χρόνο πτήσης. Τα αμφότερα αμαυρώνουν τη φήμη των αεροπορικών εταιρειών και συχνά οδηγούν σε απώλεια ζήτησης από τους επιβάτες.

Στόχος της παρούσας διπλωματικής εργασίας είναι αρχικά η διερευνητική ανάλυση δεδομένων πτήσεων για την εξόρυξη χρήσιμων πληροφοριών που σχετιζονται με τις καθυστερήσεις πτήσεων και στη συνέχεια η ανάπτυξη μοντέλων μηχανικής και βαθιάς μάθησης για την πρόβλεψη των καθυστερήσεων.

Πιο συγκεκριμένα, η διερευνητική ανάλυση δεδομένων πτήσεων χωρίζεται σε τέσσερις φάσεις. Αρχικά εύρεση πηγής δεδομένων. Έπειτα ο καθαρισμός τους και στη συνέχεια η δημιουργία χαρακτηριστικών που θα βοηθήσουν στην διαδικασία της πρόβλεψης των καθηστερήσεων. Τέλος για την πρόβλεψη των καθυστερήσεων το οποίο θα αντιμετωπιστεί ως ένα πρόβλημα κατηγοριοποίησης θα εφαρμοστούν πρώτον τεχνικές επιβλεπόμενης μάθησης με αλγορίθμους όπως εκμάθηση συνόλου, συσκευασίας και δεύτερον βαθιά μάθηση με χρήση πολυεπιπεδού νευρωνικού δικτύου.

Λέξεις Κλειδιά

Κατηγοριοποίηση ή Ταξινόμηση, Μηχανική Μάθηση, Βαθιά Μάθηση, Νευρωνικά Δίκτυα, Διερευνητική Ανάλυση Δεδομένων, Αλγόριθμος Εκμάθησης Συνόλου, Αλγόριθμος Συσκευασίας, Επιβλεπόμενη Μάθηση

Abstract

Flight delays are a very important issue for air transportation around the world because of the associated financial losses that the aviation industry constantly suffers.

These delays not only cause inconvenience to airlines, but also to passengers. With increased travel time comes increased costs associated with food and accommodation and this leads to additional stress among passengers. Airlines also suffer from additional costs such as those related to crews, changing aircraft, increased fuel consumption trying to reduce the total flight time and many others that tarnish the reputation of the airlines and often lead to a loss of demand from passengers.

The objective of this thesis is firstly the exploratory data analysis (EDA) of flights to extract useful information related to flight delays and then the development of machine learning and deep learning models to predict delays.

More specifically, the exploratory analysis of flight data is divided into four phases. First find a data source. Then cleaning them up and then creating features that will help in the delay prediction process. Finally, for delay prediction which will be treated as a classification problem, supervised learning techniques with algorithms such as ensemble and bagging learning and secondly deep learning using multi-layer neural network (MLP) will be applied.

Keywords

Classification, Machine Learning, Deep Learning, Neural Networks, Exploratory Data Analysis, Ensemble Learning Algorithm, Bagging Algorithm, Supervised Learning

στους γονείς μου

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω την επιβλέπουσα καθηγήτρια Ε.Μ.Π. κα. Θεοδώρα Βαρβαρίγου για την ευκαιρία που μου προσέφερε να εκπονήσω τη συγκεκριμένη εργασία καθώς και για την εμπιστοσύνη που μου έδειξε. Παράλληλα, θα ήθελα να ευχαριστήσω τον συνεπιβλέποντα κ. Αναστάσιο Νικολακόπουλο για την εξαιρετική συνεργασία που είχαμε. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για την καθοδήγηση και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Αθήνα, Δεκέμβριος 2022

Ευάγγελος Παπαγιαννόπουλος

Περιεχόμενα

Περίληψη **1**

Abstract **3**

Ευχαριστίες **7**

Πρόλογος **15**

1 Εισαγωγή **17**

1.1 Αντικείμενο της διπλωματικής 17

1.2 Οργάνωση του τόμου 18

I Θεωρητικό Μέρος **19**

2 Θεωρητικό υπόβαθρο Μηχανικής Μάθησης **21**

2.1 Μηχανική Μάθηση 21

2.1.1 Ορισμός και μέθοδοι Μηχανικής Μάθησης 21

2.2 Επιβλεπόμενη Μάθηση 22

2.2.1 Δέντρα Απόφασης (Decision Trees) 23

2.3 Εκμάθηση Συνόλου (Ensemble learning) 24

2.3.1 Γιατί δουλεύουν οι τεχνικές εκμάθησης συνόλου· 24

2.4 Η Τεχνική της Ενίσχυσης (Boosting) 25

2.4.1 AdaBoost - Αλγόριθμος Προσαρμοσμένης Ενίσχυσης : 26

2.4.2 Gradient Boost : Ενίσχυση Κλίσης 26

2.4.3 XGBoost : Ακραία Ενίσχυση κλίσης 27

2.5 Η Τεχνική της Συσκευασίας (Bagging) 27

2.5.1 Τυχαία Δάση (RF - Random Forests) 28

2.6 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANN) 29

2.6.1 Perceptrons 30

2.6.2 Multi Layer Perceptrons - MLP 31

2.6.3 Συναρτήσεις Ενεργοποίησης (Activation Functions) 32

2.6.4 Εκπαίδευση Πολυστρωματικού Perceptron - MLP 34

2.6.5 Μετρικές Αξιολόγησης – Evaluation Metrics 34

3 Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA) 37
3.1 Τι είναι η EDA; . . . 37
3.1.1 Χαρακτηριστικά της EDA . . . 37
3.2 Περιγραφική στατιστική . . . 38

II Πρακτικό Μέρος 41

4 Διερευνητική Ανάλυση στο Σύνολο των Δεδομένων 43
4.1 Συλλογή Δεδομένων . . . 43
4.2 Περιγραφή Δεδομένων . . . 43
4.3 Εργαλεία Ανάλυσης . . . 45
4.4 Καθαρισμός Δεδομένων . . . 46
4.4.1 Συγχώνευση Δεδομένων . . . 46
4.5 Διερευνητική Ανάλυση Δεδομένων - (EDA) . . . 46
4.5.1 Univariate Analysis . . . 49
4.5.2 Bivariate Analysis . . . 53
4.6 Exploratory Data Analysis for weather data . . . 56
4.7 Τεχνικές Προεπεξεργασίας Δεδομένων για Modeling . . . 56
4.7.1 One hot Encoding . . . 56
4.7.2 Integer encoding . . . 57
4.8 Υπερδειγματοληψία και Υποδειγματοληψία . . . 57

5 Πειραματικά αποτελέσματα Μηχανικής Μάθησης 59
5.1 Βελτιστοποίηση Υπερ-Παραμέτρων . . . 59
5.2 Υπερπροσαρμογή (Overfitting) . . . 60
5.3 Αποτελέσματα . . . 60
5.3.1 Αποτελέσματα Decision Trees . . . 61
5.3.2 Αποτελέσματα Bagged Trees . . . 61
5.3.3 Αποτελέσματα Random Forest . . . 62
5.3.4 Αποτελέσματα ADAptive Boost . . . 64
5.3.5 Αποτελέσματα Gradient Boosted Trees . . . 65
5.3.6 Αποτελέσματα eXtreme gradient Boosting . . . 66
5.3.7 Αποτελέσματα Multi Layer Perceptron . . . 66
5.3.8 Γενικές παρατηρήσεις . . . 71

III Επίλογος 73

6 Επίλογος 75
6.1 Συμπεράσματα . . . 75
6.2 Μελλοντικές Επεκτάσεις . . . 75

Παραρτήματα 77

Α΄ Παραδείγματα Βιβλιογραφικών Αναφορών 79

Βιβλιογραφία 83

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια 85

Κατάλογος Σχημάτων

2.1 Παράδειγμα ενός RL δικτύου (Πηγή: [1]) 22
2.2 Παράδειγμα ενός Decision Tree (Πηγή:[2]) 23
2.3 Αναπαράσταση των τύπων μεθόδων συνόλου. Εδώ, το m αντιπροσωπεύει έναν αδύναμο μαθητή. d1, d2, d3, d4 είναι τα τυχαία δείγματα από τα δεδομένα D. d', d", d" είναι ενημερωμένα δεδομένα εκπαίδευσης με βάση τα αποτελέσματα από τον προηγούμενο αδύναμο μαθητή.(Πηγή:[3]) 24
2.4 Training Boosting Models(Πηγή:[4]) 25
2.5 Training of an AdaBoost classifier. The first classifier trains on unweighted data, then reweights the data for the next and so on to produce the final classifier. (Πηγή:[5]) 26
2.6 Algorithm 1. Friedman's Gradient Boost algorithm.(Πηγή:[6]) 27
2.7 Bagging (independent predictors) vs. Boosting (sequential predictors).(Πηγή:[7]) 28
2.8 The Random Forest Algorithm (Πηγή:[8]) 29
2.9 Random Forest Classifier (Πηγή:[9]) 29
2.10 Simple neural network (Πηγή:[10]) 30
2.11 παράδειγμα ενός Perceptron (Πηγή:[11]) 31
2.12 MLP με 2 κρυφά επίπεδα (Πηγή: [12]) 32
2.13 Logistic Curve Function (Πηγή: Wikipedia) 33
2.14 Hyperbolic Tangent Function (Πηγή: Wikipedia) 34
2.15 ReLU (Πηγή: [13]) 34
2.16 (Πηγή: [14]) 35

3.1 parts of EDA 38

4.1 Flights on time (status 0), Flights slightly delayed(status 1), Flights highly delayed status 2) 47
4.2 Correlation Matrix for all the features 48
4.3 Histograms for all the weather features 48
4.4 Percentage of delayed Flights 50
4.5 Delayed Flights per day 50
4.6 Delayed Flights per month 51
4.7 Origin Airports count plots 51
4.8 airlines piechart 52
4.9 airlines countplots 52
4.10 Most Popular Destinations and their Average Arrival Delays 53

4.11 How does the origin airport impact delays? Analysis of Airports and Airlines 54
4.12 Taxi in and Taxi out by Airlines 55
4.13 Arrival delays by Airline 55
4.14 Impact of Delays 55
4.15 Different kind of delays 56
4.16 One hot encoding example[15] 57

5.1 Classification report for Decision Tree Classifier without Departure Delay accounted 61
5.2 Classification report for Decision Tree Classifier with Departure Delay accounted 61
5.3 Classification report for Bagged Trees Classifier without Departure Delay . 61
5.4 Classification report for Bagged Trees Classifier with Departure Delay . . . 62
5.5 Classification report for Random Forest Classifier without Departure Delay 62
5.6 Classification report for Random Forest Classifier with Departure Delay . . 62
5.7 Feature Importance for Random Forrest Classifier without Departure Delay 63
5.8 Feature Importance for Random Forrest Classifier with Departure Delay . . 64
5.9 Classification report for ADAptive Boost Classifier without Departure Delay 64
5.10 Classification report for ADAptive Boost Classifier with Departure Delay . . 65
5.11 Cross validation score for ADAboost without Departure Delay 65
5.12 Cross validation score for ADAboost with Departure Delay 65
5.13 Classification report for Gradient Boosted Trees Classifier without Departure Delay 65
5.14 Classification report for Gradient Boosted Trees Classifier Departure Delay 65
5.15 Cross validation score for Gradient Boosted Trees without Departure Delay 66
5.16 Classification report for eXtreme Gradient Boosting Classifier without Departure Delay 66
5.17 Classification report for MLP model1 with Departure Delay 67
5.18 Training results for MLP model1 with Departure Delay 68
5.19 Confusion Matrix for MLP model1 with Departure Delay 68
5.20 Classification report for MLP model2 with Departure Delay 69
5.21 Training results for MLP model2 with Departure Delay 69
5.22 Confusion Matrix for MLP model2 with Departure Delay 70
5.23 Classification report for MLP model3 with Departure Delay 70
5.24 Training results for MLP model3 with Departure Delay 71
5.25 Confusion Matrix for MLP model3 with Departure Delay 71

Πρόλογος

Η συγκεκριμένη διπλωματική εργασία αναπτύχθηκε σε ερευνητικό πλαίσιο για να μελετήσουμε ποια είναι τα χαρακτηριστικά εκείνα και σε ποιον βαθμο επηρεάζουν την καθυστέρηση μιας πτήσης. Η ερευνα εκτέλέστηκε σε συνεργασία με διπλωματικούς ερευνητές και καθοδηγητές του ΕΜΠ του τμήματος της κας Βαρβαρίγου. Το εργαστήριο βρίσκεται στον τομέα Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής της Σχολής ΕΜΠ.

Κεφάλαιο 1

Εισαγωγή

Η αεροπορική βιομηχανία είναι σημαντικός τομέας της υποδομής της κάθε χώρας που προσφέρει ευκολία στην εγχώρια και διεθνή μεταφορά των ανθρώπων και σημαντικά οικονομικά οφέλη για την κοινωνία. Η βιωσιμότητα του συστήματος μεταφορών διαδραματίζει επίσης σημαντικό ρόλο στη μείωση της χρήσης ενέργειας και των εκπομπών στην ατμόσφαιρα [16].

Με την ανάπτυξη της κοινωνίας και τις ταξιδιωτικές απαιτήσεις, περισσότεροι άνθρωποι επιλέγουν τις πτήσεις ως τον προτιμότερο τρόπο μεταφοράς λόγω του χρόνου ταξιδιού, της ευκολίας ή του κόστους. Συνοπτικά, η βελτίωση της αξιοπιστίας των υπηρεσιών πτήσεων όχι μόνο θα εξοικονομούσε χρόνο στους πελάτες και θα βελτίωνε την ικανοποίηση των πελατών, αλλά θα οδηγούσε επίσης σε βιώσιμη ανάπτυξη και εξοικονόμηση πόρων, όπως κάυσιμα.

Ένα ενδιαφέρον στατιστικό στοιχείο σύμφωνα με στοιχεία από το Γραφείο Στατιστικών Μεταφορών των Ηνωμένων Πολιτειών είναι οτι περισσότερο από το 20% των πτήσεων στις ΗΠΑ καθυστέρησαν το 2018. Αυτές οι καθυστερήσεις έχουν σοβαρό οικονομικό αντίκτυπο που ισοδυναμεί με 40,7 δισεκατομμύρια δολάρια ετησίως.

Όλα τα παραπάνω καθιστούν επιτακτική την ανάγκη για ανάλυση και ακριβή πρόβλεψη καθυστερήσεων στην εναέρια κυκλοφορία.

1.1 Αντικείμενο της διπλωματικής

Στην παρούσα διπλωματική εργασία θα ασχοληθούμε με ακριβώς αυτά τα δύο μέρη. Τα δεδομένα πτήσεων θα τα αντλήσουμε απο το Γραφείο Στατιστικών Μεταφορών των Ηνωμένων Πολιτειών για το έτος 2019.

1. Το πρώτο μέρος αφορά την διερευνητική ανάλυση (EDA) του συνόλου δεδομένων. Σκοπός εδώ είναι να ερευνήσουμε σχέσεις μεταξύ μεταβλητών-χαρακτηριστικών του συνόλου των δεδομένων. Η ανάλυση είναι σημαντικό κομμάτι διότι οι καθυστερήσεις στα αεροδρόμια μπορεί να προκύψουν από τις διάφορες λειτουργίες αεροπορικών εταιρειών, τη συμφόρηση εναέριας κυκλοφορίας, τις καιρικές συνθήκες, τις πρωτοβουλίες οσον αφορά τη διαχείριση εναέριας κυκλοφορίας κ.λπ. Οι περισσότεροι από τους λόγους είναι στοχαστικά φαινόμενα που είναι δύσκολο να προβλεφθούν έγκαιρα και με ακρίβεια.

2. Το δεύτερο μέρος αφορά την πρόβλεψη των καθυστερήσεων το οποίο θα αντιμετωπιστεί

ως ένα πρόβλημα κατηγοριοποίησης. Συνεπώς από την οπτική της επιβλεπόμενης μάθησης με τη χρήση αλγορίθμων όπως εκμάθηση συνόλου, συσκευασίας θα δούμε τι τιμές ακρίβειας, precision recall , f1-score λαμβάνουμε. Στη συνέχεια θα δούμε αν μπορούμε να οδηγηθούμε σε παρεμφερή αποτελέσματα με τη χρήση πολυεπιπεδων νευρωνικών δικτύων και βαθιάς μάθησης.

Υπάρχουν πολλά παραδείγματα λογισμικού ανοιχτού κώδικα, εν μέρει στην Python, που διευκολύνουν τους αναλυτές να εκπαιδεύουν, να δοκιμάζουν και να επικυρώνουν μοντέλα. Παραδείγματος χάριν το Scikit-learn είναι μια εργαλειοθήκη της Python που μπορεί να χρησιμοποιηθεί για να συγκρίνει την εγκυρότητα και την αξιοπιστία των μοντέλων πρόβλεψης και να επιλέξει τα με την καλύτερη απόδοση[17].

1.2 Οργάνωση του τόμου

Η εργασία αυτή είναι οργανωμένη σε έξι κεφάλαια: Στο Κεφάλαιο 2 δίνεται το θεωρητικό υπόβαθρο των βασικών τεχνολογιών που σχετίζονται με τη διπλωματική αυτή, όσον αφορά το κομμάτι της της Μηχανικής Μάθησης. Αρχικά περιγράφονται η έννοια της Μηχανικής Μάθησης και διάφορες μέθοδοι της. Στη συνέχεια ασχολούμαστε συγκεκριμένα με την Επιβλεπόμενη Μάθηση και τα Βαθιά Νευρωνικά Δίκτυα που αποτελούν θεμέλια της εργασίας όπου γίνεται εις βάθος εισαγωγή στα συγκεκριμένα θεωρητικά και πρακτικά εργαλεία με τα οποία θα προσεγγίσουμε το θέμα. Στο κεφάλαιο 3 γίνεται σύντομη εισαγωγή στο ρόλο της Διερευνητικής Ανάλυσης δεδομένων και τη σημασία της. Στο Κεφάλαιο 4 παρουσιάζεται το σύνολο των δεδομένων που θα χρησιμοποιήσουμε για την εκπαίδευση , καθώς και η ανάλυση των χαρακτηριστικών-μεταβλητών με την διευρενητική ανάλυση δεδομένων. Στο κεφάλαιο 5, γίνεται αναλυτική περιγραφή της πειραματικής διαδικασίας καθώς και σχολιασμός των αποφάσεων που πήραμε για τις παραμέτρου του συστήματος. Ύστερα έχουμε την ανάλυση των αποτελεσμάτων που προέκυψαν. Τέλος στο Κεφάλαιο 6 παραθέτονται τα τελικά συμπεράσματα καθώς και μελλοντικές επεκτάσεις.

Μέρος I

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό υπόβαθρο Μηχανικής Μάθησης

2.1 Μηχανική Μάθηση

2.1.1 Ορισμός και μέθοδοι Μηχανικής Μάθησης

Η Μηχανική Μάθηση (Machine Learning - ML) είναι ένα πεδίο της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI) που ασχολείται με το σχεδιασμό προγραμμάτων τα οποία μπορούν να μαθαίνουν κανόνες από δεδομένα και μπορούν να προσαρμόζονται στις αλλαγές και να βελτιώνουν την απόδοση τους με την εμπειρία χωρίς ανθρώπινη παρέμβαση [18]. Ο Tom M. Mitchell πρότεινε έναν πιο επίσημο ορισμό που χρησιμοποιείται ευρέως: "Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει απο εμπειρία Ε ως προς μια κλάση εργασιών Τ και ένα μέτρο επίδοσης Π, αν η επίδοση του σε εργασίες της κλάσης Τ, όπως αποτιμάται απο το μέτρο Π, βελτιώνεται με την εμπειρία Ε" [19].

Εκτός από το ότι είναι ένα από τα αρχικά όνειρα της Επιστήμης Υπολογιστών, η Μηχανική Μάθηση έχει καθοριστική σημασία, σήμερα, καθώς οι υπολογιστές προβλέπεται να επιλύουν όλο και πιο περίπλοκα προβλήματα και να ενοποιούνται περισσότερο στην καθημερινότητά μας. Η σημασία της Μηχανικής Μάθησης έγκειται στο ότι πολλά από τα προβλήματα που θέλουμε να λύσουμε δεν είναι εργασίες που γνωρίζουμε ρητά πως να πούμε σε έναν υπολογιστή να τις λύσει.

Οι αλγόριθμοι μηχανικής μάθησης μπορούν να ταξινομηθούν σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση του εκπαιδευτικού «σήματος» ή την «ανατροφοδότηση» που είναι διαθέσιμα σε ένα σύστημα εκμάθησης. [20]. Αυτές είναι:

- **Επιβλεπόμενη μάθηση (Supervised Learning)**: Ονομάζεται επίσης κατηγοριοποίηση ή επαγωγική μάθηση. Ένας αλγόριθμος επιβλεπόμενης μάθησης αναλύει τα δεδομένα εκπαίδευσης και παράγει μια συνάρτηση που αντιστοιχεί ένα σύνολο μεταβλητών εισόδου Χ σε με μια επιθυμητή μεταβλητή εξόδου Ψ. Σκοπός είναι η πρόβλεψη των εξόδων για δεδομένα που δεν έχουμε δει ακόμα [21]. Οι αλγόριθμοι της Επιβλεπόμενης Μάθησης μπορούν να χωριστούν σε δύο κατηγορίες με βάση την επιθυμητή έξοδο.
  - **Κατηγοριοποίηση (Classification)**, που χρησιμοποιείται όταν οι έξοδοι ανήκουν σε ένα διακριτό σύνολο τιμών οι οποίες θα λέγονται κλάσεις. Στην πιο απλή περίπτωση όπου έχουμε δύο πιθανές κλάσεις το πρόβλημα θα ονομάζεται δυαδική

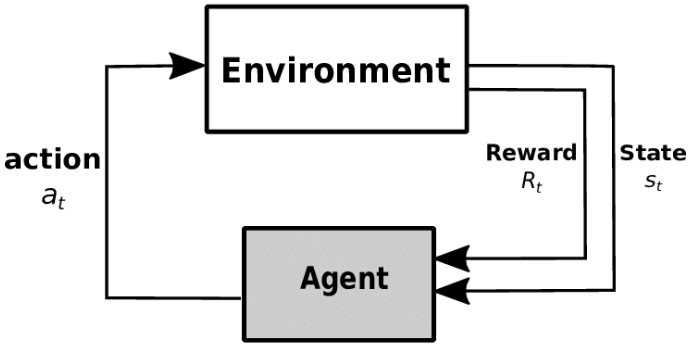
κατηγοριοποίηση (binary classification)

– **Παλινδρόμηση (Regression)**, που χρησιμοποιείται για την πρόβλεψη εξόδων που ανήκουν σε ένα σύνολο πραγματικών αριθμών.

Περισσότερες πληροφορίες για την επιβλεπόμενη μάθηση αναφέρονται στο επόμενο υποκεφάλαιο μιας και είναι το αντικείμενο που ταιριάζει στο θέμα της παρούσας διπλωματικής όπου οι πτήσεις θα κατηγοριοποιούνται σε καθυστερημένες και μη.

- **Μη Επιβλεπόμενη μάθηση (Unsupervised Learning)**: Ένας αλγόριθμος μη επιβλεπόμενης μάθησης λαμβάνει ένα σύνολο δεδομένων το οποίο περιέχει μόνο εισόδους και στοχεύει να βρει μία ομαδοποίηση ή συσταδοποίηση τους εντοπίζοντας κοινά σημεία. Κατά μία έννοια, η μη επιβλεπόμενη μάθηση μπορεί να θεωρηθεί ως εύρεση προτύπων στα δεδομένα πέρα από αυτό που θεωρείται καθαρός μη δομημένος θόρυβος [22]. Δύο κλασικά παραδείγματα μη επιβλεπόμενηω μάθησηω είναι η ομαδοποίηση (clustering) και η μείωση διαστάσεων (dimensionality reduction).

- **Ενισχυτική Μάθηση (Reinforcement Learning- RL)**: Ένα μοντέλο RL αποτελείται από πέντε δομικά στοιχεία και πιο συγκεκριμένα τους *πράκτορες* (agents), το *περιβάλλον* (environment), τις *καταστάσεις* (states), τις *δράσεις* (actions) και τις *ανταμοιβές* (rewards). Ο τρόπος με τον οποίο αλληλεπιδρούν αυτά τα δομικά στοιχεία φαίνεται στο Σχήμα 2.1 Είναι το πρόβλημα που αντιμετωπίζει ένας πράκτορας λογισμικού που πρέπει να μάθει να προσαρμόζει τη συμπεριφορά του μέσω αλληλεπιδράσεων δοκιμής και λάθους με ένα δυναμικό περιβάλλον μεγιστοποιώντας ένα αριθμητικό σήμα ανταμοιβής[23]. Στις πιο ενδιαφέρουσες και προκλητικές περιπτώσεις, ο αλγόριθμος θα εκτελέσει ενέργειες που μπορεί να επηρεάσουν όχι μόνο την άμεση ανταμοιβή αλλά και την επόμενη κατάσταση και, μέσω αυτής, όλες τις επόμενες ανταμοιβές.



Σχήμα 2.1: *Παράδειγμα ενός RL δικτύου (Πηγή: [1])*

2.2 Επιβλεπόμενη Μάθηση

Η παρούσα διπλωματική εργασία ασχολείται με το θέμα της πρόβλεψης καθυστερήσεων σε πτήσεις το οποιο ανήκει στην κατηγορία κλασικής επιβλεπόμενης μάθησης. Κρίνεται απαραίτητο λοιπόν να παρουσιάσουμε και να εξηγήσουμε σε βάθος τις τεχνικές επιβλεπόμενης μάθησης που θα χρησιμοποιηθούν για το σκοπό της εργασίας.

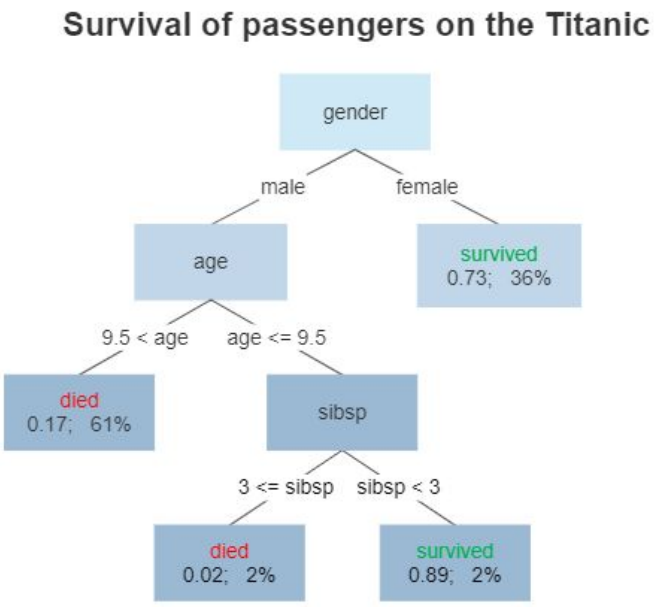
2.2.1 Δέντρα Απόφασης (Decision Trees)

Τα δέντρα απόφασης είναι μηχανισμοί πρόβλεψης γενικού σκοπού που εφαρμόζονται ευρέως ως ταξινομητές σε προβλήματα κατηγοριοποίησης. Ο "De Ville Barry" [24] αναφέρει ότι είναι ένα υψηλά διεπιστημονικό εργαλείο, με πεδία εφαρμογής όπως : τεχνητή νοημοσύνη, μηχανική μάθηση, ανακάλυψη γνώσης και εξόρυξη δεδομένων .

Ένα δέντρο απόφασης χρησιμοποιείται για τη δημιουργία μοντέλων που μιμούνται την δομή ενός δέντρου. Ταξινομεί τα δεδομένα αναδρομικά σε υποσύνολα : από τον ¨κορμό δέντρου¨ (tree trunk), σε ¨κλαδιά¨ (branches) έως ¨φύλλα¨ (leafs). Κάθε κόμβος στο δέντρο καθορίζει μια δοκιμή κάποιου χαρακτηριστικού του στιγμιότυπου και κάθε κλάδος που κατέρχεται από αυτόν τον κόμβο αντιστοιχεί σε μία από τις πιθανές τιμές για αυτό το χαρακτηριστικό. Ένα στιγμιότυπο ταξινομείται ξεκινώντας από τον ριζικό κόμβο του δέντρου, δοκιμάζοντας το χαρακτηριστικό που καθορίζεται από αυτόν τον κόμβο και μετά μετακινώντας προς τα κάτω τον κλάδο δέντρου που αντιστοιχεί στην τιμή του χαρακτηριστικού στο συγκεκριμένο παράδειγμα. Αυτή η διαδικασία στη συνέχεια επαναλαμβάνεται για το υποδέντρο που έχει ρίζες στον νέο κόμβο. [19].

Το παρακάτω σχήμα αποτελεί ένα παράδειγμα δέντρου απόφασης το οποίο σκοπεύει στην λήψη απόφασης ζωής ή θανάτου για κάποιον άνθρωπο που υπήρξε στον Τιτανικό, λαμβάνοντας υπόψιν χαρακτηριστικά όπως φύλο,ηλικία κ.ά.

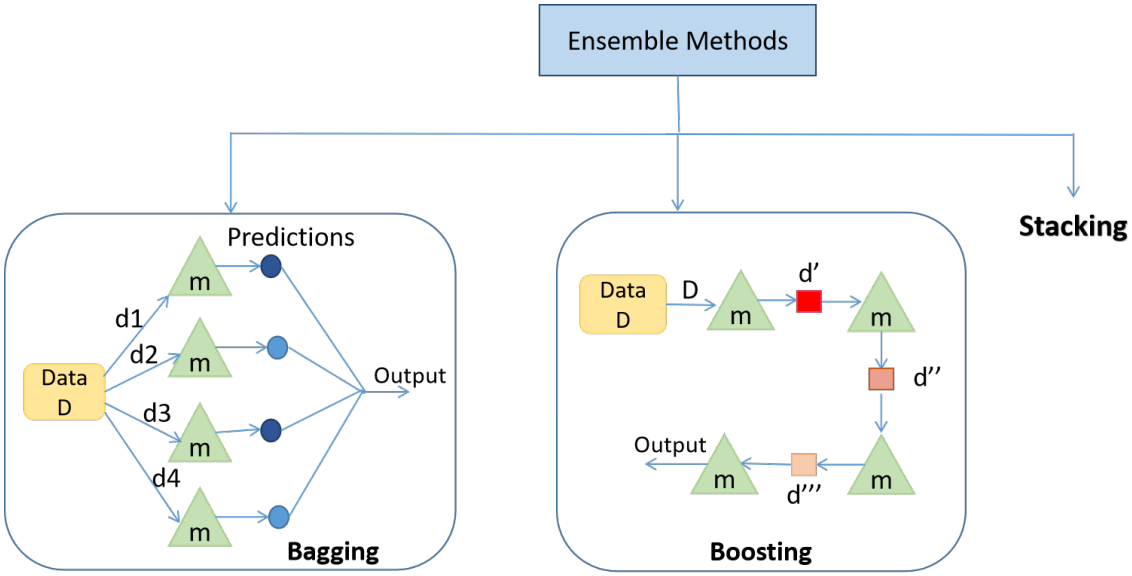
Ένα μεγάλο μειονέκτημα του αλγορίθμου των δέντρων απόφασης είναι ότι οδηγεί εύκολα σε υπερπροσαρμογή (overfitting), αφού υπάρχει η πιθανότητα δημιουργίας μεγάλου πλήθους κλαδιών λόγω θορύβου ή ακραίων τιμών (outliers). Σε μια τέτοια περίπτωση παρόλο που το μοντέλο μπορεί να έχει καλή απόδοση στα δεδομένα εκπαίδευσης θα έχει κακή απόδοση σε καινούρια δεδομένα.



Σχήμα 2.2: *Παράδειγμα ενός Decision Tree (Πηγή :[2])*

2.3 Εκμάθηση Συνόλου (Ensemble learning)

Η εκμάθηση συνόλου είναι ένας γενικός όρος (umbrella term) για μεθόδους που συνδυάζουν πολλαπλούς επαγωγείς (inducers) για να λάβουν μια απόφαση, συνήθως σε εργασίες επιβλεπόμενης μηχανικής μάθησης [25]. Ένας επαγωγέας, που αναφέρεται επίσης ως βασικός μαθητής (base learner), είναι ένας αλγόριθμος που λαμβάνει ένα σύνολο επισημασμένων (labeled) παραδειγμάτων ως είσοδο και παράγει ένα μοντέλο (π.χ. ταξινομητή ή παλινδρομητή) που γενικεύει αυτά τα παραδείγματα. Χρησιμοποιώντας το παραχθέν μοντέλο, μπορούν να γίνουν προβλέψεις για νέα παραδείγματα μη επισημασμένα. Ένας επαγωγέας συνόλου μπορεί να είναι αλγόριθμος οποιουδήποτε τύπου μηχανικής μάθησης (π.χ. δέντρα αποφάσεων, νευρωνικά δίκτυα, μοντέλα παλινδρόμησης, κ.λπ.). Η κύρια υπόθεση της εκμάθησης συνόλου είναι ότι συνδυάζοντας πολλαπλά μοντέλα, τα σφάλματα ενός μόνο επαγωγέα πιθανότατα θα αντισταθμιστούν από άλλους επαγωγείς, και ως αποτέλεσμα, η συνολική απόδοση πρόβλεψης του συνόλου θα είναι καλύτερη από αυτή ενός μεμονωμένου επαγωγέα. Το παραπάνω εξηγεί ουσιαστικά τον λόγο που οι μέθοδοι εκμάθησης συνόλου θεωρούνται τελευταίας τεχνολογίας (state-of-the-art) λύση για πολλές προκλήσεις της μηχανικής μάθησης . Οι επαγωγείς πολλές φορές αναφέρονται και ως μαθητές ή αδύναμοι μαθητές. Ο συνδυασμός πολλών αδύναμων μαθητών δημιουργεί έναν δυνατό μαθητή.



Σχήμα 2.3: *Αναπαράσταση των τύπων μεθόδων συνόλου. Εδώ, το m αντιπροσωπεύει έναν αδύναμο μαθητή. d1, d2, d3, d4 είναι τα τυχαία δείγματα από τα δεδομένα D. d', d", d" είναι ενημερωμένα δεδομένα εκπαίδευσης με βάση τα αποτελέσματα από τον προηγούμενο αδύναμο μαθητή.(Πηγή :[3])*

2.3.1 Γιατί δουλεύουν οι τεχνικές εκμάθησης συνόλου·

Υπάρχουν διάφοροι λόγοι για τους οποίους οι μέθοδοι συνόλου συχνά βελτιώνουν την προγνωστική απόδοση:

- **Αποφυγή υπερπροσαρμογής**: Όταν είναι διαθέσιμος μόνο ένας μικρός όγκος δεδο-

μένων, ένας αλγόριθμος μάθησης είναι επιρρεπής στην εύρεση πολλών διαφορετικών υποθέσεων που προβλέπουν τέλεια όλα τα δεδομένα εκπαίδευσης, ενώ κάνει κακές προβλέψεις για περιπτώσεις που δεν έχει δει[25].

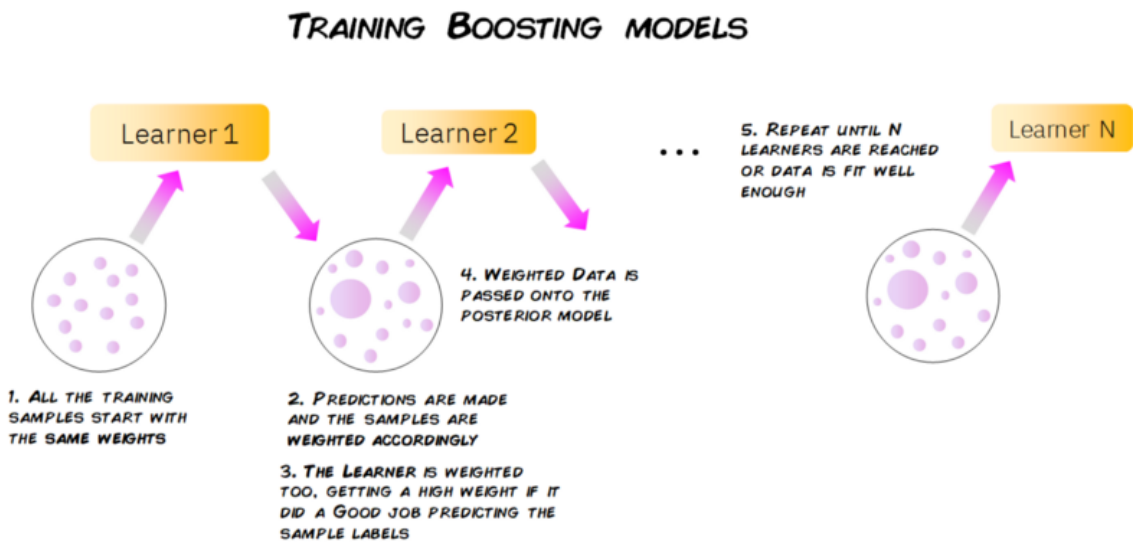
Ο μέσος όρος διαφορετικών υποθέσεων μειώνει τον κίνδυνο επιλογής λανθασμένης υπόθεσης και επομένως βελτιώνει τη συνολική απόδοση του συστήματος πρόβλεψης.

- **Υπολογιστικό πλεονέκτημα**: Οι μεμονωμένοι μαθητές που πραγματοποιούν τοπικές αναζητήσεις μπορεί να κολλήσουν στο τοπικό βέλτιστο. Με το συνδυασμό πολλών μαθητών, οι μέθοδοι συνόλου μειώνουν τον κίνδυνο απόκτησης ενός τοπικού ελάχιστου[25].

- **Αναπαράσταση**: Η βέλτιστη υπόθεση μπορεί να βρίσκεται εκτός του χώρου οποιουδήποτε μεμονωμένου μοντέλου. Συνδυάζοντας διαφορετικά μοντέλα, ο χώρος αναζήτησης μπορεί να επεκταθεί και ως εκ τούτου, επιτυγχάνεται καλύτερη προσαρμογή στο χώρο δεδομένων[25].

2.4 Η Τεχνική της Ενίσχυσης (Boosting)

Η Ενίσχυση είναι μια μέθοδος συνόλου στην οποία κάθε προγνωστικός παράγοντας μαθαίνει από τα προηγούμενα λάθη των προβλέψεων για να κάνει καλύτερες προβλέψεις στο μέλλον. Η τεχνική συνδυάζει αρκετούς αδύναμους μαθητές(weak learners) που είναι διατεταγμένοι με διαδοχικό τρόπο έτσι ώστε ο καθένας να μαθαίνει από τα λάθη του προηγούμενου ώστε να δημιουργηθεί ένα καλύτερο μοντέλο πρόβλεψης. Ως εκ τούτου, ένας ισχυρός μαθητής διαμορφώνεται μέσω της σημαντικής βελτίωσης της προβλεψιμότητας των μοντέλων. Παρακάτω φαίνεται ο τρόπος εκπαίδευσης ενός μοντέλου φτιαγμένο με την τεχνική της ενίσχυσης.



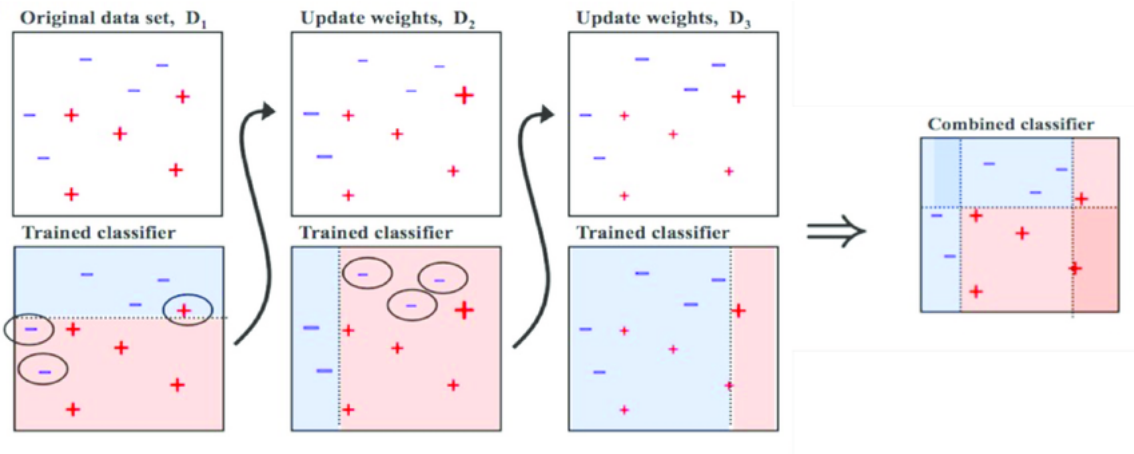
Σχήμα 2.4: *Training Boosting Models(Πηγή :[4])*

Παραδείγματα γνωστών αλγορίθμων ενίσχυσης αποτελούν οι:

2.4.1 AdaBoost - Αλγόριθμος Προσαρμοσμένης Ενίσχυσης :

Η κύρια ιδέα του AdaBoost είναι να επικεντρωθεί σε περιπτώσεις που προηγουμένως είχαν ταξινομηθεί εσφαλμένα κατά την εκπαίδευση ενός νέου επαγωγέα. Το επίπεδο εστίασης που δίνεται καθορίζεται από ένα βάρος που αποδίδεται σε κάθε περίπτωση στο σετ προπόνησης. Η πιο απλή περίπτωση προσαρμοσμένης ενίσχυσης είναι αυτή που χρησιμοποιεί δέντρα απόφασης ως weak learners ή επαγωγείς. Σε αυτή την περίπτωση κάθε αδύναμος μαθητής είναι ένα Decision Tree Stump που χωρίζει τα δεδομένα σε 2 κατηγορίες. Το κάθε δέντρο απόφασης λοιπόν αξιολογεί ένα χαρακτηριστικό του συνολικού Dataset. Συνδυασμός μεγάλου αριθμού τέτοιων επαγωγέων δημιουργεί ένα μοντέλο που συγκλίνει σε μεγάλη ακρίβεια πρόβλεψης.

Ο αλγόριθμος έχει ως εξής: Στην πρώτη επανάληψη, το ίδιο βάρος αποδίδεται σε όλες τις περιπτώσεις. Σε κάθε επανάληψη, τα βάρη των λανθασμένων περιπτώσεων αυξάνονται, ενώ τα βάρη των σωστά ταξινομημένων περιπτώσεων μειώνονται. Επιπλέον, βάρη αποδίδονται και στους εκάστοτε base learners με βάση την συνολική προγνωστική απόδοση.



Σχήμα 2.5: *Training of an AdaBoost classifier. The first classifier trains on unweighted data, then reweights the data for the next and so on to produce the final classifier. (Πηγή :[5])*

2.4.2 Gradient Boost : Ενίσχυση Κλίσης

Σε μηχανές ενίσχυσης κλίσης, ή απλά, GBM, η διαδικασία μάθησης προσαρμόζει διαδοχικά νέα μοντέλα για να παρέχει μια πιο ακριβή εκτίμηση της μεταβλητής εξόδου. Η βασική ιδέα πίσω από αυτόν τον αλγόριθμο είναι η κατασκευή των νέων βασικών μαθητών base learners με τρόπο τέτοιο ώστε ο τωρινός μαθητής να είναι πιο αποδοτικός απο τον προηγούμενο του. Σε αντίθεση με την περίπτωση της Προσαρμοσμένης Ενίσχυσης εδώ τα βάρη δεν αυξάνουν αλλά βελτιστοποιούμε σε κάθε βήμα την συνάρτηση απώλειας του προηγούμενου μαθητή. Αυτό επιτυγχάνεται όταν οι νέοι βασικοί μαθητές συσχετίζονται στο μέγιστο με την αρνητική κλίση της συνάρτησης απώλειας, η οποία αφορά το σύνολο (ensemble) [26].Η επαναληπτική διαδικασία σταματά είτε όταν το σφάλμα δεν αλλάζει.

Algorithm 1 Friedman's Gradient Boost algorithm

Inputs:

- input data $(x, y)_{i=1}^{N}$
- number of iterations M
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

1: initialize $\widehat{f}_0$ with a constant
2: **for** $t = 1$ to M **do**
3: compute the negative gradient $g_t(x)$
4: fit a new base-learner function $h(x, \theta_t)$
5: find the best gradient descent step-size ρ_t:

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi\left[y_i, \widehat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right]$$

6: update the function estimate:

$$\widehat{f}_t \leftarrow \widehat{f}_{t-1} + \rho_t h(x, \theta_t)$$

7: **end for**

Σχήμα 2.6: *Algorithm 1. Friedman's Gradient Boost algorithm.(Πηγή :[6])*

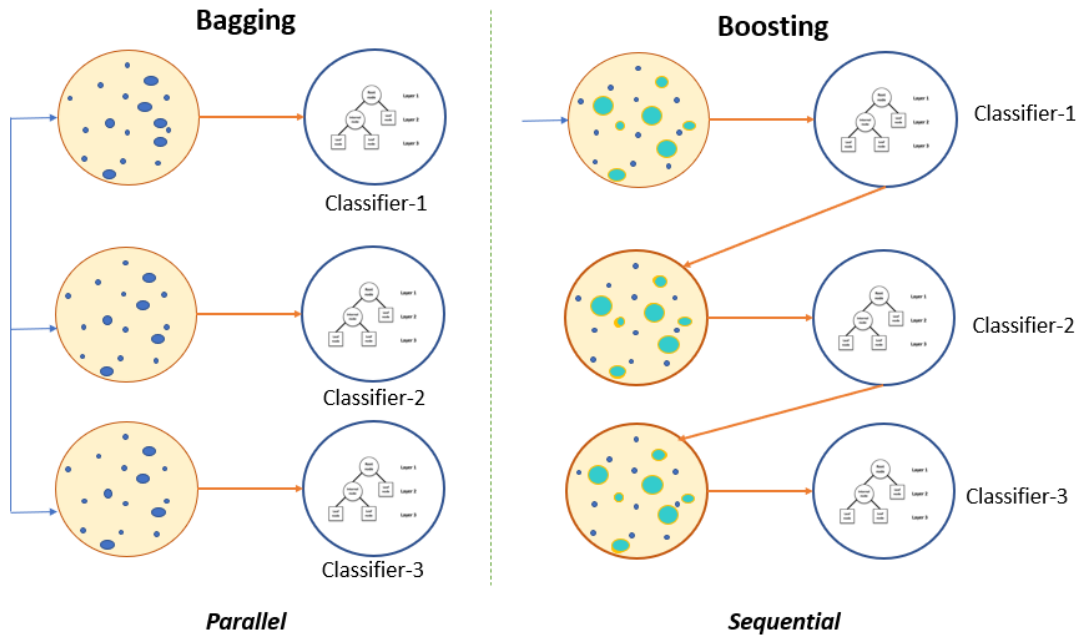
2.4.3 XGBoost : Ακραία Ενίσχυση κλίσης

Το XGBoost είναι συντομογραφία για το πακέτο eXtreme Gradient Boost. Είναι μια αποτελεσματική και επεκτάσιμη εφαρμογή του αλγορίθμου ενίσχυσης κλίσης από τον [27]. Το πακέτο περιλαμβάνει επιλυτή γραμμικών μοντέλων και αλγόριθμο εκμάθησης δέντρων. Υποστηρίζει διάφορες συναρτήσεις σκοπού, συμπεριλαμβανομένης της παλινδρόμησης, της ταξινόμησης και της κατάταξης. Είναι σχεδιασμένο ώστε να επιτυχαχύνει την υπολογιστική πολυπλοκότητα και να δημιουργεί αποδοτικά μοντέλα. Με το XGBoost, τα δέντρα δημιουργούνται παράλληλα, αντί να είναι διαδοχικά όπως το Gradient Boosting Decision Trees - (GBDT) . Αρχικά, δημιουργεί ένα μοντέλο βασισμένο σε ένα υποσύνολο των δεδομένων. Με αυτό το μοντέλο γίνονται προβλέψεις σε ολόκληρο το σύνολο των δεδομένων εκπαίδευσης και κατόπιν υπολογίζεται το βελτιστοποιεί το σφάλμα μέσω της αρνητική κλίσης της συνάρτησης απώλειας. Ύστερα ακολουθεί μια στρατηγική επιπεδο ανά επίπεδο, σαρώνοντας τις τιμές κλίσης και χρησιμοποιώντας αυτά τα επιμέρους αθροίσματα για την αξιολόγηση της ποιότητας των διαχωρισμών σε κάθε πιθανό διαχωρισμό στο σετ εκπαίδευσης.

2.5 Η Τεχνική της Συσκευασίας (Bagging)

Η τεχνική της Συσκευασίας Bagging [28] είναι μια απλή αλλά αποτελεσματική προσέγγιση για τη δημιουργία ενός συνόλου ανεξάρτητων μοντέλων στα οποία κάθε επαγωγέας εκπαιδεύεται χρησιμοποιώντας ένα δείγμα στιγμιότυπων που λαμβάνονται από το αρχικό σύνολο δεδομένων με αντικατάσταση. Προκειμένου να διασφαλιστεί επαρκής αριθμός στιγμιότυπων ανά επαγωγέα, κάθε δείγμα περιέχει συνήθως τον ίδιο αριθμό στιγμιότυπων όπως στο αρχικό σύνολο δεδομένων. Στην συνέχεια γίνεται ψηφοφορία voting και η πλειο-

ψηφία των προβλέψεων των επαγωγέων καθορίζει την τελική πρόβλεψη ενός αόρατουunseen στιγμιότυπου [29]. Εφόσον η δειγματοληψία γίνεται με αντικαταστάσεις, ορισμένα από τα αρχικά στιγμιότυπα είναι πιθανό να εμφανιστούν περισσότερες από μία φορές όταν εκπαιδεύετε ο ίδιος επαγωγέας, ενώ άλλα στιγμιότυπα μπορεί να μην περιλαμβάνονται καθόλου. Δεδομένου ότι οι επαγωγείς εκπαιδεύονται ανεξάρτητα, το bagging μπορεί εύκολα να εφαρμοστεί με παράλληλο τρόπο εκπαιδεύοντας κάθε επαγωγέα χρησιμοποιώντας διαφορετικές υπολογιστικές μονάδες.



Σχήμα 2.7: *Bagging (independent predictors) vs. Boosting (sequential predictors).(Πηγή :[7])*

2.5.1 Τυχαία Δάση (RF - Random Forests)

Μία απο τις πιο διαδεδομένες μεθόδους εκμάθησης συνόλου (Ensemble Learning) είναι τα Τυχαία Δάση. Χρησιμοποιούνται συνήθως για προβλήματα ταξινόμησης(ή κατηγοριοποίησης) και παλινδρόμησης. Τα RF εισήχθησαν από τον Leo Breiman [30] ο οποίος εμπνεύστηκε από παλαιότερες εργασίες των Amit και Geman και είναι μια επέκταση της ιδέας του Breiman για το Bagging που αναπτύχθηκε ως ανταγωνιστής της ενίσχυσης. Ο αλγόριθμος RF χρησιμοποιεί έναν μεγάλο αριθμό ανεξάρτητων δέντρων απόφασης. Τα εκάστοτε δέντρα κατασκευάζονται με τη διαδικασία που περιγράφεται στον αλγόρθμο του Σχήματος 2.8. Η παράμετρος εισόδου N αντιπροσωπεύει τον αριθμό των μεταβλητών εισόδου που χρησιμοποιούνται για τον προσδιορισμό της απόφασης σε έναν κόμβο του δέντρου. Προκειμένου να αναπτυχθεί ένα τυχαίο δέντρο που εξακολουθεί να είναι επαρκώς ακριβές, η τυχαιότητα εγχέεται στους επαγωγείς του δέντρου απόφασης χρησιμοποιώντας δύο διαδικασίες τυχαιοποίησης: (α) εκπαίδευση κάθε δέντρου σε διαφορετικό δείγμα στιγμιοτύπων(τυχαίο υποσύνολο από τα συνολικά δεδομένα) ή (β) αντί να επιλέγει τον καλύτερο διαχωρισμό σε κάθε κόμβο, ο επαγωγέας δειγματοληπτεί τυχαία ένα υποσύνολο των χαρακτηριστικών και επιλέγει τον καλύτερο διαχωρισμό μεταξύ τους. Μετά την διαδικασία κατασκευής των δέντρων που θα συμβει κατά την εκπαίδευση του RF, επέρχεται όπως υποδηλώνει και η τεχνική του Bagging

ψηφοφορία των δέντρων ώστε να καθοριστεί η συμπεριφορά του μοντέλου σε δεδομένα που δεν έχουμε δει (unseen data).

Algorithm 2: The Random Forest algorithm

Input: IDT(a decision tree inducer),T(the number of iterations), S(training set), μ(the subsample size), N(number of attributes used in each node)

Output: M_t:$\forall t = 1, ..., T$

for *each* t *in* $1, ..., T$ **do**

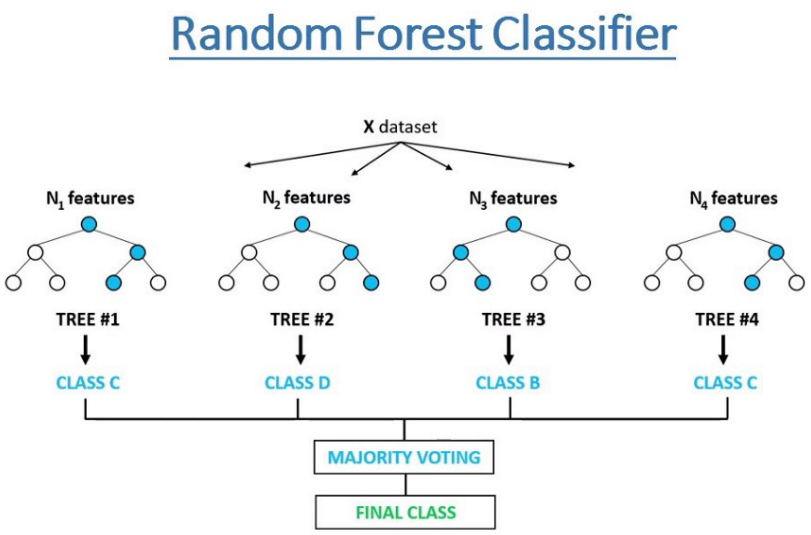
  $S_t \leftarrow$ Sample μ instances from S with replacement.

  Build classifier M_t using $IDT(N)$ on S_t

  $t++$

end

Σχήμα 2.8: *The Random Forest Algorithm (Πηγή :[8])*



Σχήμα 2.9: *Random Forest Classifier (Πηγή :[9])*

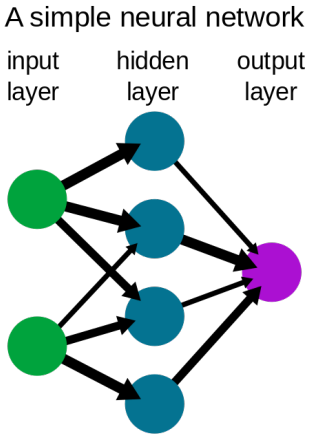
2.6 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANN)

Τα τεχνητά νευρωνικά δίκτυα (ANN), που συνήθως ονομάζονται απλά νευρωνικά δίκτυα (NN) , είναι υπολογιστικά συστήματα εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα που συνιστούν τους εγκεφάλους των ζώων.

Ένα ANN βασίζεται σε μια συλλογή συνδεδεμένων μονάδων ή κόμβων που ονομάζονται τεχνητοί νευρώνες, οι οποίοι μοντελοποιούν τους νευρώνες σε έναν βιολογικό εγκέφαλο [31].

Κάθε σύνδεση, όπως οι συνάψεις σε έναν βιολογικό εγκέφαλο, μπορεί να μεταδώσει ένα σήμα σε άλλους νευρώνες.

Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.



Σχήμα 2.10: *Simple neural network (Πηγή :[10])*

2.6.1 Perceptrons

Ο νευρώνας Perceptron είναι ένα είδος τεχνητού νευρωνικού δικτύου που εφευρέθηκε το 1957 στο Αεροναυτικό Εργαστήριο του Κορνέλλ από τον Φρανκ Ρόζενμπλατ.

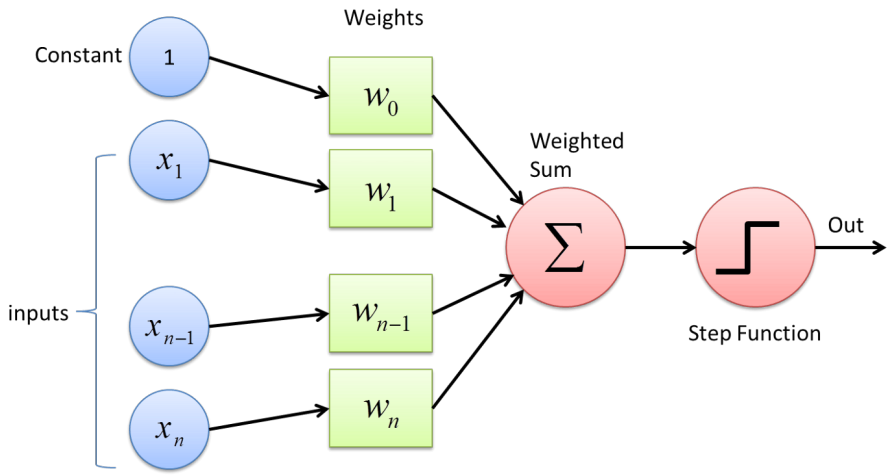
Ο Perceptron είναι ένας δυαδικός ταξινομητής, δηλαδή μία συνάρτηση η οποία απεικονίζει την είσοδο x (ένα διάνυσμα με πραγματικές τιμές) σε μία τιμή εξόδου f(x) (μία και μοναδική δυαδική τιμή)

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b \geq 0 \\ 0 & \text{else} \end{cases}$$

που w είναι ένα διάνυσμα από βάρη με πραγματικές τιμές και w*x είναι το εσωτερικό γινόμενο μεταξύ των διανυσμάτων w και εισόδων x. Το b είναι το 'bias', ένας σταθερός όρος ο οποίος δεν εξαρτάται από καμία τιμή εισόδου.

Η τιμή της f(x) θα είναι 1 ή 0 και χρησιμοποιείται για να ταξινομήσει το x είτε ως θετικό ή αρνητικό στιγμιότυπο, στην περίπτωση ενός δυαδικού προβλήματος ταξινόμησης. Το bias χρησιμοποιείται για την μετατόπιση της συνάρτησης ενεργοποίησης ή για να δώσει στον νευρώνα εξόδου ένα βασικό επίπεδο δραστηριότητας. Αν το b είναι αρνητικό τότε ο βεβαρημένος συνδυασμός των εισόδων πρέπει να παραγάγει μία θετική τιμή μεγαλύτερη του -b έτσι ώστε να αναγκάσει τον νευρώνα που ταξινομεί να έχει τιμή άνω του κατωφλίου 0. Χωρικά, το bias μεταβάλει την θέση (αλλά όχι τον προσανατολισμό) του συνόρου απόφασης.

Εφόσον οι είσοδοι τροφοδοτούνται στο δίκτυο άμεσα μέσω των βεβαρημένων συνδέσεων, ο νευρώνας μπορεί να θεωρηθεί ως ένα απλό είδος νευρωνικού δικτύου εμπρός τροφοδότησης.



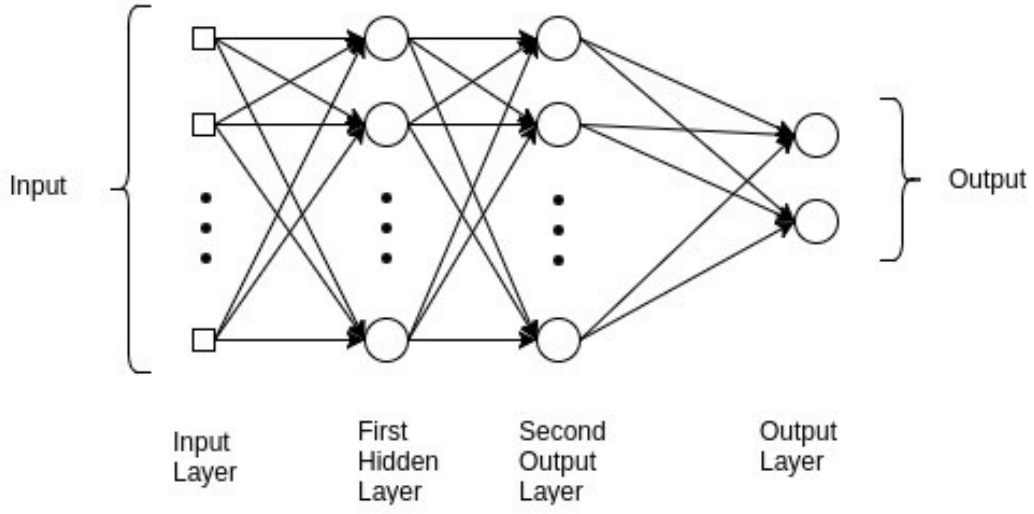
Σχήμα 2.11: *παράδειγμα ενός Perceptron (Πηγή :[11])*

2.6.2 Multi Layer Perceptrons - MLP

Ένα *πολυεπίπεδο Perceptron* (Multi-layer Perceptrons - MLP) είναι ένα πλήρως συνδεδεμένο τεχνητό νευρωνικό δίκτο (ANN) πρόσθιας τροφοδότησης (feed forward). Αποτελείται από πολλά *επίπεδα* (layers) επεξεργασίας μεταξύ της εισόδου και εξόδου. Σε ένα MLP υπάρχουν τουλάχιστον τρία επίπεδα νευρώνων, όπου οι νευρώνες ανήκουν σε ένα από τρία διαφορετικά είδη επιπέδου:

- Επίπεδο εισόδου (input layer)
- Κρυφά επίπεδα (hidden layers)
- Επίπεδο Εξόδου (output layer)

Όλα τα επίπεδα έχουν κόμβους και κάθε κόμβος έχει ένα βάρος το οποίο λαμβάνεται υπόψη κατά την επεξεργασία πληροφοριών από το ένα επίπεδο στο επόμενο επίπεδο. Ένα τυπικό MLP φαίνεται στο Σχήμα 2.12. Κάθε κόμβος του δικτύου είναι ένας νευρώνας, ο οποίος ενεργοποιείται από μη-γραμμική συνάρτηση (λ.χ. την ημι-γραμμική) η οποία λέγεται συνάρτηση ενεργοποίησης. Για την εκπαίδευση του δικτύου χρησιμοποιείται ο αλγόριθμος της προς τα πίσω διάδοσης του σφάλματος (back-propagation).



Σχήμα 2.12: *MLP με 2 κρυφά επίπεδα (Πηγή: [12])*

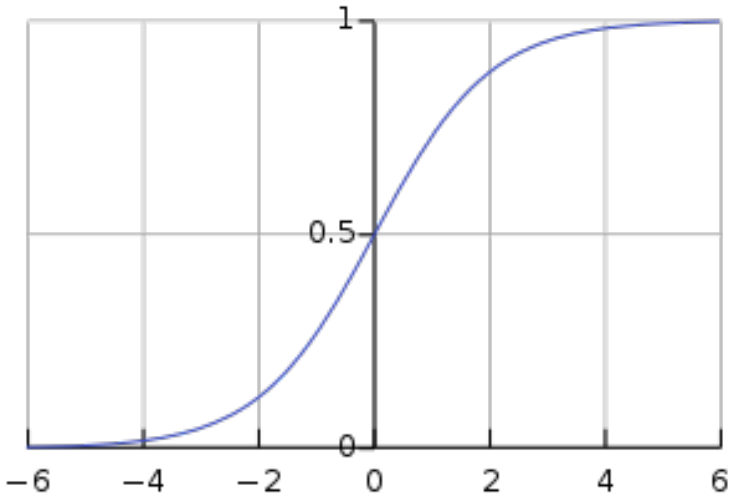
2.6.3 Συναρτήσεις Ενεργοποίησης (Activation Functions)

Τα σήματα που παράγονται απο έναν τεχνητό νευρώνα ενός νευρωνικού δικτύου όπως το perceptron έχουν την μορφή y = f($\sum_{i=1}^{n} w_i x_i + b$). Η f(x) ονομάζεται συνάρτηση ενεργοποίησης όπως αναφέραμε στο προηγούμενο κεφάλαιο. Εάν μια συνάρτηση ενεργοποίησης δεν χρησιμοποιείται σε ένα νευρωνικό δίκτυο, τότε το σήμα εξόδου θα είναι απλώς μια απλή γραμμική συνάρτηση που είναι απλώς ένα πολυώνυμο βαθμού ένα [32].

Μια γραμμική εξίσωση είναι απλή και εύκολη στην επίλυση αλλά η πολυπλοκότητά της είναι περιορισμένη και τα ANN που θα παραχθούν δεν έχουν τη δυνατότητα να μάθουν και να αναγνωρίσουν σύνθετες αντιστοιχίσεις από δεδομένα. Το νευρωνικό δίκτυο χωρίς συναρτήσεις ενεργοποίησης λειτουργεί ως μοντέλο γραμμικής παλινδρόμησης με περιορισμένη απόδοση και ισχύ τις περισσότερες φορές. Είναι επιθυμητό λοιπόν ένα νευρωνικό δίκτυο όχι μόνο να μαθαίνει και να υπολογίζει μια γραμμική συνάρτηση αλλά να εκτελεί εργασίες πιο περίπλοκες από αυτές όπως η μοντελοποίηση περίπλοκων τύπων δεδομένων όπως εικόνες, βίντεο, ήχος, ομιλία, κείμενο κ.λπ. Οι συναρτήσεις αυτές είναι εξαιρετικά χρήσιμες διότι μπορούν να οριστούν με κατάλληλο τρόπο ώστε η έξοδος τους να είναι οποιοσδήποτε αριθμός.

Η λειτουργία της συνάρτησης, μπορεί να χαρακτηριστεί ως λειτουργία φίλτρου που μεταφέρει την έξοδο στο διάστημα που κρίνεται βολικότερο. Υπάρχουν διάφορες συναρτήσεις ενεργοποίησης, παρακάτω παρουσιάζονται κάποιες από τις συνηθέστερες.

- Σιγμοειδής Συνάρτηση: Είναι η πιο ευρέως χρησιμοποιούμενη συνάρτηση ενεργοποίησης. Είναι μια μη γραμμική συνάρτηση. Ορίζεται ως: f(x) = $\frac{1}{1+exp(-x)}$. Η σιγμοειδής συνάρτηση μετασχηματίζει τις τιμές στην περιοχή από 0 έως 1 , είναι συνεχώς διαφοροποιήσιμη και είναι μια ομαλή συνάρτηση σε σχήμα S. Επίσης, η σιγμοειδής συνάρτηση δεν είναι συμμετρική ως προς το μηδέν, πράγμα που σημαίνει ότι τα πρόσημα όλων των τιμών εξόδου των νευρώνων θα είναι ίδια.



Σχήμα 2.13: *Logistic Curve Function (Πηγή: Wikipedia)*

- Binary Step Function : Όταν έχουμε μια συνάρτηση ενεργοποίησης, το πιο σημαντικό πράγμα που πρέπει να λάβουμε υπόψη είναι το κατώφλι του ταξινομητή δηλαδή αν η τιμή του γραμμικού μετασχηματισμού οφείλει να ενεργοποιήσει τον νευρώνα ή να τον απενεργοποιήσει. Μπορούμε να πούμε ότι ένας νευρώνας ενεργοποιείται εάν η είσοδος στη συνάρτηση ενεργοποίησης είναι μεγαλύτερη από μια τιμή κατωφλίου, διαφορετικά απενεργοποιείται. Σε αυτή την περίπτωση η έξοδος δεν τροφοδοτείται ως είσοδος στο επόμενο στρώμα. Η λειτουργία της Βηματικής Συνάρτησης είναι η απλούστερη ενεργοποίηση που υπάρχει για αυτήν την δουλειά. Ορίζεται ως

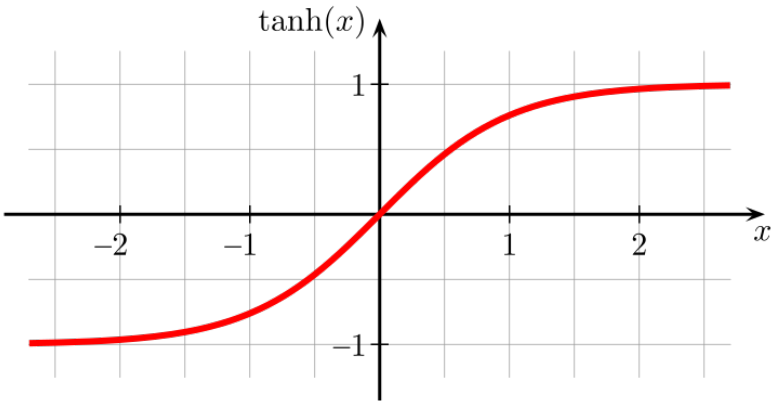
$$U(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

- Συνάρτηση Υπερβολικής Εφαπτομένης - (Tanh) : Η συνάρτηση Tanh είναι παρόμοια με τη σιγμοειδή συνάρτηση , αλλά είναι συμμετρική γύρω από το μηδέν. Αυτό έχει ως αποτέλεσμα διαφορετικά πρόσημα εξόδων από τα προηγούμενα στρώματα να τροφοδοτηθούν ως είσοδος στο επόμενο στρώμα. Ορίζεται ως: f(x) = $\frac{exp(2x)-1}{exp(2x)+1}$

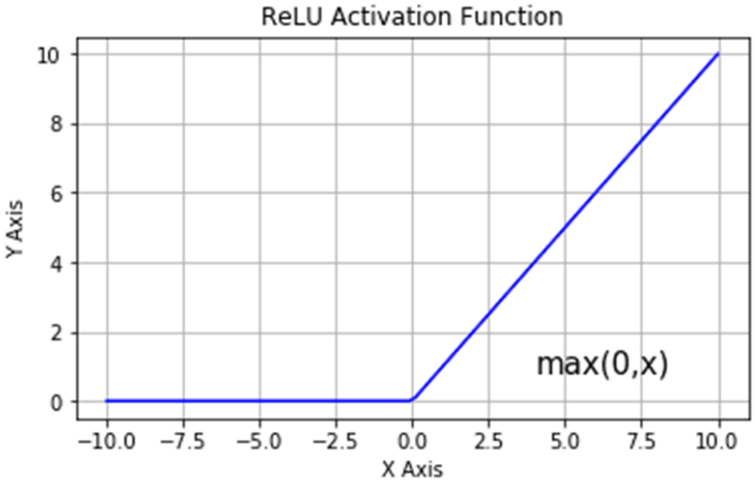
  Η συνάρτηση Tanh είναι συνεχής, διαφοροποιήσιμη και οι τιμές βρίσκονται στην περιοχή -1 έως 1.

- Μονάδα Γραμμικής Ανόρθωσης (Rectified Linear Unit – ReLU) :

  Η Μονάδα Γραμμικής Ανόρθωσης είναι μια μη γραμμική συνάρτηση ενεργοποίησης που χρησιμοποιείται ευρέως σε νευρωνικά δίκτυα. Το μεγάλο πλεονέκτημα της χρήσης της συνάρτησης ReLU είναι ότι όλοι οι νευρώνες δεν ενεργοποιούνται ταυτόχρονα. Αυτό σημαίνει ότι ένας νευρώνας θα απενεργοποιηθεί μόνο όταν η έξοδος του γραμμικού μετασχηματισμού είναι μηδέν. Ορίζεται μαθηματικά ως: f(x) = max(0,x)



Σχήμα 2.14: *Hyperbolic Tangent Function (Πηγή: Wikipedia)*



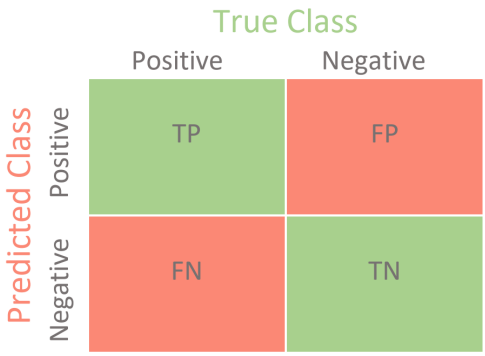
Σχήμα 2.15: *ReLU (Πηγή: [13])*

2.6.4 Εκπαίδευση Πολυστρωματικού Perceptron - MLP

Ο κυριότερος αλγόριθμος που χρησιμοποιείται στην εκπαίδευση MLPs είναι ο αλγόριθμος οπίσθιας διάδοσης (backpropagation). Ο αλγόριθμος εφαρμόζει την ιδέα της ελαφριάς μεταβολής των βαρών σε κάθε βήμα εκπαίδευσης, ανάλογα με τη συνεισφορά τους στη συνάρτηση σφάλματος που χρησιμοποιείται, προς την κατεύθυνση που ελαχιστοποιεί την τιμή σφάλματος. Ως βήμα εκπαίδευσης θεωρείται η τροφοδότηση του μοντέλου με ένα δεδομένα ή μια μικρή ομάδα των δεδομένων (batch) και προσαρμογή των βαρών σύμφωνα με τον κανόνα εκπαίδευσης.

2.6.5 Μετρικές Αξιολόγησης – Evaluation Metrics

Η επιλογή της κατάλληλης μετρικής αξιολόγησης είναι ένα κρίσιμο βήμα για να διασφαλίσουμε ότι η εκπαίδευση του μοντέλου του ταξινομητή είναι βέλτιστη[33]. Αυτές οι μετρικές εφαρμόζονται στα δεδομένα αξιολόγησης (test set). Για προβλήματα δυαδικής ταξινόμησης, η διάκριση της καλύτερης (βέλτιστης) λύσης κατά την εκπαίδευση ταξινόμησης μπορεί να οριστεί με βάση τον πίνακα σύγχυσης όπως φαίνεται στο Σχήμα 2.16.



Σχήμα 2.16: *(Πηγή: [14])*

- True Positive – TP: Αληθώς θετική ονομάζεται μια πρόβλεψη όπου το μοντέλο προβλέπει σωστά τη θετική κλάσση.
- False Positive – FP: Ψευδώς θετική. Ονομάζεται επίσης σφάλμα τύπου 1, ένα αποτέλεσμα όπου το μοντέλο προβλέπει λανθασμένα τη θετική κλάση όταν είναι στην πραγματικότητα αρνητική.
- True Negative – TN: Αληθώς αρνητική ονομάζεται μια πρόβλεψη όπου το μοντέλο προβλέπει σωστά την αρνητική τάξη.
- False Negative – FN: Ψευδώς αρνητική. Ονομάζεται επίσης σφάλμα τύπου 2, ένα αποτέλεσμα όπου το μοντέλο προβλέπει λανθασμένα την αρνητική κλάση όταν είναι πραγματικά θετική.

Συνολικά χρησιμοποιούμε τις παρακάτω μετρήσεις για την αξιολόγηση των αποτελεσμάτων του μοντέλου.

- Ορθότητα - Accuracy : Γενικά, μετρά την αναλογία των σωστών προβλέψεων ως προς τον συνολικό αριθμό των στιγμιοτύπων που αξιολογήθηκαν

  Accuracy = $\frac{tp+tn}{tp+tn+fp+fn}$
- Ακρίβεια – Precision : Είναι ο λόγος των σωστών αποτελεσμάτων πρόβλεψης μιας κλάσης προς τον συνολικό αριθμό των προβλέψεων αυτής της κλάσης. Η μετρική αυτή συνοψίζει την ικανότητα του μοντέλου να επιστρέφει ως αποτελέσματα δείγματα συναφή με την συγκεκριμένη κλάση.

  Precision = $\frac{tp}{tp+fp}$
- Ανάκληση - Recall : Είναι ο λόγος των σωστών προβλέψεων μιας κλάσης προς το σύνολο των δειγμάτων που πραγματικά ανήκουν σε αυτήν.

  Recall = $\frac{tp}{tp+fn}$
- F1 Score: : Αποτελεί τον αρμονικό μέσο των μετρήσεων :Precision και Recall.

  F1 score = $\frac{2Precision*Recall}{Precision+Recall}$

Κεφάλαιο 3

Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA)

3.1 Τι είναι η EDA;

Η Διερευνητική ανάλυση δεδομένων ή EDA είναι μέρος της διαδικασίας της επιστήμης δεδομένων. Αφορά τη διαδικασία διεξαγωγής αρχικών ερευνών σε δεδομένα με σκοπό να ανακαλύπτονται χαρακτηριστικά και μοτίβα, να εντοπίζονται ανωμαλίες και να ελέγχονται υποθέσεις με τη βοήθεια περιγραφικών στατιστικών (descriptive statistics) και γραφικών αναπαραστάσεων. Η EDA είναι πολύ σημαντική πριν από την μηχανική χαρακτηριστικών (feature engineering) και την μοντελοποίηση (modeling), επειδή σε αυτό το στάδιο πρέπει πρώτα να κατανοήσουμε το περιεχόμενων των δεδομένων όπως την κατανομή τους, τη συχνότητα, τη συσχέτιση και άλλα. Στην πράξη, η περιέργεια είναι πολύ σημαντική σε αυτή τη διαδικασία [34]. Τέλος Η EDA δεν καθοδηγείται από την επιθυμία να επιβεβαιώσει την παρουσία ενός συγκεκριμένου αποτελέσματος και δεν υποστηρίζεται από ένα στατιστικό μοντέλο που ενσωματώνει μια μαθηματική έκφραση για ένα τέτοιο αποτέλεσμα. [35].

Μία από τις βασικές αναφορές για την EDA είναι το περιεκτικό βιβλίο του Tukey [36]. Ο Tukey, στην εργασία του, είχε ως στόχο να δημιουργήσει ένα πλαίσιο ανάλυσης δεδομένων όπου η οπτική εξέταση συνόλων δεδομένων, μέσω στατιστικά σημαντικών αναπαραστάσεων, παίζει τον κεντρικό ρόλο για να βοηθήσει τον αναλυτή να διατυπώσει υποθέσεις που θα μπορούσαν να δοκιμαστούν σε νέα σύνολα δεδομένων. Η έμφαση σε δύο έννοιες, όπως ο δυναμικός πειραματισμός σε δεδομένα (π.χ. αξιολόγηση των αποτελεσμάτων σε διαφορετικά υποσύνολα του ίδιου συνόλου δεδομένων, υπό διαφορετικές συνθήκες προεπεξεργασίας δεδομένων) και οι εξαντλητικές δυνατότητες οπτικοποίησης προσφέρουν στους ερευνητές τη δυνατότητα να εντοπίσουν ακραίες τιμές, τάσεις και μοτίβα στα δεδομένα, πάνω στις οποίες μπορούν να οικοδομηθούν νέες θεωρίες και υποθέσεις

3.1.1 Χαρακτηριστικά της EDA

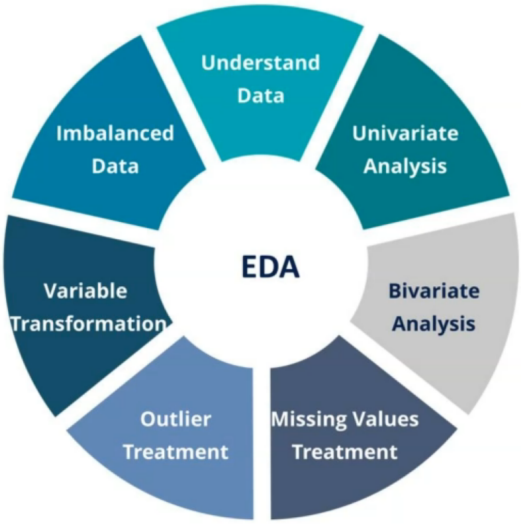
Η EDA σαν διεργασία διέπεται απο συγκεκριμένα χαρακτηριστικά.

- Κατανόηση των δεδομένων.
- Διαχείρηση μη ισορροπημένων δεδομένων

- Διαχείρηση ακραίων τιμών
- Μετασχηματισμός μεταβλητών
- Έλεγχος υποθέσεων
- Έλεγχος Multicollinearity

Γενικά, η EDA πραγματοποιείται με διάφορες μεθόδους:

- **Μονομεταβλητή ανάλυση** : Περιγραφική ανάλυση με μία μεταβλητή
- **Ανάλυση δύο μεταβλητών** : Ανάλυση σχέσης με δύο μεταβλητές
- **Ανάλυση πολλών μεταβλητών** : Ανάλυση χρησιμοποιώντας περισσότερες από ή ίσες με τρεις μεταβλητές



Σχήμα 3.1: *parts of EDA*

3.2 Περιγραφική στατιστική

Όλα εκείνα τα εργαλεία οπτικοποίησης που επιτρέπουν την εξερεύνηση μονο- και ολιγοβαρών δεδομένων μπορούν να θεωρηθούν ως όργανα περιγραφικής στατιστικής [37]. Η περιγραφική στατιστική ορίζεται συνήθως ως ένας τρόπος σύνοψης/εξαγωγής πληροφοριών από μία ή λίγες μεταβλητές: σε σύγκριση με τις στατιστικές συμπερασμάτων, των οποίων ο στόχος είναι η αξιολόγηση της εγκυρότητας μιας υπόθεσης που γίνεται σε δεδομένα μέτρησης. Η περιγραφική στατιστική είναι απλώς διερευνητική. Συγκεκριμένα, ορισμένα σημαντικά γεγονότα μπορούν να εξαχθούν σχετικά με μια μεταβλητή:

- Μέτρο της κεντρικής τάσης, δηλαδή της κεντρικής θέσης μιας κατανομής συχνότητας μιας ομάδας δεδομένων. Με άλλα λόγια, ένας αριθμός που είναι καλύτερα κατάλληλος για να αντιπροσωπεύει την τιμή των δειγμάτων που ερευνήθηκαν σε σχέση με τη μετρούμενη ιδιότητα. Τα τυπικά στατιστικά στοιχεία είναι ο μέσος όρος, ο διάμεσος και ο τρόπος λειτουργίας.

- Ένα μέτρο εξάπλωσης, που περιγράφει πόσο διασκορπισμένες είναι οι μετρούμενες τιμές γύρω από την κεντρική τιμή. Τα τυπικά στατιστικά στοιχεία είναι το εύρος, τα τεταρτημόρια και η τυπική απόκλιση.

- Ένα μέτρο ασυμμετρίας (λοξής) και αιχμής (κύρτωσης) της κατανομής συχνότητας, δηλαδή εάν η εξάπλωση των δεδομένων γύρω από την κεντρική τιμή είναι συμμετρική και στις δύο κατευθύνσεις αριστερά/δεξιά και πόσο έντονη/επίπεδη είναι η κατανομή στην κεντρική θέση , αντίστοιχα.

Αν και είναι χρήσιμα, αυτά τα στατιστικά στοιχεία είναι μόνο μια σύνοψη δεδομένων και δεν προσφέρουν άμεσο πλεονέκτημα ερμηνείας σε σύγκριση με μια γραφική αναπαράσταση των δεδομένων. Τα κύρια γραφικά εργαλεία για περιγραφικές στατιστικές είναι τα ιστογράμματα συχνότητας, τα γραφήματα κουτιού και τα διαγράμματα διασποράς. Αυτά τα εργαλεία είναι χρήσιμα για την επιθεώρηση των στατιστικών που αναφέρθηκαν προηγουμένως, την παρουσία ακραίων τιμών και πολλαπλών καταστάσεων στα δεδομένα (ιστογράμματα), για την επισήμανση των αλλαγών τοποθεσίας και διακύμανσης μεταξύ διαφορετικών ομάδων δεδομένων ή μεταξύ πολλών μεταβλητών (box-whisker), για την αποκάλυψη σχέσεων ή συσχετίσεις μεταξύ δύο μεταβλητών (διαγράμματα διασποράς), καθώς και για να τονίσει την εξάρτηση σε σχέση με κάποιο κριτήριο ταξινόμησης, όπως σειρά εκτέλεσης, χρόνος, θέση κ.λπ.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 4

Διερευνητική Ανάλυση στο Σύνολο των Δεδομένων

Στο κεφάλαιο αυτό παρουσιάζεται το αρχικό μέρος της ΔΑΔ που αφορά την συλλογή, την κατανόηση και ύστερα τον καθαρισμό και την προεπεξεργασία των δεδομένων πτήσεων και μετεωρολογικών στοιχείων.

4.1 Συλλογή Δεδομένων

Στα πλαίσια διεξαγωγής της παρούσας εργασίας αρχικά είχαμε δεδομένα για εγχώριες πτήσεις με τα αντίστοιχα μετεωρολογικά στοιχεία τους ωστόσο σε ένα σημείο αυτά χάθηκαν για λόγους που έχουν να κάνουν με τον σερερ της σχολής. Η κατεύθυνση λοιπόν που ακολουθήσαμε ύστερα ήταν η αναζήτηση ανοιχτών δεδομένων στο διαδίκτυο. Τα τελικά δεδομένα που χρησιμοπιήθηκαν λοιπόν είναι αυτά από το Bureau of Transportation Statistics (BTS) του υπουργείου Μεταφορών των ΗΠΑ. https://www.bts.gov/. Μέσω της πλατφόρμας δεδομένων του BTS υπάρχει η δυνατότητα επιλογής των χαρακτηριστικών που επιθυμεί ο χρήστης να συλλέξει. Επιλέγηκαν όσα χαρακτηριστικά είχαν να κάνουν με ώρα και ημερομηνία της πτήσης, αεροδρόμιο αναχώρησης, προγραμματισμένες και πραγματικές ώρες αναχώρησης, απόσταση, χρόνο στον αέρα, συνολικά λεπτά καθυστέρησης και λεπτά καθυστέρησης ανά κατηγορία καθυστέρησης. Το dataset περιέχει έγκυρα στοιχεία που αναφέρθηκαν από πιστοποιημένους αερομεταφορείς των ΗΠΑ που αντιπροσωπεύουν τουλάχιστον το ένα τοις εκατό των εσόδων τακτικών επιβατών εσωτερικού. Τα τελικά δεδομένα είναι μια μίξη των δεδομένων "Reporting Carrier On-Time Performance" του 2019 και των δεδομένων καιρού για αεροδρόμια για το 2019 που βρίκαμε από το National Centers for Environmental Information https://www.ncdc.noaa.gov/cdo-web/datasets

4.2 Περιγραφή Δεδομένων

Στην ενότητα αυτή περιγράφονται τα στοιχεία των dataset για τις αεροπορικές πτήσεις και τις καιρικές συνθήκες.

Πτήσεις

Όσον αφορά τις πτήσεις έχουμε:

- 'YEAR', 'MONTH', 'DAY-OF-MONTH', 'DAY-OF-WEEK' στοιχεία που αφορούν τον χρόνο που διεξάγεται η πτήση

- 'AIRLINE', 'ORIGIN-AIRPORT-ID','ORIGIN','ORIGIN-CITY-NAME','DEST-AIRPORT-ID', 'DEST','DEST-CITY-NAME' στοιχεία που αφορούν την αεροπορική εταιρεία που διεξάγει την πτήση, την προέλευση και τον προορισμό

- 'SCHEDULED-DEP-TIME', 'DEP-TIME', 'DEP-DELAY', 'SCHEDULED-ARR-TIME','ARR-TIME', 'ARR-DELAY', 'SCHEDULED-ELAPSED-TIME', 'ACTUAL-ELAPSED-TIME' στοιχεία που αφορούν τον προβλεπόμενο και τον πραγματικό χρόνο αναχώρησης και άφιξης και τις καθυστερήσεις

- 'WHEELS-OFF', 'WHEELS-ON', στοιχεία που αφορούν τη χρονική στιγμή όπου το αεροπλάνο αφήνει τις ρόδες του απο το έδαφος (διαφέρει απο τη στιγμή αναχώρησης κατα ελάχιστο) και τη στιγμή που οι ρόδες του αεροπλάνου ακουμπούν το έδαφος κατα την προσγείωση

- 'TAXI-OUT', 'TAXI-IN', στοιχεία που αφορούν τη χρονική διάρκεια μεταξύ της αναχώρησης και του wheels-off δηλαδή στοιχεία που αναφέρονται στο χρόνο εδάφους

Τέλος έχουμε και πληροφορία για τις κατηγορίες καθυστερησης. Αυτές διακρίνονται σε πέντε διαφορετικές:

- CARRIER-DELAY = Delay caused by the airline in minutes. Η αιτία της ακύρωσης ή καθυστέρησης οφείλεται σε παράγοντες που βρίσκονται υπό τον έλεγχο του αερομεταφορέα (π.χ. προβλήματα συντήρησης ή πληρώματος, καθαρισμός αεροσκαφών, φόρτωση αποσκευών, τροφοδοσία κ.λπ.).

- WEATHER-DELAY = Delay caused by weather. Ακραίες μετεωρολογικές συνθήκες (πραγματικές ή προβλέψεις) που κατά την κρίση του αερομεταφορέα, καθυστερούν ή αποτρέπουν τη λειτουργία πτήσης όπως ανεμοστρόβιλοι, χιονοθύελλες ή τυφώνες.

- NAS-DELAY = Delay caused by air system. Καθυστερήσεις και ακυρώσεις που οφείλονται στο εθνικό αεροπορικό σύστημα που αναφέρονται σε ένα ευρύ φάσμα συνθηκών, όπως μη ακραίες καιρικές συνθήκες, λειτουργίες αεροδρομίου, βαριά κυκλοφορία και έλεγχος εναέριας κυκλοφορίας.

- SECURITY-DELAY = caused by security reasons. Καθυστερήσεις ή ακυρώσεις που προκαλούνται από εκκένωση τερματικού σταθμού ή χώρου του αεροδρομίου, επανεπιβίβαση σε αεροσκάφος λόγω παραβίασης ασφάλειας, μη λειτουργικού εξοπλισμού ελέγχου και / ή μεγάλων γραμμών που υπερβαίνουν τα 29 λεπτά σε περιοχές ελέγχου.

- LATE-AIRCRAFT-DELAY = Delay caused by security. Μια προηγούμενη πτήση με το ίδιο αεροσκάφος αφίχθει αργά, με αποτέλεσμα την αργοπορημένη αναχώρηση της παρούσας πτήσης.

Καιρικές Συνθήκες

Όσον αφορά τις καιρικές συνθήκες έχουμε :

- DATE στοιχείο που αφορά τον χρόνο που διεξάγεται η πτήση
- PRCP: Precipitation (i.e., rain, melted snow) Στην ουσία είναι ο κωδικός κακοκαιρίας
- SNOW: Snowfall
- SNWD: Snow depth
- TMAX: Daily maximum temperature
- AWND: Average daily wind speed
- ORIGIN-AIRPORT-ID
- AIRPORT-NAME

Σημειώνεται επίσης, πως δεδομένα διατίθενται σε αρχεία με comma separated values (.csv).

4.3 Εργαλεία Ανάλυσης

Υπάρχουν πολλά παραδείγματα λογισμικού ανοιχτού κώδικα, εν μέρει στην Python, που διευκολύνουν τους αναλυτές να εκπαιδεύουν, να δοκιμάζουν και να επικυρώνουν μοντέλα. Παραδείγματος χάριν το Scikit-learn είναι μια εργαλειοθήκη της Python που μπορεί να χρησιμοποιηθεί για να συγκρίνει την εγκυρότητα και την αξιοπιστία των μοντέλων πρόβλεψης και να επιλέξει τα με την καλύτερη απόδοση[17]. Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε σε ολόκληρη την εργασία είναι η Python3. Η γλώσσα Πψτηον σε συνδυασμό με συγκεκριμένες βιβλιοθήκες χρησιμοποιήθηκε τόσο για την προεπεξεργασία, καθαρισμό και συγχώνευση των δεδομένων, όσο και για τη δημιουργία και εκπαίδευση μοντέλων Μηχανικής Μάθησης. Όλα τα πειράματα εκτελέστηκαν σε Cloud service , Google Colab. Η βιβλιοθήκη που χρησιμοποιήθηκε κατα κόρον είναι η Pandas η οποία προσφέρει μεθόδους για εύκολη διαχείριση και οργάνωση μεγάλου όγκου δεδομένων. Η βιβλιοθήκη NumPy χρησιμοποιείται για πράξεις μεταξύ πινάκων. Διαθέτει πληθώρα συναρτήσεων στο πεδίο της γραμμικής άλγεβρας, μετασχηματισμούς κ.α. Η Matplotlib είναι μια εύχρηστη βιβλιοθήκη απεικόνισης (plotting library), η οποία διαθέτει συναρτήσεις απεικόνισης γραφημάτων αλλά και εικόνων. Η Scipy περιέχει επιπλέον επιστημονικές μεθόδους. Η βιβλιοθήκη Scikit-Learn είναι μία βιβλιοθήκη ανοιχτού κώδικα (open source), η οποία χρησιμοποιήθηκε σε διάφορα στάδια της εκπαίδευσης αλγορίθμων ML. Παραδείγματα της χρήσης της είναι η χρήση της train-test-split για τον διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης (train test) και ελέγχου (test set). Περιέχει μεθόδους κανονικοποίησης δεδομένων . Η σημαντικότερη συνεισφορά της βιβλιοθήκης είναι βέβαια η πληθώρα των έτοιμων ταξινομητών που διαθέτει τους οποίους χρησιμοποιήσαμε και εκπαιδεύσαμε για την δημιουργία προβλέψεων για το πρόβλημα μας. Επιπλέον χρησιμοποιήθηκε η βιβλιοθήκη XGBoost και ADABoost για την υλοποίηση και εκπαίδευση των αντίστοιχων μοντέλων.

4.4 Καθαρισμός Δεδομένων

Σε αυτην την ενότητα ασχολούμαστε με τον καθαρισμό των δεδομένων και την προεπεργασία τους ώστε να είναι έτοιμα για ανάλυση.

Μία πτήση όπως ορίζεται απο το Υπουργείο Μεταφορών των ΗΠΑ είναι καθυστερημένη εάν είχε προγραμματιστεί να φτάσει σε κάποια ώρα και η προγραματική άφιξη είναι τουλάχιστον δεκαπέντε λεπτά αργότερα.

Αρχικά θα πρέπει να αφαιρέσουμε την πληροφορία που μας είναι περιττή. Θα ασχοληθούμε με πτήσεις που ειναι delayed , όχι με τις cancelled, άρα δεν χρειαζόμαστε να συμπεριλάβουμε τις cancelled στο dataset. Μία canceled πτήση είναι μη καθυστερημένη εξ ορισμού αφού δεν έχει φύγει από ούτε φτάσει σε κάποιο αεροδρόμιο. Ο λόγος που είναι cancelled πάλι δεν μας ενδιαφέρει στα πλαίσια της εργασίας.

Έπειτα διαχειριζόμαστε τις diverted πτήσεις. Σε μία diverted πτήση το αεροπλάνο έχει προσγειωθεί σε διαφορετικό αεροδρόμιο απο αυτό που ήταν αρχικά προγραμματισμένο. Αυτό φυσικά είναι προσωρινό και η αεροπορική εταιρεία έχει την ευθύνη να μεταφέρει τους επιβάτες στον τελικό τους προορισμό χωρίς επιπλέον κόστος. Συνεπώς δεν χρειάζεται να αφαιρέσουμε τις diverted πτήσεις αλλά μόνο την στήλη αφού αποτελεί πρόσθετη πληροφορία που δεν χρειαζόματε.

Τέλος θέλουμε να διαχειριστούμε τις στήλες από τα δεδομένα μας που έχουν NAN values, δηλαδή τιμές απροσδιόριστες που δεν γνωρίζουμε εκ των προτέρων. Υπάρχουν πτήσεις ¨φαντάσματα¨ στις οποίες δεν έχουμε τιμές για την ώρα αναχώρησης, ώρα άφιξης, ούτε για τους χρόνους εδάφους (taxi-in , taxi-out). Αποτελούν τιμές που δεν μπορούμε να εφαρμόσουμε κάποια τεχνική filling για να προσθέσουμε δικές μας τιμές χωρις να αλλοιώσουμε τα στατιστικά στοιχεία όλου του δείγματος. Συνεπώς αφαιρούνται.

4.4.1 Συγχώνευση Δεδομένων

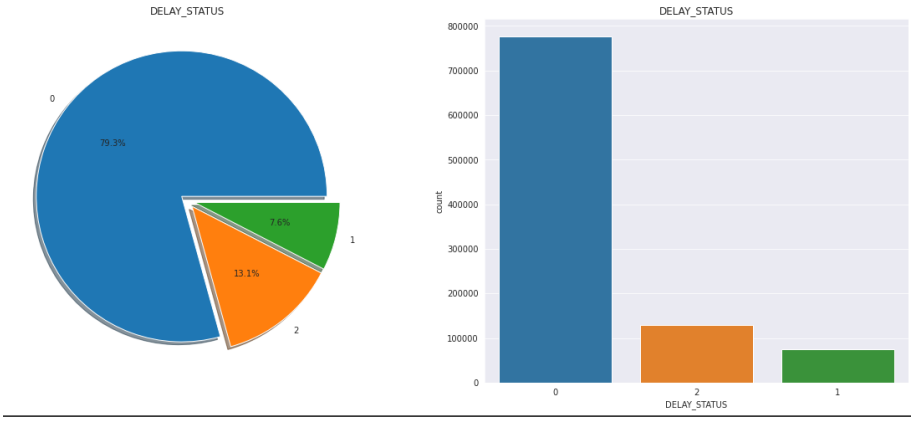
Για την πιο αποτελεσματική εκπαίδευση των μοντέλων είναι απαραίτητο να συγχωνευσουμε διαθέσιμη πληροφορία για τον καιρό. Έχοντας καθαρίσει τα 2 σύνολα δεδομένων που αφορούν τις πτήσεις και τον καιρό προχωράμε στη συγχώνευση τους σε ένα σύνολο δεδομένων, στο οποίο κάθε καταχώρηση φέρει πληροφορία για όλα στα στοιχεία μιας πτήσης και επιπλέον πληροφορία για τον καιρό την ώρα άφιξης. Συγχωνεύουμε εσωτερικά (Inner Join) με βάση το Origin-Airport-ID και την ημερομηνία της πτήσης.

4.5 Διερευνητική Ανάλυση Δεδομένων - (EDA)

Ξεκινώντας την EDA θέλουμε να ερευνήσουμε τα δεδομένα με σκοπό να βρούμε ενδιαφέροντα χαρακτηριστικά και μοτίβα και ύστερα να εκτελέσουμε διαδικασίες προεπεξεργασίας. Ως δεδομένα θα θεωρήσουμε τώρα τις πρώτες 500000 σειρές του dataset για διευκόλυνση στο χρόνο εκτέλεσης.

Πτήσεις στην ώρα τους ή επαρκώς καθυστερημένες :

Στο παρακάτω Σχήμα 4.1 μπορούμε να πάρουμε μια εικόνα για την ποσόστωση των πτήσεων που έχουν φθάσει στην ώρα τους (status 0), τις πτήσεις που άργησαν το πολύ 30 λεπτά(status 1), και τις πτήσεις που άργησαν πάνω απο 30 λεπτά(status 2)

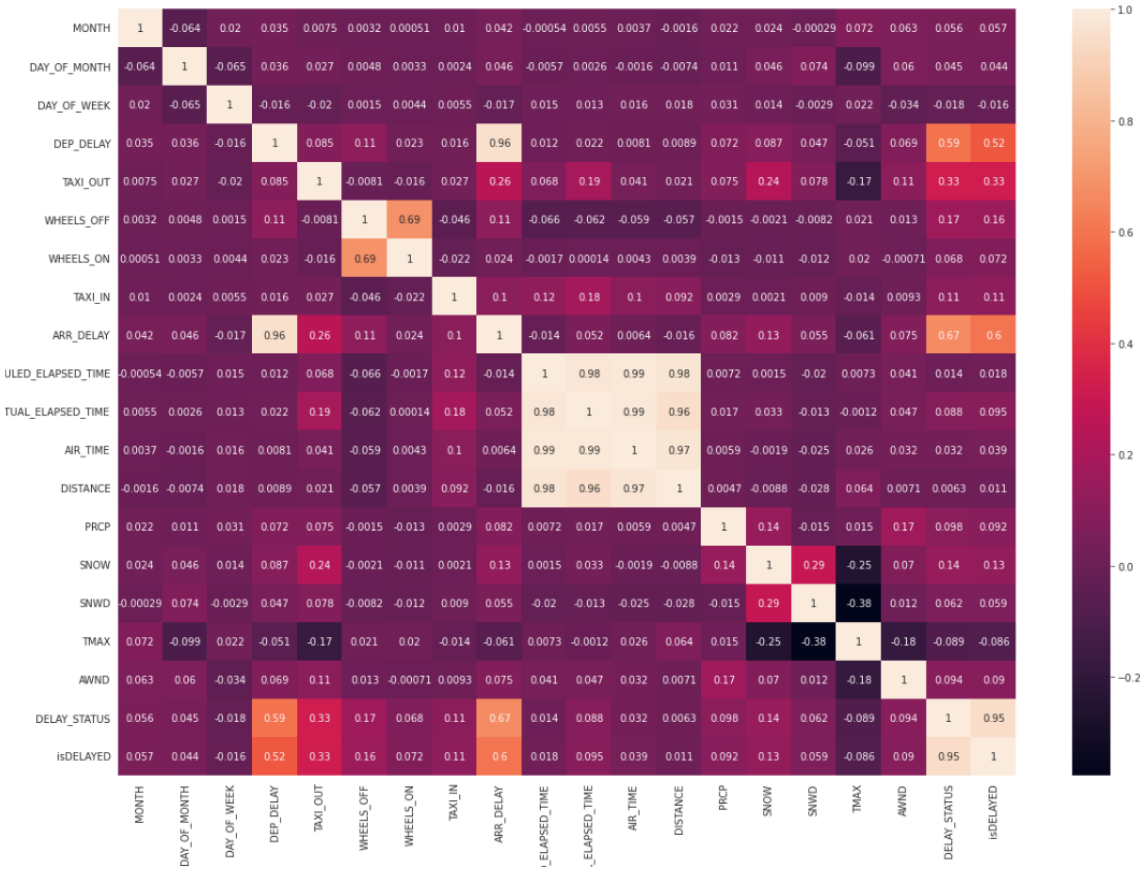


Σχήμα 4.1: *Flights on time (status 0), Flights slightly delayed(status 1), Flights highly delayed status 2)*

Βαθμός συσχέτισεις χαρακτηριστικών του συνόλου δεδομένων

Μπορούμε να δούμε το βαθμό συσχέτισης των χαρακτηριστικών του dataset για να κρίνουμε την σημαντικότητα τους στην ανάλυση μας και αργότερα στο αλγόριθμο ML Η συσχέτιση μετρά την ισχύ και την κατεύθυνση μεταξύ δύο στηλών σε ένα σύνολο δεδομένων. Η συσχέτιση χρησιμοποιείται συχνά για να βρεθεί η σχέση μεταξύ ενός χαρακτηριστικού και του στοιχείου-στόχου, που για μας εδώ στόχος είναι η δυαδική μεταβλητή που δείχνει αν η πτήση είναι καθυστερημένη ή οχι.

Με βάση το παρακάτω πίνακα μπορούμε να εξάγουμε πολλά συμπεράσματα



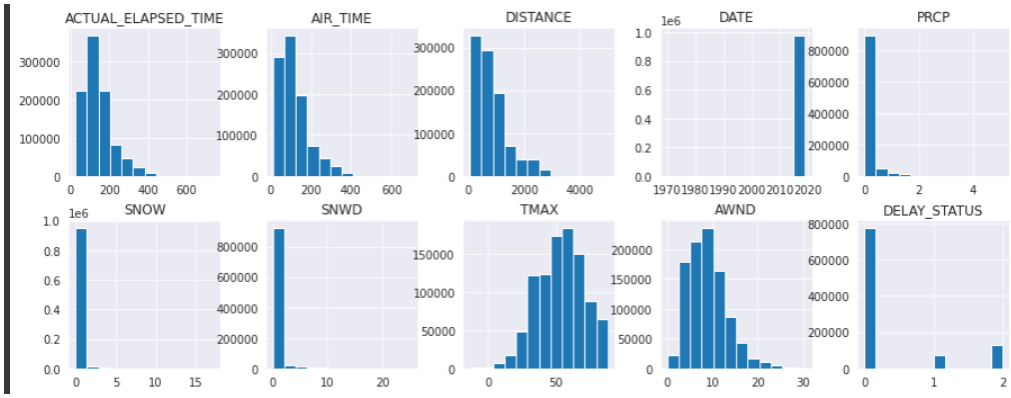
Σχήμα 4.2: *Correlation Matrix for all the features*

Τα χαρακτηριστικά που λείπουν είναι οι κατηγορικές στήλες. Αν θεωρήσουμε ένα όριο στο 50 τοις εκατο στον πίνακα συσχέτισης, φαίνεται ότι έχουμε συσχέτιση στα ακόλουθα:

ARR-DELAY με DEP-DELAY, WHEELS-ON με WHEELS-OFF ,SCHEDULED-ELAPSED-TIME με ACTUAL-ELAPSED-TIME, AIR-TIME, DISTANCE και DELAY-STATUS με is-DELAYED

Υπάρχει λοιπόν μια πιθανή πολυσυγγραμμικότητα με αυτές τις σχέσεις.

Η κατανομή των καιρικών συνθηκών είναι η ακόλουθη:



Σχήμα 4.3: *Histograms for all the weather features*

4.5.1 Univariate Analysis

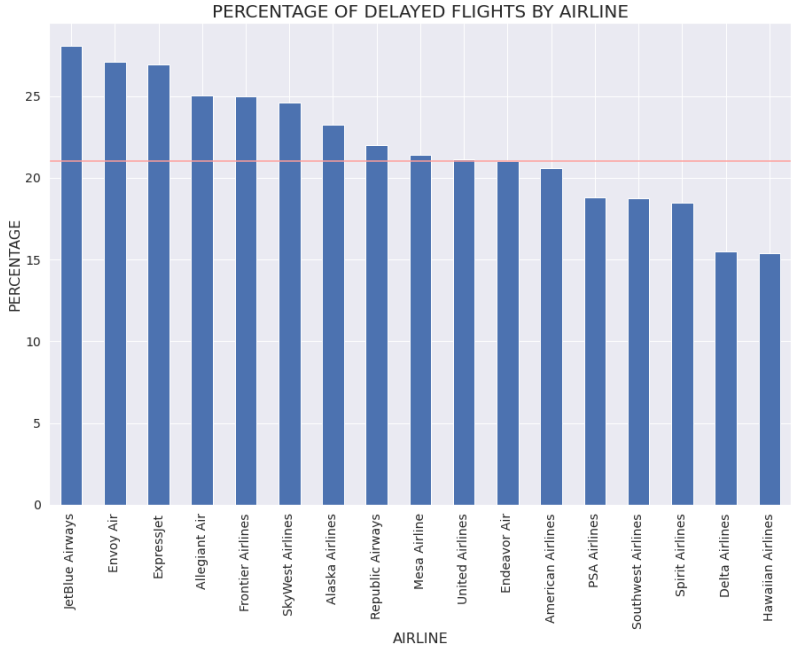
Μονομεταβλητή Περιγραφική Στατιστική αφορά τους τρόπους που μπορούμε να περιγράψουμε μοτίβα που βρίσκονται σε μονομεταβλητά δεδομένα. Περιλαμβάνουν την κεντρική τάση (μέσος όρος, λειτουργία και διάμεσος) και διασπορά : εύρος , διακύμανση, μέγιστο, ελάχιστο, τεταρτημόρια (συμπεριλαμβανομένου του διατεταρτημορίου εύρους) και τυπική απόκλιση.

Υπάρχουν πολλές επιλογές για την περιγραφή δεδομένων με μονομεταβλητά δεδομένα. Αναφορικά :

- Πίνακες κατανομής συχνότητας.
- Ραβδογράμματα.
- Ιστογράμματα.
- Πολύγωνα Συχνότητας.
- Διαγράμματα πίτας.

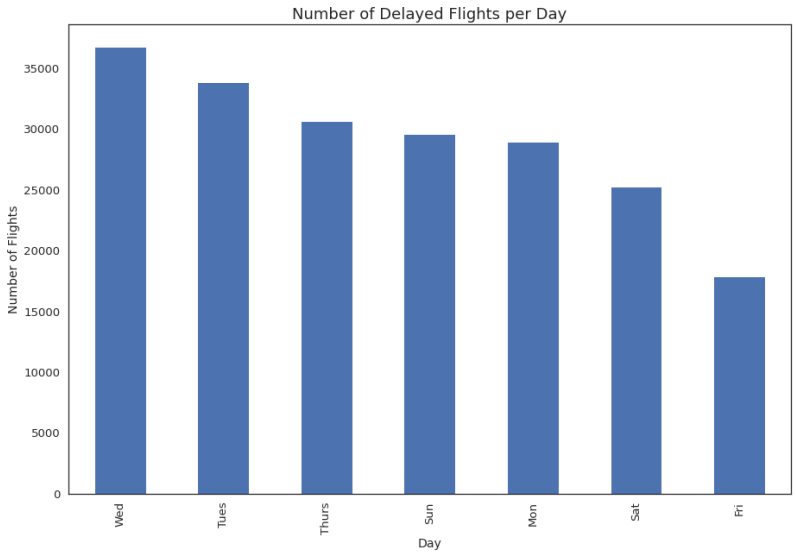
Ποσοστό καθυστερημένων πτήσεων ανά αεροπορική εταιρεία :

Το πρώτο πράγμα που πρέπει να υπολογίσουμε είναι το συνολικό ποσοστό των καθυστερημένων πτήσεων για να μπορούμε να κάνουμε σωστές συγκρίσεις με κάθε αεροπορική εταιρεία. Αυτό μπορούμε να το δούμε στο Σχήμα 4.4. Οι αεροπορικές εταιρείες θα ήθελαν να είναι κάτω από το μέσο ποσοστό καθυστερημένων πτήσεων για να βρίσκονται σε αποδεκτή θέση, έτσι ώστε ο μαγικός αριθμός να αντιπροσωπεύει το κατώφλι μας. Οι αεροπορικές εταιρείες που βρίσκονται πάνω από την κόκκινη γραμμή (21%) θα ήταν κατά κοινή λογική αυτές που οι ταξιδιώτες θα ήθελαν να αποφύγουν, καθώς σημαίνει ότι θα έχουν τις περισσότερες καθυστερήσεις.

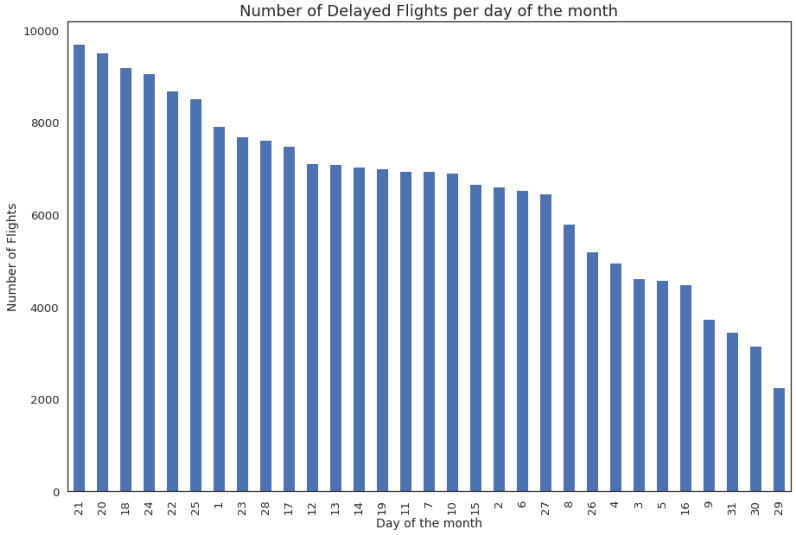


Σχήμα 4.4: *Percentage of delayed Flights*

Ας εξετάσουμε ποιες είναι οι χειρότερες και καλύτερες μέρες για ταξίδι με βάση τις καθυστερήσεις πτήσεων:

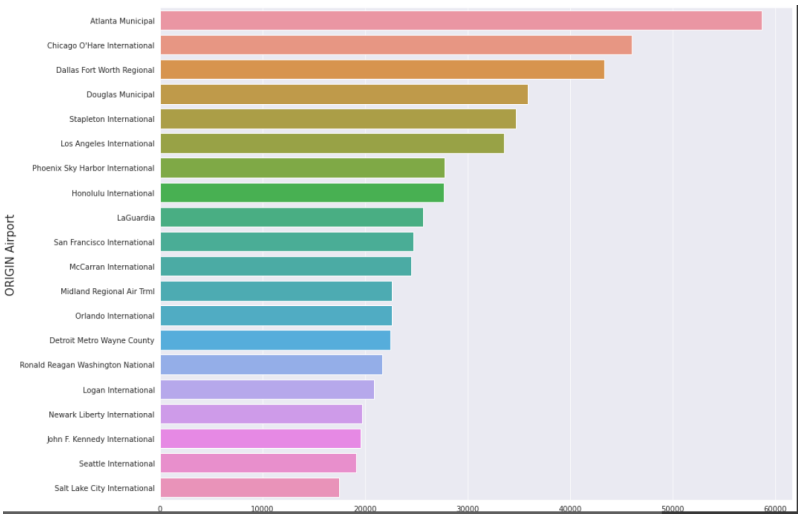


Σχήμα 4.5: *Delayed Flights per day*

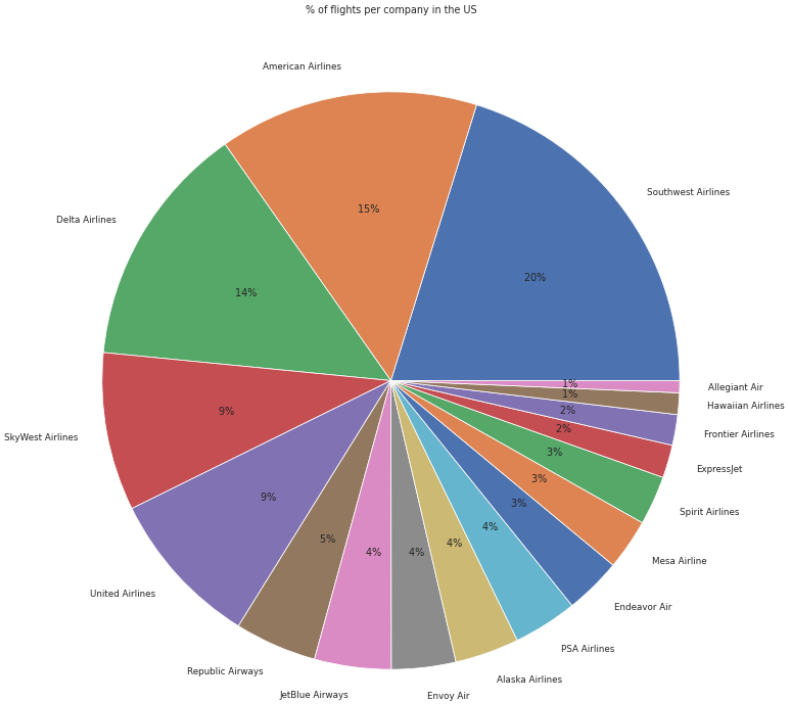


Σχήμα 4.6: *Delayed Flights per month*

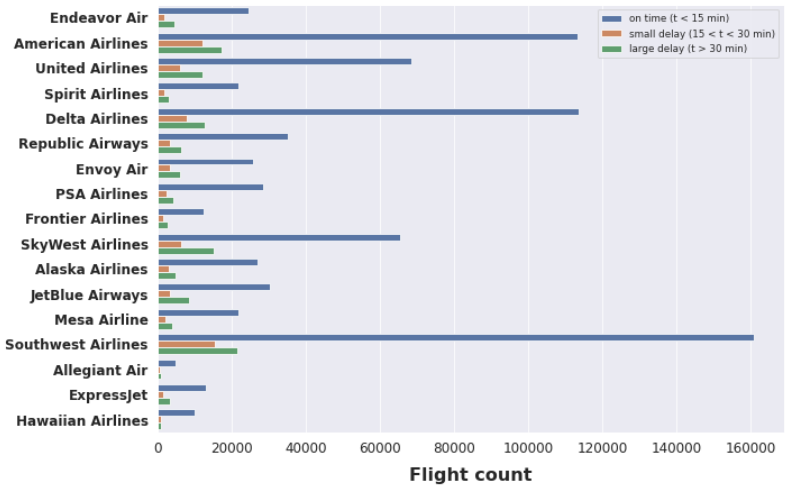
Ας εξετάσουμε τα αεροδρόμια απο τα οποία ξεκινούν οι πτήσεις (origin Airports). Βλέπουμε στο Σχήμα 4.7 για την κατανομή των καιρικών συνθηκών ότι το Atlanta Municipal και το Chicago O'Hare International έχει τις περισσότερες πτήσεις.



Σχήμα 4.7: *Origin Airports count plots*

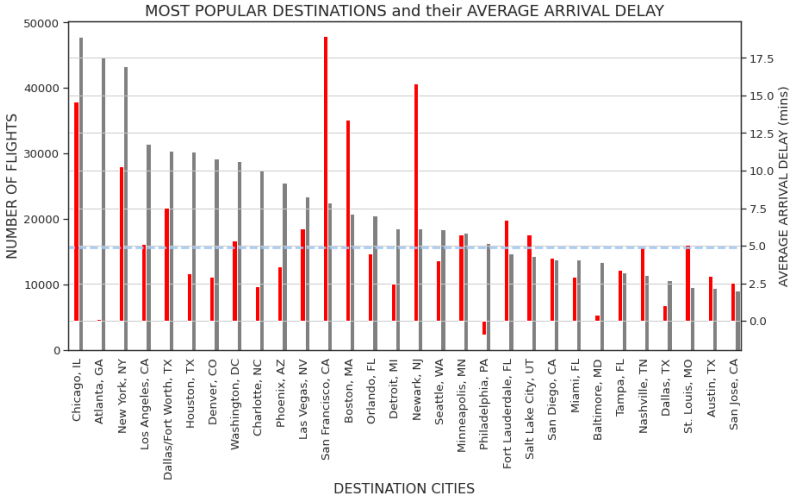


Σχήμα 4.8: *airlines piechart*



Σχήμα 4.9: *airlines countplots*

Τέλος θα ερευνήσουμε το πως η επιλογή κάποιου προορισμου μπορεί να επηρεάσει τις καθυστερήσεις. Στο σχήμα 4.10 βλέπουμε λοιπόν τους πιο δημοφιλής προορισμούς μαζί με τις καθυστερήσεις των πτήσεων.



Σχήμα 4.10: *Most Popular Destinations and their Average Arrival Delays*

4.5.2 Bivariate Analysis

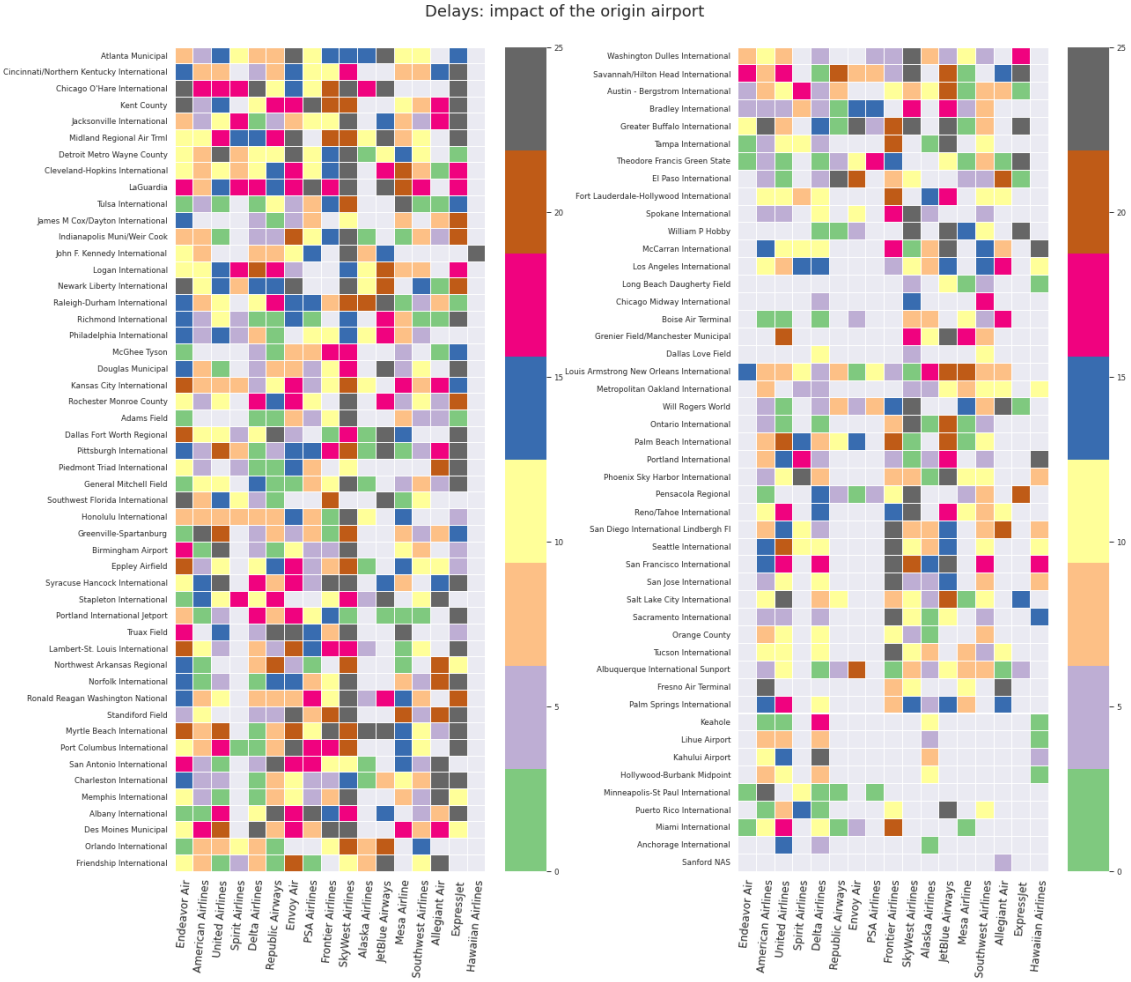
Ονομάζεται η ανάλυση διμεταβλητών δεδομένων. Είναι μια από τις απλούστερες μορφές στατιστικής ανάλυσης, που χρησιμοποιείται για να διαπιστωθεί εάν υπάρχει σχέση μεταξύ δύο συνόλων τιμών.

Σε αυτήν την ενότητα, θα ερευνήσουμε τις παραλλαγές των καθυστερήσεων σε σχέση με το αεροδρόμιο προέλευσης και για κάθε αεροπορική εταιρεία. Το πρώτο βήμα λοιπόν συνίσταται στον προσδιορισμό του μέσου όρου των καθυστερήσεων ανά αεροδρόμιο. Ύστερα με βάση αυτούς τους μέσους όρους θα εμφανίσουμε ένα heatmap (βλέπε Σχήμα 4.11 για την καλύτερη αναπαράσταση του τελικού αποτελέσματος.

Αυτό το σχήμα επιτρέπει την εξαγωγή ορισμένων συμπερασμάτων. Πρώτον, εξετάζοντας τα δεδομένα που σχετίζονται με τις διάφορες αεροπορικές εταιρείες, βρίσκουμε τη συμπεριφορά που παρατηρήσαμε προηγουμένως: για παράδειγμα, αν εξετάσουμε το αριστερά πλαίσιο, θα φανεί ότι η στήλη που σχετίζεται με τις Skywest Airlines και ExpressJet αναφέρει ως επί το πλείστον μεγάλες καθυστερήσεις, ενώ η στήλη που σχετίζεται με την Ρεπυβλις Αιρωαψ σχετίζεται κυρίως με καθυστερήσεις μικρότερες των 5 λεπτών.

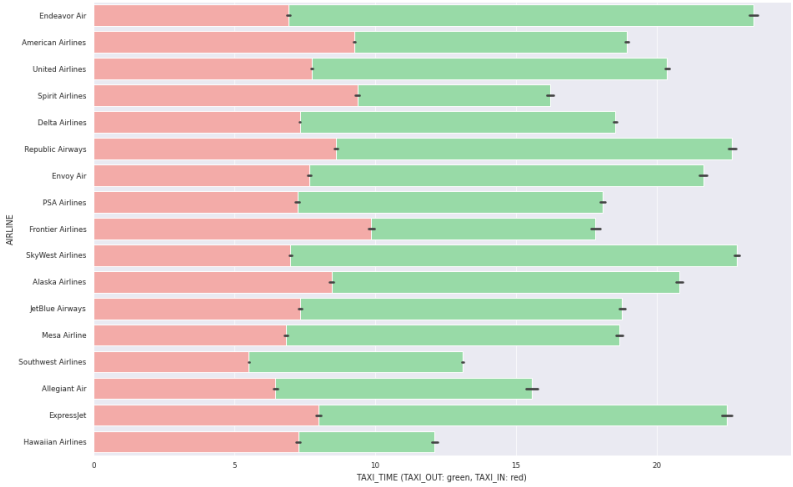
Τέλος, μπορούμε να συμπεράνουμε από αυτές τις παρατηρήσεις ότι υπάρχει μεγάλη μεταβλητότητα στις μέσες καθυστερήσεις, τόσο μεταξύ των διαφορετικών αεροδρομίων όσο και μεταξύ των διαφορετικών αεροπορικών εταιρειών. Αυτό είναι σημαντικό γιατί υπονοεί ότι για να μοντελοποιηθούν με ακρίβεια οι καθυστερήσεις, θα είναι απαραίτητο να υιοθετηθεί ένα μοντέλο που να είναι συγκεκριμένο για την εταιρεία και το αεροδρόμιο της χώρας προέλευσης.

Ενδιαφέρον παρουσιάζει η ανάλυση των χρόνων εδάφους Taxi-In και Taxi-out. Αναφέρουμε ξανά πως ως taxi-in νοείται ο χρόνος από τη στιγμή που οι ρόδες του αεροπλάνου ακουμπούν το έδαφος μέχρι να φτάσει στην πύλη του αεροδρομίου και ταξι-out το αντίθετο. Στη γραφική παράσταση 4.12 παρατηρούμε ότι το taxi in είναι πάντα μικρότερο κατά μέση τιμή απο το taxi out. Αυτό σημαίνει πως όταν τα αεροπλάνα αργουν συνήθως θα εξοικονομούν χρόνο κατα την προσγείωση. Πως συμβαίνει όμως αυτο στην πράξη· Πολλά αεροπλάνα επιταχύνουν στο στάδιο του taxi-in δηλαδή αφού οι ρόδες τους ακουμπήσουν το έδαφος και

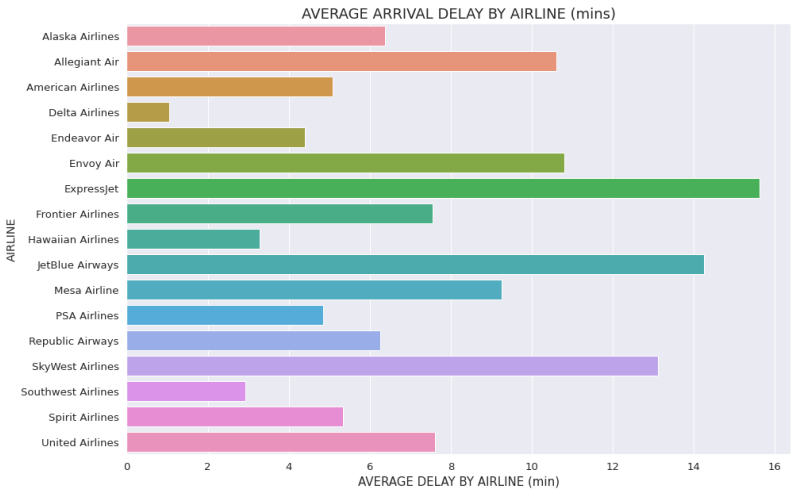


Σχήμα 4.11: *How does the origin airport impact delays? Analysis of Airports and Airlines*

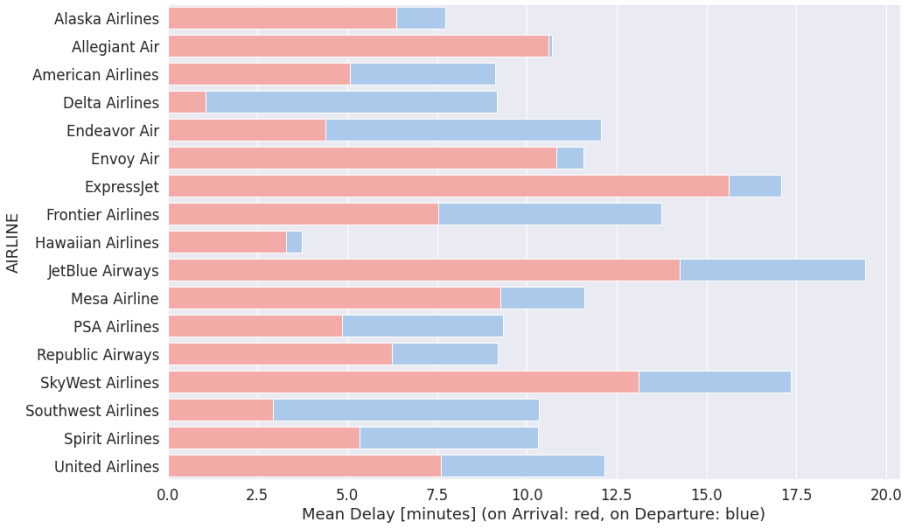
κινηθούν μέχρι την πύλη του αεροδρομίο, ακόμη και σε κατάσταση ρελαντί, και ειδικά με τους δύο κινητήρες σε λειτουργία (τα περισσότερα είναι διπλά αεριωθούμενα). Εάν το αεροσκάφος έχει εγκεκριμένη διαδικασία για taxi-in με μόνο έναν κινητήρα, ο πιλότος μπορεί να επιλέξει να το κάνει και μετά να ξεκινήσει τον δεύτερο κινητήρα κοντά στον διάδρομο προσγείωσης.



Σχήμα 4.12: *Taxi in and Taxi out by Airlines*



Σχήμα 4.13: *Arrival delays by Airline*

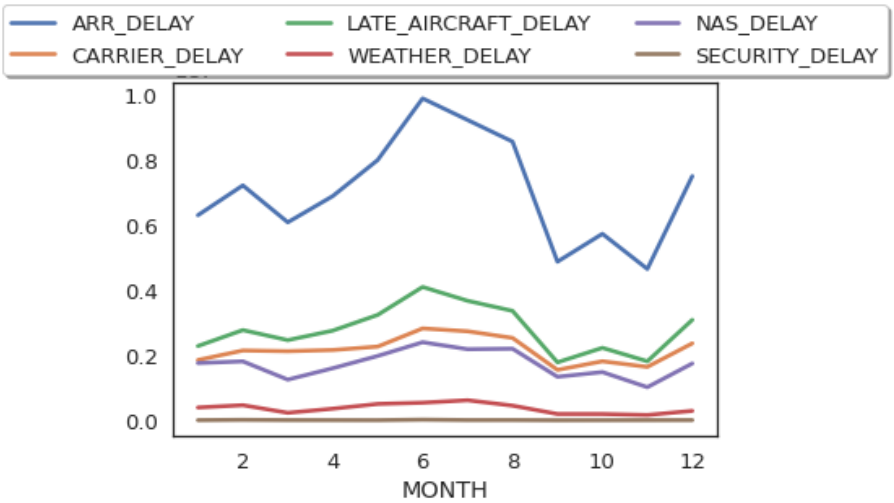


Σχήμα 4.14: *Impact of Delays*

Μπορούμε ακόμα να αναλύσουμε την επίπτωση των καθυστερήσεων (καθυστέρηση αναχώρησης έναντι καθυστέρησης άφιξης). Σε αυτήν την γραφική 4.14, μπορούμε να δούμε ότι οι καθυστερήσεις άφιξης είναι γενικά χαμηλότερες από τις καθυστερήσεις αναχώρησης για την ίδια αεροπορική εταιρεία. Αυτό υποδηλώνει ότι οι αεροπορικές εταιρείες προσαρμόζουν την ταχύτητα της πτήσης τους προκειμένου να μειώσουν τις καθυστερήσεις κατά την άφιξη.

4.6 Exploratory Data Analysis for weather data

Σε αυτήν την ενότητα, θα ερευνήσουμε τις μεταβλητές καιρού που σχετίζονται με τις καθυστερήσεις. Θα δούμε επίσης την κατηγοριοποίηση της συνολικής καθυστέρησης για την εκαστοτε γραμμή σε CARRIER-DELAY, LATE-AIRCRAFT-DELAY, WEATHER-DELAY, NAS-DELAY, SECURITY-DELAY.

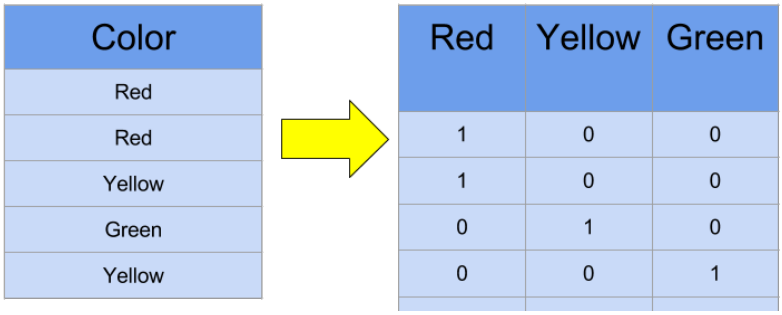


Σχήμα 4.15: *Different kind of delays*

4.7 Τεχνικές Προεπεξεργασίας Δεδομένων για Modeling

4.7.1 One hot Encoding

Για κατηγορικές μεταβλητές η κωδικοποίηση σε ακέραιους δεν είναι αρκετή. Σε αυτή την περίπτωση χρησιμοποιούμε κωδικοποίηση One-Hot κατά την οποία αφαιρείται η κατηγορική μεταβλητή και προστίθεται μια νέα δυαδική μεταβλητή που αναφέρεται σε κάθε μοναδική τιμή της μεταβλητής αυτής. Στην εργασία αυτή τέτοια κωδικοποίηση χρησιμοποιήθηκε για τις αεροπορικές εταιρείες, τα αεροδρόμια Δεν είναι, προφανώς, μια εξαιρετική επιλογή για την κωδικοποίηση κατηγορικών μεταβλητών από την οπτική γωνία της μηχανικής μάθησης. Το πιο προφανές είναι η μεγάλη ποσότητα διαστάσεων που προσθέτει, και είναι γνωστό ότι γενικά μια μικρότερη ποσότητα διαστάσεων είναι καλύτερη. Για παράδειγμα, εάν επρόκειτο να έχουμε μια στήλη που αντιπροσωπεύει μια πολιτεία των ΗΠΑ (π.χ. Καλιφόρνια, Νέα Υόρκη), ένα σύστημα κωδικοποίησης μίας κλήσης θα είχε ως αποτέλεσμα πενήντα επιπλέον διαστάσεις.



Σχήμα 4.16: *One hot encoding example[15]*

4.7.2 Integer encoding

Για τις στήλες που έχουν πληροφορία για χρόνο συνηθως σε κατηγορική μορφή όπως πχ, τον προβλεπόμενο χρόνο αναχώρησης και άφιξης (SCHEDULED-DEP-TIME, SCHEDULED-ARR-TIME) χρησιμοποιούμε μια τεχνική μετατροπης του κατηγορήματος σε ακέραιο. Θα εφαρμοστή κωδικοποίηση One-Hot σε αυτές τις στήλες, τέτοια ώστε να ταξινομείται η ώρα σε 4 χρονικά slices της ημέρας. Επομένως θα χωρίσουμε τον χρόνο σε 4 τέταρτα των ημερών που σημαίνει 6 ώρες το καθένα. Επιπλέον θα εφαρμόσουμε την ίδια μέθοδο, στους μήνες και στις μέρες.

4.8 Υπερδειγματοληψία και Υποδειγματοληψία

Σε ένα σύνολο δεδομένων όταν τα πλήθη δειγμάτων των κλάσεων διαφέρουν σημαντικά μεταξύ τους χρησιμοποιούμε τον όρο Μη-Ισορροπημένο (imbalanced) Dataset. Αν και δεν υπάρχει κάποιος ορισμός, εμπειρικά όταν ο λόγος μεταξύ του αριθμού των δειγμάτων σε ένα σύνολο δεδομένων δύο κλάσεων ξεπερνά τα 2:3 θεωρούμε το σύνολο μη ισορροπημένο. Οι περισσότεροι ταξινομητές τείνουν να εκπαιδεύονται πιο αποτελεσματικά εάν το πλήθος των δειγμάτων στις διαφορετικές κλάσεις είναι ισορροπημένο. Στο σύνολο δεδομένων που έχουμε κατασκευάσει ο λόγος των κλάσεων non-delayed και delayed της στήλης “is-Delayed” είναι 4:1. Αυτό σημαίνει ότι το σύνολο μας δεν είναι ισορροπημένο. Παρόλα αυτά το σύνολο μας αποτελείται από μεγάλο αριθμό δειγμάτων (6558649 δείγματα) με αποτέλεσμα η ισορροπία των κλάσεων στη συγκεκριμένη περίπτωση να μην είναι καθοριστικής σημασίας αφού υπάρχουν αρκετά δεδομένα για αποτελεσματική εκπαίδευση. Υπάρχουν δύο μέθοδοι για να εξισορροπήσουμε τις κλάσεις σε ένα σύνολο δεδομένων. Στη μέθοδο της υπέρδειγματοληψίας (oversampling) επιλέγονται τυχαία δείγματα από την κλάση που υστερεί σε πλήθος δειγμάτων και επαναλαμβάνονται μέχρι να εξισωθεί ο αριθμός των δειγμάτων μεταξύ των κλάσεων. Αντίθετα, στην υποδειγματοληψία (undesrampling), επιλέγεται η κλάση με τα περισσότερα δείγματα και αφαιρούνται από αυτήν δείγματα μέχρι να φτάσουν το πλήθος των δειγμάτων της άλλης κλάσης. Στα πλαίσια της πειραματικής διαδικασίας χρησιμοποιήθηκαν οι μέθοδοι της βιβλιοθήκες imbalanced learn, RandomOverSampler και RandomUnderSampler αντίστοιχα. Παρατηρήθηκε ότι με τη χρήση και των δύο μεθόδων τα μοντέλα μας εκπαιδεύονταν λιγότερο αποτελεσματικά και είχαμε μια πτώση της ορθότητας (accuracy) της τάξης του 1

Προτού περάσουμε στη μηχανική μάθηση πρέπει να τονισουμε ότι έχουν εξεταστεί μόνο χαρακτηριστικά που γνωρίζει κάποιος πριν απογειωθεί. Τα υπόλοιπα έχουν αφαιρεθεί. Με αυτόν τον τρόπο αυτό που προβλέπουμε είναι a priori γνώση επιβίβασης, δηλαδή για το πριν επιβιβαστείτε στο αεροπλάνο και όχι ενώ βρίσκεστε στο αεροπλάνο στον αέρα, κάτι που δεν θα ήταν πολύ χρήσιμο καθώς θα θέλατε να μάθετε αν θα αργήσετε πριν επιβιβαστείτε στο αεροπλάνο. Η προσθήκη οποιουδήποτε από τα χαρακτηριστικά TAXI-OUT, TAXI-IN, WHEELS-OFF, WHEELS-ON, ARR-DELAY, ACTUAL-ELAPSED-TIME θα αύξανε την ακρίβειά σας σε τουλάχιστον 80%, κάτι που ακούγεται υπέροχο, αλλά και πάλι, ποιο είναι το νόημα αν είστε ήδη στον αέρα;

Κεφάλαιο 5

Πειραματικά αποτελέσματα Μηχανικής Μάθησης

Στο κεφάλαιο αυτό θα εφαρμόσουμε όλες τις τεχνικές Μηχανικής Μάθησης που είδαμε στο κεφάλαιο 2.1. Σε κάθε μοντέλο για την παραγωγή των βέλτιστων δυνατών αποτελεσμάτων έπρεπε να γίνει βελτιστοποίηση υπερ-παραμέτρων (hyperparameter tuning). Στο παρόν κεφάλαιο εξηγείται η διαδικασία βελτιστοποίησης των υπερ-παραμέτρων των μοντέλων, καθώς και παρατίθενται και σχολιάζονται πίνακες με τις τιμές κάθε μετρικής για κάθε μοντέλο όπως προέκυψαν στην αξιολόγηση του συνόλου ελέγχου (τεστ σετ).

5.1 Βελτιστοποίηση Υπερ-Παραμέτρων

Η απόδοση όλων των πιθανών συνδυασμών υπερ-παραμέτρων στο εκάστοτε μοντέλο γίνεται με τη μέθοδο της αναζήτησης πλέγματος (Grid Search). Η αναζήτηση πλέγματος είναι απλά η εξαντλητική αναζήτηση όλων των συνδυασμών ενός ορισμένου συνόλου τιμών για κάθε υπερ-παράμετρο του μοντέλου. Οι τιμές αυτές ορίζονται χειροκίνητα βάσει δοκιμών και εμπειρικής γνώσης. Η μέθοδος πρέπει να καθοδηγείται από μια μετρική επίδοσης η οποία αποτιμάται πάνω στο σύνολο ελέγχου (test set) με τη χρήση διασταυρωμένης επικύρωσης (Cross Validation). Η υλοποίηση της μεθόδου έγινε μέσω της συνάρτησης GridSearchCV της βιβλιοθήκης Sci-kit Learn. Κατά την τεχνική αυτή τα δεδομένα εκπαίδευσης (training set) χωρίζονται σε ένα σταθερό αριθμό πτυχών (folds) στον οποίο θα γίνει το Cross Validation. Σε αυτή τη μέθοδο, χωρίζουμε το σετ εκπαίδευσης σε υποσύνολα ίσου μεγέθους K ή σε σύνολα δειγμάτων που ονομάζονται φολδς. Η εκπαιδευτική διαδικασία αποτελείται από μια σειρά επαναλήψεων. Κατά τη διάρκεια κάθε επανάληψης, τα βήματα είναι:

1. Διατηρήστε ένα υποσύνολο ως δεδομένα επικύρωσης και εκπαιδεύστε το μοντέλο μηχανικής εκμάθησης στα υπόλοιπα υποσύνολα K-1.

2. Παρατηρήστε πώς λειτουργεί το μοντέλο στο δείγμα επικύρωσης.

3. Βαθμολογήστε την απόδοση του μοντέλου με βάση την ποιότητα των δεδομένων εξόδου.

Χρησιμοποιώντας αυτή τη μέθοδο έχουμε αποτελεσματικότερη εκπαίδευση και αποφεύγουμε το πρόβλημα της υπερπροσαρμογής (οερφιττινγ)

Για την εκπαίδευση των μοντέλων το σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Έγινε χρήση της μεθόδου τραιν_τεστ_σπλιτ από τη βιβλιοθήκη Σχι-κιτ λεαρν. Το σύνολο ελέγχου αποτελεί το 30% των δεδομένων, και το υπόλοιπο 70% χρησιμοποιήθηκαν για την εκπαίδευση. Για κάθε μοντέλο εξετάστηκαν οι ακόλουθες υπερπαραμέτροι:

- Decision Trees : criterion, max_depth, max_features, min_samples_leaf, min_samples_split, splitter
- Random Forest : bootstrap, max_depth, max_features, min_samples leaf, min_samples_split, n_estimators
- AdaBoost : algorithm, learning_rate, n_estimators
- XGBoost : colsample_bytree, eta, gamma, learning_rate, max_depth, min_child_weight, subsample
- Multi_Layer Perceptron : activation, alpha, hidden_layer_sizes,learning_rate,solver

5.2 Υπερπροσαρμογή (Overfitting)

Η υπερπροσαρμογή είναι μια ανεπιθύμητη συμπεριφορά μηχανικής μάθησης που εμφανίζεται όταν το μοντέλο μηχανικής εκμάθησης παρέχει ακριβείς προβλέψεις για δεδομένα εκπαίδευσης αλλά όχι για νέα δεδομένα. Όταν οι επιστήμονες δεδομένων χρησιμοποιούν μοντέλα μηχανικής μάθησης για να κάνουν προβλέψεις, εκπαιδεύουν πρώτα το μοντέλο σε ένα γνωστό σύνολο δεδομένων. Στη συνέχεια, με βάση αυτές τις πληροφορίες, το μοντέλο προσπαθεί να προβλέψει τα αποτελέσματα για νέα σύνολα δεδομένων. Ένα μοντέλο υπερπροσαρμογής μπορεί να δώσει ανακριβείς προβλέψεις και δεν μπορεί να αποδώσει καλά για όλους τους τύπους νέων δεδομένων. Για την αποφυγή του προβλήματος αυτού εφαρμόζουμε την μέθοδο της διασταυρούμενης επικύρωσης (k-fold Cross Validation).

5.3 Αποτελέσματα

Για την αξιολόγηση των πειραμάτων χρησιμοποιήθηκαν οι μετρικές που αναφέρθηκαν στο Κεφάλαιο 2.6.5. Η βάση (baseline) για την αξιολόγηση των προβλέψεων στα δεδομένα μας είναι το 79.6%. Ο λόγος είναι ότι το ποσοστό αυτό αντιστοιχεί στη κλάση των πτήσεων non_-delayed, δηλαδή στις πτήσεις με is_Delayed status = 0. Σημειώνεται ότι οι μετρικές precision και recall αναφέρονται με σταθμισμένο μέσο όρο λόγω της ανισορροπίας των κλάσεων, για μια πιο ορθή εικόνα των αποτελεσμάτων.

Πριν περάσουμε στα αποτελέσματα πρέπει να επισυμάνουμε ότι έχουν εξεταστεί μόνο χαρακτηριστικά για το σετ δεδομένων που γνωρίζει κάποιος πριν απογειωθεί. Τα υπόλοιπα έχουν αφαιρεθεί. Αυτό μας οδηγεί στη σκέψη ότι το dep_delay αν και highly correlated με το arrival_delay χαρακτηριστικό των πτήσεων -γεγονός που διαπιστώσαμε κατά την διερευνητική ανάλυση δεδομένων- είναι μάλλον απαραίτητο. Συνεπώς στα αποτελέσματα του εκάστοτε αλγορίθμου μηχανικής μάθησης που χρησιμοποιήθηκε έχει ερευνηθεί πρώτον η περίπτωση

που έχουμε διατηρήσει το dep_delay σαν χαρακτηριστικό και δεύτερον η περίπτωση που έχει αφαιρεθεί απο το dataset .

Μετά το πέρας της εκπαίδευσης και της βελτιστοποίησης υπερ-παραμέτρων, συνολικά λοιπόν τα μοντέλα μας έδωσαν τα ακόλουθα αποτελέσματα:

5.3.1 Αποτελέσματα Decision Trees

```
Classification Report
---------------------
              precision    recall  f1-score   support

           0       0.86      0.66      0.75   4238273
           1       0.27      0.54      0.36   1008646

    accuracy                           0.64   5246919
   macro avg       0.57      0.60      0.55   5246919
weighted avg       0.74      0.64      0.67   5246919
```

Σχήμα 5.1: *Classification report for Decision Tree Classifier without Departure Delay accounted*

```
Classification Report
---------------------
              precision    recall  f1-score   support

           0       0.95      0.93      0.94   4238273
           1       0.74      0.81      0.78   1008646

    accuracy                           0.91   5246919
   macro avg       0.85      0.87      0.86   5246919
weighted avg       0.91      0.91      0.91   5246919
```

Σχήμα 5.2: *Classification report for Decision Tree Classifier with Departure Delay accounted*

5.3.2 Αποτελέσματα Bagged Trees

```
Classification Report
---------------------
              precision    recall  f1-score   support

           0       0.81      0.99      0.89   4238273
           1       0.58      0.04      0.07   1008646

    accuracy                           0.81   5246919
   macro avg       0.69      0.52      0.48   5246919
weighted avg       0.77      0.81      0.74   5246919
```

Σχήμα 5.3: *Classification report for Bagged Trees Classifier without Departure Delay*

```
Classification Report
---------------------
              precision    recall  f1-score   support

           0       0.94      0.98      0.96   4238273
           1       0.91      0.71      0.80   1008646

    accuracy                           0.93   5246919
   macro avg       0.92      0.85      0.88   5246919
weighted avg       0.93      0.93      0.93   5246919
```

Σχήμα 5.4: *Classification report for Bagged Trees Classifier with Departure Delay*

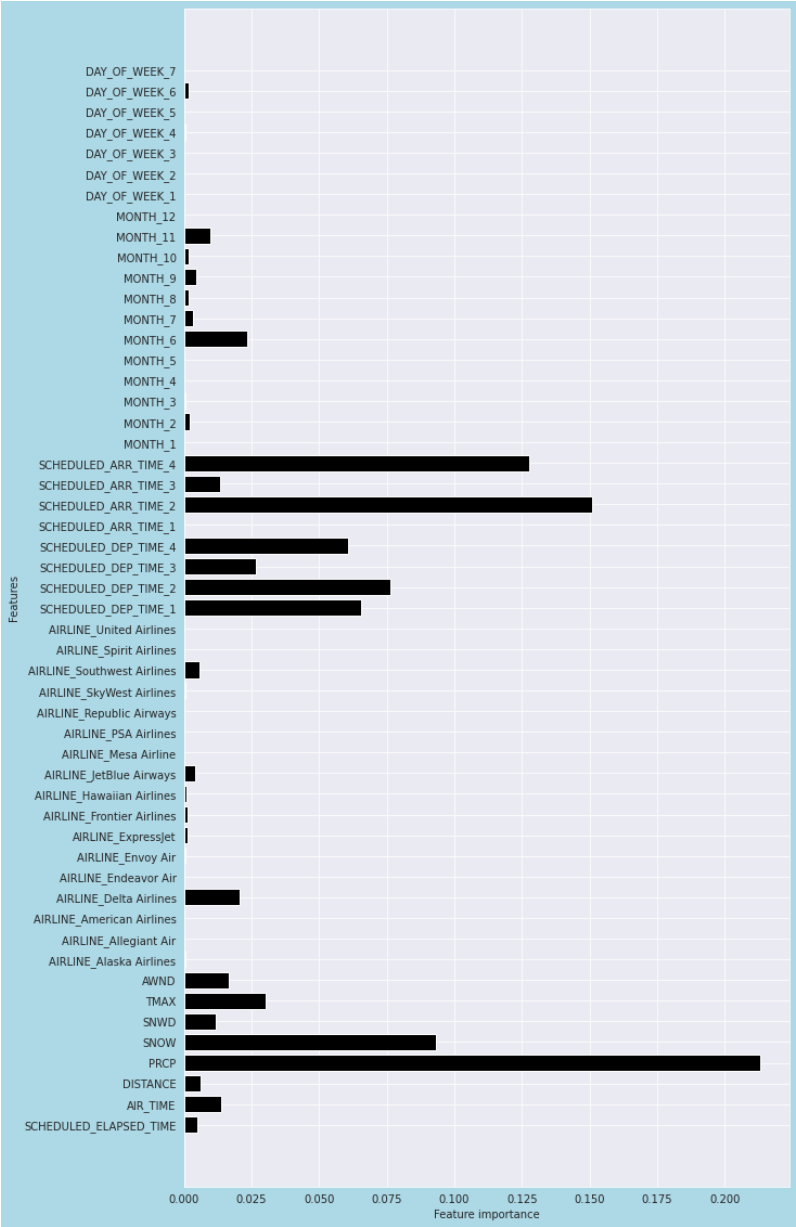
5.3.3 Αποτελέσματα Random Forest

```
Classification Report
---------------------
              precision    recall  f1-score   support

           0       0.81      1.00      0.89   4238273
           1       0.90      0.00      0.00   1008646

    accuracy                           0.81   5246919
   macro avg       0.85      0.50      0.45   5246919
weighted avg       0.83      0.81      0.72   5246919
```

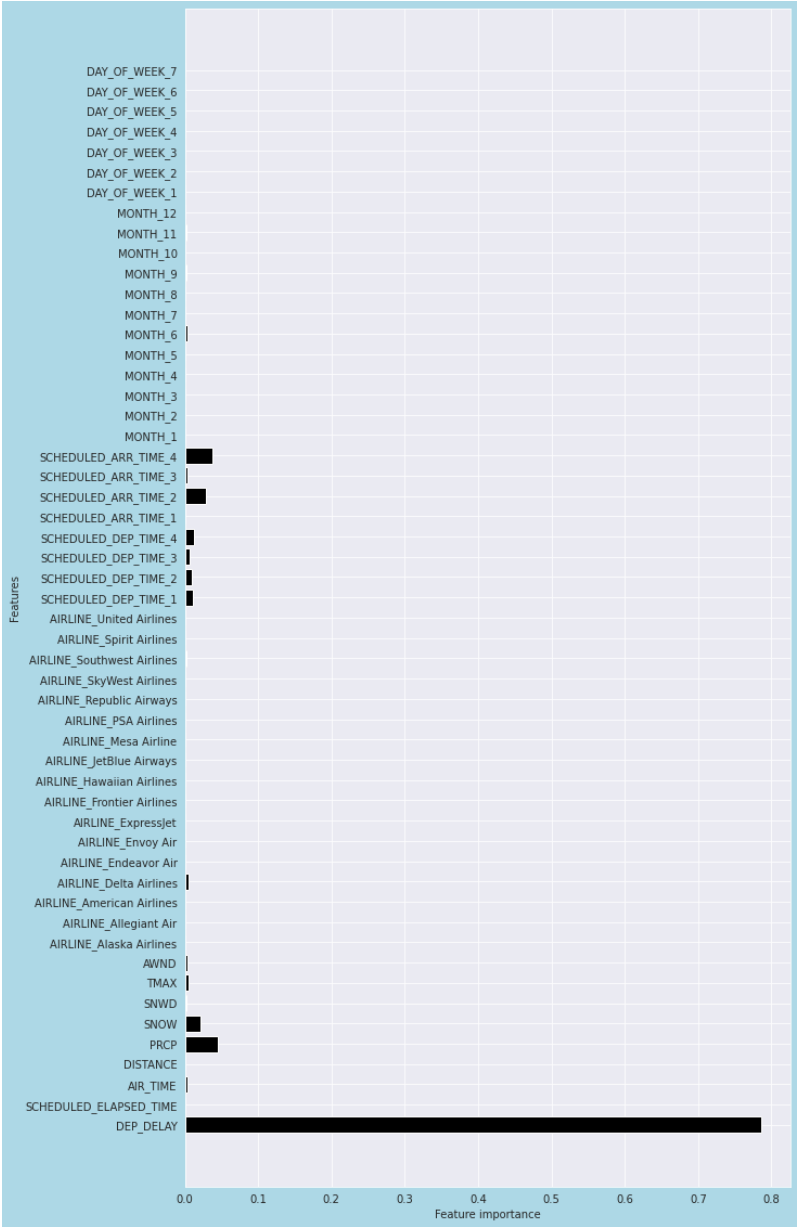
Σχήμα 5.5: *Classification report for Random Forest Classifier without Departure Delay*

```
Classification Report
---------------------
              precision    recall  f1-score   support

           0       0.92      0.99      0.95   4238273
           1       0.92      0.63      0.75   1008646

    accuracy                           0.92   5246919
   macro avg       0.92      0.81      0.85   5246919
weighted avg       0.92      0.92      0.91   5246919
```

Σχήμα 5.6: *Classification report for Random Forest Classifier with Departure Delay*



Σχήμα 5.7: *Feature Importance for Random Forrest Classifier without Departure Delay*



Σχήμα 5.8: *Feature Importance for Random Forrest Classifier with Departure Delay*

5.3.4 Αποτελέσματα ADAptive Boost

	precision	recall	f1-score	support
0	0.81	0.99	0.89	3973520
1	0.58	0.03	0.06	945466
accuracy			0.81	4918986
macro avg	0.70	0.51	0.48	4918986
weighted avg	0.77	0.81	0.73	4918986

Σχήμα 5.9: *Classification report for ADAptive Boost Classifier without Departure Delay*

	precision	recall	f1-score	support
0	0.94	0.98	0.96	3973520
1	0.90	0.72	0.80	945466
accuracy			0.93	4918986
macro avg	0.92	0.85	0.88	4918986
weighted avg	0.93	0.93	0.93	4918986

Σχήμα 5.10: *Classification report for ADAptive Boost Classifier with Departure Delay*

```
Mean Adaboost Cross-Val Score (k=5):
0.7195117606590815
```

Σχήμα 5.11: *Cross validation score for ADAboost without Departure Delay*

```
Mean Adaboost Cross-Val Score (k=5):
0.8503592853260367
```

Σχήμα 5.12: *Cross validation score for ADAboost with Departure Delay*

5.3.5 Αποτελέσματα Gradient Boosted Trees

	precision	recall	f1-score	support
0	0.81	0.99	0.90	3973520
1	0.67	0.04	0.08	945466
accuracy			0.81	4918986
macro avg	0.74	0.52	0.49	4918986
weighted avg	0.79	0.81	0.74	4918986

Σχήμα 5.13: *Classification report for Gradient Boosted Trees Classifier without Departure Delay*

	precision	recall	f1-score	support
0	0.94	0.98	0.96	3973520
1	0.92	0.73	0.81	945466
accuracy			0.94	4918986
macro avg	0.93	0.86	0.89	4918986
weighted avg	0.94	0.94	0.93	4918986

Σχήμα 5.14: *Classification report for Gradient Boosted Trees Classifier Departure Delay*

```
Mean GBT Cross-Val Score (k=5):
0.8547926676738606
```

Σχήμα 5.15: *Cross validation score for Gradient Boosted Trees without Departure Delay*

5.3.6 Αποτελέσματα eXtreme gradient Boosting

```
Training Accuracy: 93.48%
Validation accuracy: 93.49%
```

```
xgboost_classification_report = classification_report(y_test, val_preds)
print(xgboost_classification_report)
```

```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96   3973520
           1       0.92      0.72      0.81    945466

    accuracy                           0.93   4918986
   macro avg       0.93      0.85      0.89   4918986
weighted avg       0.93      0.93      0.93   4918986
```

Σχήμα 5.16: *Classification report for eXtreme Gradient Boosting Classifier without Departure Delay*

5.3.7 Αποτελέσματα Multi Layer Perceptron

Συνθήκες που πρέπει να ισχύουν για ένα νευρωνικό δίκτυο:

1. Τα δεδομένα πρέπει να είναι καθαρά αριθμητικά
2. Τα δεδομένα δεν μπορούν να περιέχουν τιμές που λείπουν
3. Τα δεδομένα πρέπει να κανονικοποιηθούν

Δοκιμάσαμε να χτίσουμε διάφορα ακολουθιακό μοντέλα. Ένα ακολουθιακό μοντέλο είναι κατάλληλο για μια απλή στοίβα στρωμάτων όπου κάθε στρώμα έχει ακριβώς έναν τανυστή εισόδου και έναν τανυστή εξόδου.

Για την δημιουργία των στρωμάτων χρησιμοποιήθηκε η βιβλιοθήκη Keras του Tensorflow. Ως στρώματα χρησιμοποιούμε πυκνά (densely) συνδεδεμένα στρώματα.

Τι είναι ένα πυκνό στρώμα;

Σε οποιοδήποτε νευρωνικό δίκτυο, ένα πυκνό στρώμα είναι ένα στρώμα που είναι βαθιά συνδεδεμένο με το προηγούμενο στρώμα του, που σημαίνει ότι οι νευρώνες του στρώματος συνδέονται με κάθε νευρώνα του προηγούμενου στρώματός του. Αυτό το επίπεδο είναι το πιο συχνά χρησιμοποιούμενο στρώμα σε τεχνητά νευρωνικά δίκτυα. Για την δημιουργία του καλούμε την συνάρτηση Δενσε του Κερασ. Κάθε πυκνό στρώμα υλοποιεί τη λειτουργία:

output = activation(dot(input, kernel) + bias), ακριβώς όπως είδαμε στο κεφάλαιο 2.6.1. Ο τρόπος ενεργοποίησης (activation) του στρώματος(νευρώνα) δηλώνεται ως όρισμα κατά τη δημιουργία του και είναι μία εκ των συναρτήσεων ενεργοποίησης που είδαμε στο κεφάλαιο 2.6.3. Ο πυρήνας (kernel) είναι ένας πίνακας βαρών που δημιουργείται από το επίπεδο και το bias είναι ένα διάνυσμα πόλωσης που δημιουργείται από το επίπεδο.

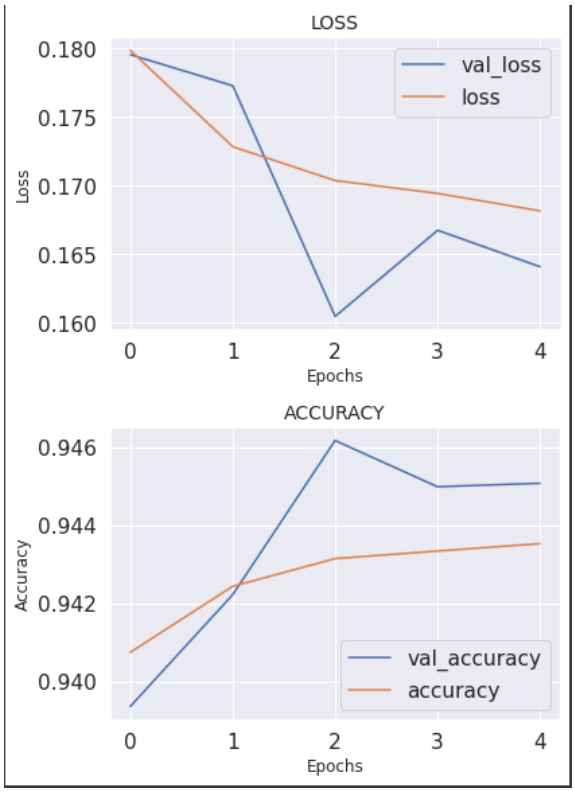
Η πρώτη μας δοκιμή είναι να δημιουργήσουμε ένα MLP 2 επιπέδων ως εξής:

```
model_1 = Sequential()

model_1.add(Dense(30, activation='tanh', input_shape=(53,)))

model_1.add(Dense(1, activation='sigmoid'))
```

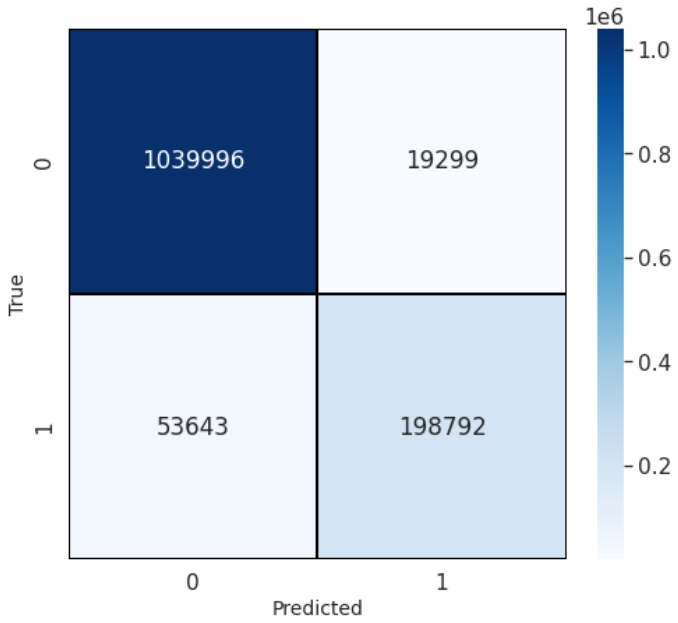
Δηλαδή εδώ θα έχουμε 2-στρώματα (input, output). Το πρώτο θα έχει 30 διαστάσεις ως έξοδο, είσοδο 54 διαστάσεις που αντιστοιχούν στις στήλες των χαρακτηριστικών του σετ δεδομένων μετά το one-hot encoding και ενεργοποίηση την Συνάρτηση Υπερβολικής Εφαπτομένης. Το δεύτερο στρώμα είναι η τελική έξοδος και θα έχει συνάρτηση ενεργοποίησης την Σιγμοειδή.

	precision	recall	f1-score	support
0	0.95	0.98	0.97	1059295
1	0.91	0.79	0.84	252435
accuracy			0.94	1311730
macro avg	0.93	0.88	0.91	1311730
weighted avg	0.94	0.94	0.94	1311730

Σχήμα 5.17: *Classification report for MLP model1 with Departure Delay*



Σχήμα 5.18: *Training results for MLP model1 with Departure Delay*



Σχήμα 5.19: *Confusion Matrix for MLP model1 with Departure Delay*

Η δεύτερη απόπειρα είναι να δημιουργήσουμε ένα MLP 3 επιπέδων ως εξής:

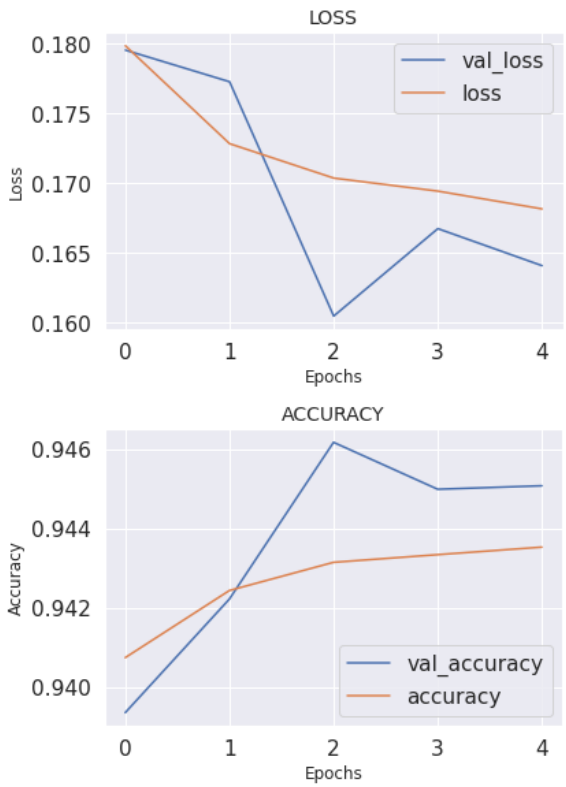
```
model_2 = Sequential()
model_2.add(Dense(30, activation='tanh', input_shape=(53,)))
model_2.add(Dense(10, activation='tanh'))
```

model_2.add(Dense(1, activation='sigmoid'))

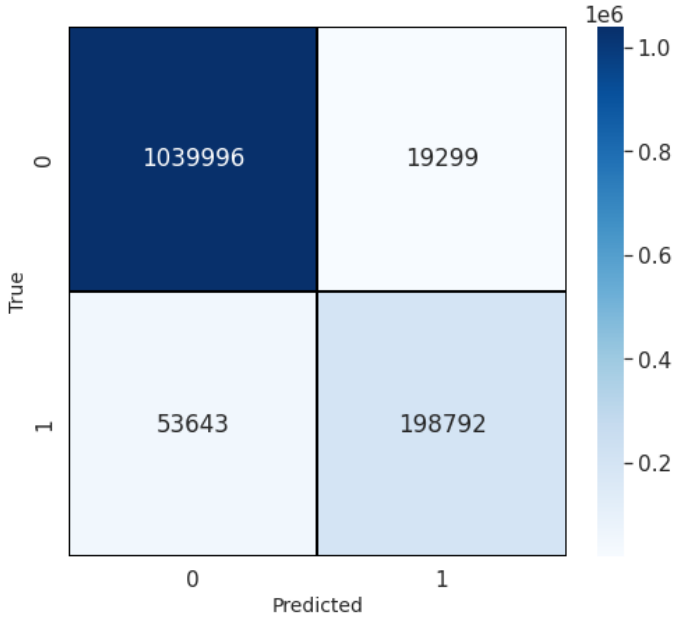
Δηλαδή εδώ θα έχουμε 3-στρώματα (input, 1-hidden, output). Το πρώτο θα έχει 30 διαστάσεις ως έξοδο, είσοδο 54 διαστάσεις που αντιστοιχούν στις στήλες των χαρακτηριστικών του σετ δεδομένων μετά το one-hot encoding και ενεργοποίηση την Συνάρτηση Υπερβολικής Εφαπτομένης. Το δεύτερο στρώμα θα έχει 10 διαστάσεις ως έξοδο, ενεργοποίηση την Συνάρτηση Υπερβολικής Εφαπτομένης. Τέλος, το τρίτο στρώμα είναι η τελική έξοδος και θα έχει συνάρτηση ενεργοποίησης την Σιγμοειδή.

	precision	recall	f1-score	support
0	0.95	0.98	0.97	1059295
1	0.91	0.79	0.84	252435
accuracy			0.94	1311730
macro avg	0.93	0.88	0.91	1311730
weighted avg	0.94	0.94	0.94	1311730

Σχήμα 5.20: *Classification report for MLP model2 with Departure Delay*



Σχήμα 5.21: *Training results for MLP model2 with Departure Delay*



Σχήμα 5.22: *Confusion Matrix for MLP model2 with Departure Delay*

Η τρίτη δοκιμή είναι να δημιουργήσουμε ένα MLP 3 επιπέδων ως εξής:

model_3 = Sequential()

model_3.add(Dense(30, activation='relu', input_shape=(53,)))

model_3.add(Dense(10, activation='relu'))

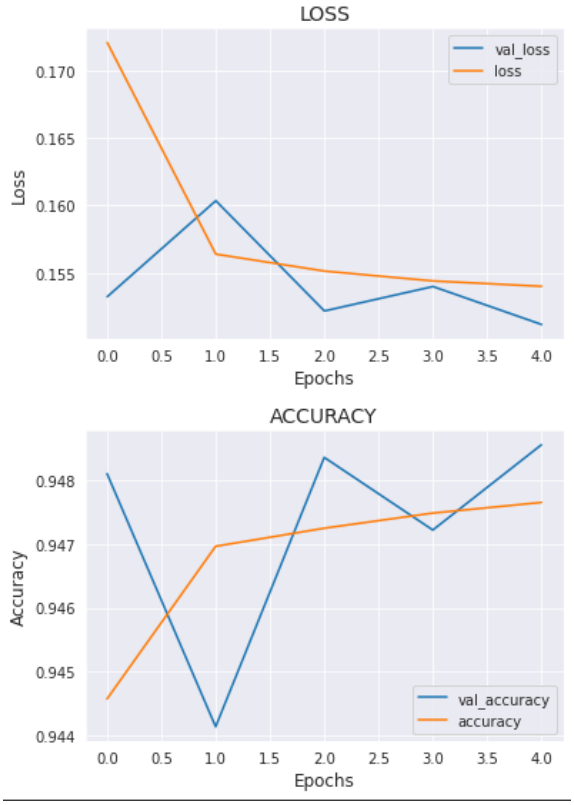
model_3.add(Dense(5, activation='relu'))

model_3.add(Dense(1, activation='sigmoid'))

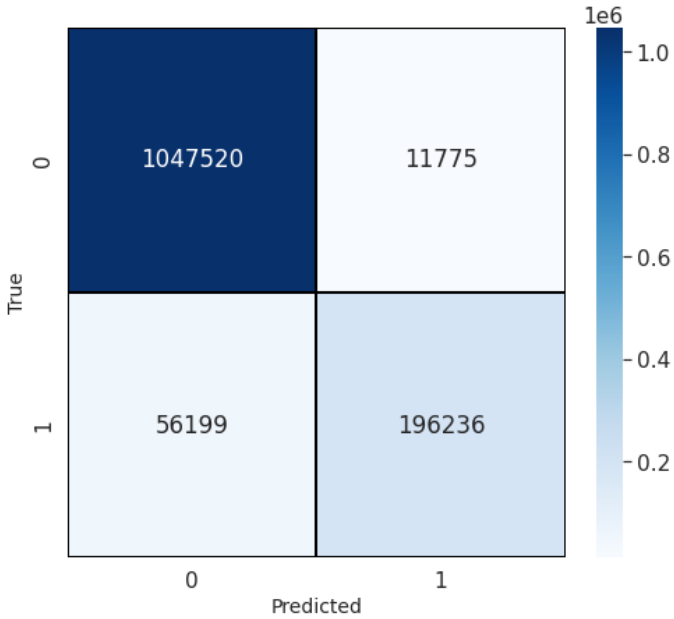
Δηλαδή εδώ θα έχουμε 4-στρώματα (input, 2-hidden, output). Το πρώτο θα έχει 30 διαστάσεις ως έξοδο, είσοδο 54 διαστάσεις που αντιστοιχούν στις στήλες των χαρακτηριστικών του σετ δεδομένων μετά το one-hot encoding και ενεργοποίηση την Συνάρτηση Υπερβολικής Εφαπτομένης. Το δεύτερο στρώμα θα έχει 10 διαστάσεις ως έξοδο, ενεργοποίηση την Συνάρτηση Υπερβολικής Εφαπτομένης. Το τρίτο στρώμα θα έχει 5 διαστάσεις ως έξοδο, ενεργοποίηση την Συνάρτηση Υπερβολικής Εφαπτομένης. Τέλος, το τέταρτο στρώμα είναι η τελική έξοδος και θα έχει συνάρτηση ενεργοποίησης την Σιγμοειδή.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	1059295
1	0.94	0.78	0.85	252435
accuracy			0.95	1311730
macro avg	0.95	0.88	0.91	1311730
weighted avg	0.95	0.95	0.95	1311730

Σχήμα 5.23: *Classification report for MLP model3 with Departure Delay*



Σχήμα 5.24: *Training results for MLP model3 with Departure Delay*



Σχήμα 5.25: *Confusion Matrix for MLP model3 with Departure Delay*

5.3.8 Γενικές παρατηρήσεις

Στα τελικά δεδομένα εκπαίδευσης που κατασκευάσαμε μπορούμε να καταλάβουμε πόσο καθοριστικό ρόλο στην πρόβλεψη καθυστέρησης κατα την άφιξη μιας πτήσης παίζει η πληροφορία του departure_delay. Στους περισσότερους ταξινομητές τα ποσοστά διαφέρουν κατα

τεράστιο βαθμό στην πρόβλεψη των καθυστερημένων πτήσεων, ενώ η ακρίβεια του κάθε ταξινομητή πλην των Random Forrest, Decision Tree που έχει εκπαιδευτεί με την πληροφορία για το dep_delay είναι πάντα πάνω απο το baseline.

Μέρος III

Επίλογος

Κεφάλαιο 6

Επίλογος

6.1 Συμπεράσματα

Στην παρούσα εργασία χρησιμοποιήθηκαν τεχνικές Επιβλεπόμενης Μάθησης με σκοπό την πρόβλεψη καθυστέρησης πτήσεων με την ταξινόμηση τους σε δύο κατηγορίες, τις καθυστερημένες και τις μη καθυστερημένες (Delayed, non_Delayed). Σκοπός ήταν η μελέτη τόσο του ενδεχομένου επιτυχούς πρόβλεψης των καθυστερήσεων χρήσει τεχνικών Μηχανικής Μάθησης, όσο και η διερεύνηση των παραγόντων εκείνων, που συμβάλλουν σημαντικότερα καθορίζοντας την επιτυχία της πρόβλεψης. Μετά τη διεξαγωγή πειραμάτων με διάφορα μοντέλα ταξινομητών, οι οποίοι χρησιμοποιούν ποικίλες τεχνικές για την επίλυση του διερευνώμενου προβλήματος, συμπεράναμε ότι οι καλύτεροι αλγόριθμοι για την αντιμετώπιση του, είναι αλγόριθμοι συνόλου Ensemble Algorithms και συγκεκριμένα αυτοί που χρησιμοποιούν την τεχνική ακραίας Ενίσχυσης και Ενισχυσης Κλίσης XGBoost και Gradient Boost. Οι τρεις αλγόριθμοι ενίσχυσης οι οποίοι εξετάστηκαν στα πλαίσια της πειραματικής διαδικασίας ήταν οι AdaBoost, Gradient Boost και XGBoost, με τον τελευταίο να αποδεικνύεται καλύτερο/αποδοτικότερο για το πρόβλημα μας. Επιπρόσθετως, πολυ καλή απόδοση έδειξαν να έχουν και τα νευρωνικά δίκτυα δηλαδή οι πολυστρωματικοί νευρώνες(MLP).

Συνεπώς διαπιστώσαμε εμπεριστατωμένα ότι η πρόβλεψη των καθυστερήσεων είναι αντικειμενική εφόσον βασίζεται σε στοχαστικά μεγέθη όπως είναι ο καιρός και οι προηγούμενες καθυστερήσεις και ακριβής σε μεγάλο ποσοστό στο οποίο συμμετέχει οχι μονον οι καιρικές συνθήκες αλλα και η γνώση παραμέτρων(όπως το departure delay) κατά την αναχώρηση της πτήσης.

6.2 Μελλοντικές Επεκτάσεις

Το σύστημα που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής εργασίας θα μπορούσε να βελτιωθεί και να επεκταθεί περαιτέρω, τουλάχιστον ως προς τρεις κατευθύνσεις. Συγκεκριμένα, αναφέρονται τα ακόλουθα:

- Στοχευμένη ανάλυση για μία πόλη ή αεροδρόμιο και πολλούς προορισμούς
- Μπορούμε να τρέξουμε ξανα τους αλγορίθμους μηχανικής μάθησης και των νευρωνικών δικτύων για το αεροδρόμιο που επιλέξαμε

- Επέκταση του σετ δεδομένων ώστε να περιέχει πληροφορία για τις προηγούμενες πτήσεις. Ένας πολύ σημαντικών λόγος όπως δείξαμε είναι η καθυστέρηση στην αναχώρηση. Εύλογο λοιπόν είναι πως η μεταβίβαση απο μια πτήση σε άλλη οταν υπάρχει ανταπόκριση να συμμετέχει και αυτή στην καθυστέρηση. Άρα θα μπορούσαμε να προσθέσουμε την ώρα μετεπιβίβασης που έχεις από την μία πτήση για την άλλη και μετά να ακολουθήσουμε σε νέα ανάλυση και modeling

Παραρτήματα

Παράρτημα Α΄

Παραδείγματα Βιβλιογραφικών Αναφορών

Τύπος βιβλιογραφικής πηγής	Αριθμός αναφοράς
Βιβλίο ξενόγλωσσο	[;]
Βιβλίο ελληνικό	[;]
Άρθρο σε επιστημονικό περιοδικό	[;]
Παρουσίαση σε επιστημονικό συνέδριο	[;]
Ιστοσελίδα	[38]
Διπλωματική εργασία	[39]
Πτυχιακή εργασία	[40]
Μεταπτυχιακή διπλωματική εργασία	[41]
Διδακτορική διατριβή	[;]
Δίπλωμα ευρεσιτεχνίας (πατέντα)	[;]
Τεχνική αναφορά	[;]

Βιβλιογραφία

[1] Roohollah Amiri, Hani Mehrpouyan, Lex Fridman, Ranjan Mallik, Arumugam Nallanathan και David Matolak. *A Machine Learning Approach for Power Allocation in HetNets Considering QoS*. https://www.researchgate.net/figure/Reinforcement-Learning-Agent-and-Environment_fig2_323867253. Ημερομηνία πρόσβασης: 28-08-2022.

[2] Gilgoldm. *A tree showing survival of passengers on the Titanic*. https://en.wikipedia.org/wiki/Decision_tree_learning. Ημερομηνία πρόσβασης: 16-08-2022.

[3] Yash Khandelwal. *Ensemble Stacking for Machine Learning and Deep Learning*. https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning/. Ημερομηνία πρόσβασης: 17-08-2022.

[4] z ai. *What is Boosting in Machine Learning?* https://towardsdatascience.com/what-is-boosting-in-machine-learning-2244aa196682. Ημερομηνία πρόσβασης: 23-08-2022.

[5] Brendan Marsh. *Multivariate Analysis of the Vector Boson Fusion Higgs Boson*. https://www.researchgate.net/figure/Training-of-an-AdaBoost-classifier-The-first-classifier-trains-on-unweighted-data-then_fig3_306054843. Ημερομηνία πρόσβασης: 23-08-2022.

[6] Alexey Natekin και Alois Knoll. *Gradient Boosting Machines, A Tutorial*. https://www.researchgate.net/figure/Algorithm-1-Friedmans-Gradient-Boost-algorithm_fig13_259653472. Ημερομηνία πρόσβασης: 23-08-2022.

[7] Aratrika Pal. *Gradient Boosting Trees for Classification: A Beginner's Guide*. https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea. Ημερομηνία πρόσβασης: 19-08-2022.

[8] Lior Rokach Omer Sagi. *Ensemble learning: A survey*. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249. Ημερομηνία πρόσβασης: 26-08-2022.

[9] Ankit Chauhan. *Random Forest Classifier and its Hyperparameters*. https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6. Ημερομηνία πρόσβασης: 26-08-2022.

[10] Wiso. *Neural network*. https://en.wikipedia.org/wiki/Neural_network. Ημερομηνία πρόσβασης: 28-08-2022.

[11] SAGAR SHARMA. *What the Hell is Perceptron?* https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53. Ημερομηνία πρόσβασης: 28-08-2022.

[12] Francke Peixoto. *A Simple overview of Multilayer Perceptron(MLP)*. https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/. Ημερομηνία πρόσβασης: 28-08-2022.

[13] Nomidl. *What is ReLU and Sigmoid activation function?* https://www.nomidl.com/deep-learning/what-is-relu-and-sigmoid-activation-function/. Ημερομηνία πρόσβασης: 31-08-2022.

[14] Bryan Shalloway. *Weighting Confusion Matrices by Outcomes and Observations*. https://www.r-bloggers.com/2020/12/weighting-confusion-matrices-by-outcomes-and-observations/. Ημερομηνία πρόσβασης: 01-09-2022.

[15] DANB. *Using Categorical Data with One Hot Encoding*. https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook. Ημερομηνία πρόσβασης: 22-10-2022.

[16] Fanxin Meng, Gengyuan Liu, Zhifeng Yang, Yan Hao και Sergio Ulgiati. *Assessment of Urban Transportation Metabolism from Life Cycle Perspective: A Multi-method Study*. *Energy Procedia*, 88:243–249, 2016. CUE 2015 - Applied Energy Symposium and Summit 2015: Low carbon cities and urban energy systems.

[17] Tony Diana. *Can machines learn how to forecast taxi-out time? A comparison of predictive models applied to the case of Seattle/Tacoma International Airport*. *Transportation Research Part E: Logistics and Transportation Review*, 119:149–164, 2018.

[18] Avrim Blum. *Machine learning theory*. *Carnegie Melon Universit, School of Computer Science*, 26, 2007.

[19] Tom M Mitchell και Tom M Mitchell. *Machine learning*, τόμος 1. McGraw-hill New York, 1997.

[20] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[21] Pádraig Cunningham, Matthieu Cord και Sarah Jane Delany. *Supervised learning*. *Machine learning techniques for multimedia*, σελίδες 21–49. Springer, 2008.

[22] Zoubin Ghahramani. *Unsupervised learning*. *Summer school on machine learning*, σελίδες 72–112. Springer, 2003.

[23] Marco A Wiering και Martijn Van Otterlo. *Reinforcement learning*. *Adaptation, learning, and optimization*, 12(3):729, 2012.

[24] Barry De Ville. *Decision trees*. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, 2013.

[25] Omer Sagi και Lior Rokach. *Ensemble learning: A survey*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):ε1249, 2018.

[26] Alexey Natekin και Alois Knoll. *Gradient boosting machines, a tutorial*. *Frontiers in neurorobotics*, 7:21, 2013.

[27] Jerome Friedman, Trevor Hastie και Robert Tibshirani. *Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)*. *The annals of statistics*, 28(2):337–407, 2000.

[28] Leo Breiman. *Bagging predictors*. *Machine learning*, 24(2):123–140, 1996.

[29] Ludmila I Kuncheva. *Classifier ensembles for changing environments*. *International Workshop on Multiple Classifier Systems*, σελίδες 1–15. Springer, 2004.

[30] Leo Breiman. *Random forests*. *Machine learning*, 45(1):5–32, 2001.

[31] Larry Hardesty. *Explained: neural networks*. *MIT News*, 14, 2017.

[32] Sagar Sharma, Simone Sharma και Anidhya Athaiya. *Activation functions in neural networks*. *towards data science*, 6(12):310–316, 2017.

[33] Mohammad Hossin και Md Nasir Sulaiman. *A review on evaluation metrics for data classification evaluations*. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[34] John T Behrens. *Principles and procedures of exploratory data analysis*. *Psychological Methods*, 2(2):131, 1997.

[35] Stephan Morgenthaler. *Exploratory data analysis*. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44, 2009.

[36] John W Tukey και others. *Exploratory data analysis*, τόμος 2. Reading, MA, 1977.

[37] Mario Li Vigni, Caterina Durante και Marina Cocchi. *Exploratory data analysis*. *Data handling in science and technology*, τόμος 28, σελίδες 55–126. Elsevier, 2013.

[38] *LaTeX Project*. http://www.latex-project.org. Ημερομηνία πρόσβασης: 13-11-2014.

[39] Ζ. Καούδη. *Πρότυπο Σύστημα Αποθήκευσης και Διαχείρισης Σχημάτων RDFS*. Διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2004.

[40] Ε. Ανδρουλάκη. *Υλοποίηση Ενεργού Μηχανισμού σε Σύστημα Ομότιμων Βάσεων*. Πτυχιακή εργασία, KDBS Lab, Εθνικό Μετσόβιο Πολυτεχνείο, 2005.

[41] Ζ. Λάσκαρη. *Κοινωνική Ανάλυση των Ταινιών της Finos Films*. Μεταπτυχιακή διπλωματική εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, 2012.

Συντομογραφίες - Αρκτικόλεξα - Ακρωνύμια

Ελληνικοί όροι:

ΜΜ	Μηχανική Μάθηση
ΤΝΔ	Τεχνητό Νευρωνικό Δίκτυο
κ.α.	και άλλα
κ.ο.κ	και ούτω καθεξής

Αγγλικοί όροι:

NAS	National Airspace System
BTS	Bureau of Transportation Statistics
FAA	Federal Aviation Administration
NTSB	National Transportation Safety Board
NOAA	National Oceanic and Atmospheric Administration
KNN	K-Nearest Neighbors
MLE	Maximum Likelihood Estimation
OLS	Ordinary Least Squares
LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
RF	Random Forest
SVM	Support Vector Machines
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
MAE	Mean Absolute Error
UTC	Coordinated Universal Time