



Ionian University
Department of Informatics

Τίτλος πτυχιακής εργασίας

«Ανάλυση Μεγάλου Όγκου Δεδομένων Αεροπορικών Πτήσεων με
Χρήση Τεχνικών Μηχανικής Μάθησης»

Βαμβακάς Παναγιώτης (Π2019118)

Επιβλέπων: Ανδρέας Καναβός

Κέρκυρα, 2023

Επιβλέπων

Ανδρέας Καναβός, *Αναπληρωτής Καθηγητής,*

Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Άννα Σωτηροπούλου, *Εργαστηριακό Διδακτικό Προσωπικό,*

Ιόνιο Πανεπιστήμιο

Ανδρέας Καναβός, *Αναπληρωτής Καθηγητής,*

Ιόνιο Πανεπιστήμιο

Στέργιος Παλαμάς, *Επίκουρος Καθηγητής (μόνιμος),*

Ιόνιο Πανεπιστήμιο

Αυτή η προσπάθεια αφιερώνεται στους φοιτητές του Τμήματος Πληροφορικής του
Ιόνιου Πανεπιστημίου

Πίνακας Περιεχομένων

Περίληψη	6
Abstract.....	7
1. Εισαγωγή.....	8
1.1 Ο κλάδος της αεροπορίας.....	8
1.2 Υπόβαθρο και Κίνητρο	9
1.3 Δήλωση Προβλήματος.....	10
1.4 Ερευνητικοί Στόχοι	11
1.5 Εμβέλεια και Περιορισμοί	12
2. Ανασκόπηση Βιβλιογραφίας Έρευνας.....	13
2.1 Επισκόπηση της Ανάλυσης Δεδομένων στην Αεροπορία	13
2.2 Προηγούμενες Μελέτες για την Πρόβλεψη Καθυστερήσεων Πτήσεων	14
2.3 Σχετικές Τεχνικές Εξόρυξης Δεδομένων και Μηχανικής Μάθησης	17
2.4 Συμπεράσματα από την Ανασκόπηση της Βιβλιογραφίας.....	18
3. Συλλογή και Προ-Επεξεργασία Δεδομένων	19
3.1 Πηγή Δεδομένων και Περιγραφή.....	19
3.2 Μεθοδολογία Συλλογής Δεδομένων	20
3.3. Εξερεύνηση Ανάλυσης Δεδομένων	22
3.4. Τεχνικές Καθαρισμού και Προ-Επεξεργασίας Δεδομένων	23
3.5. Επιλογή και Μηχανισμός Χαρακτηριστικών	24
4. Μεθοδολογία.....	26
4.1 Επισκόπηση της Προβλεπτικής Μοντελοποίησης.....	26
4.2 Επιλογή Αλγορίθμων για την πρόβλεψη Καθυστερήσεων Πτήσεων	27
4.3 Ανάπτυξη και Εκπαίδευση του Μοντέλου.....	28
4.4 Αξιολόγηση και Επικύρωση του Μοντέλου	29
5. Πειραματικά Αποτελέσματα και Ανάλυση	30
5.1 Περιγραφή Δεδομένων και Στατιστικά.....	30
5.2 Αξιολόγηση Προβλεπτικών Μοντέλων	34
5.3 Σύγκριση και Ανάλυση Επίδοσης	35
6. Συζήτηση.....	36
6.1 Ερμηνεία των Ευρημάτων.....	36

6.2	Επισημάνσεις στους Παράγοντες που Επηρεάζουν τις Καθυστερήσεις Πτήσεων .	37
6.3	Περιορισμοί της Μελέτης	38
7.	Ανάλυση κώδικα.....	39
7.1	Πηγαίος κώδικας	39
7.2	Pandas Library.....	41
7.3	scikit-learn.....	42
8.	Αποτελέσματα Έρευνας	44
8.1	Αποτελέσματα με τιμές RMSE	44
8.2	Ερμηνεία Αποτελεσμάτων	53
9.	Συμπεράσματα Έρευνας.....	54
10.	Μελλοντικές επεκτάσεις	56
11.	Βιβλιογραφία	58
12.	Παραρτήματα	60
12.1	Πηγαίος κώδικας Κώδικας εκπαίδευσης/εκτέλεσης μοντέλων	60
12.2	Ανάλυση πηγαίου κώδικα	66

Περίληψη

Ο 21^{ος} αιώνας θεωρείται η εποχή της υπέρμετρης τεχνολογικής ανάπτυξης και των καταγιστικών εξελίξεων. Ένα από τα κυριότερα χαρακτηριστικά του αποτελεί το φαινόμενο της παγκοσμιοποίησης, που έχει εκμηδενίσει τις αποστάσεις και έχει μετατρέψει την ανθρωπότητα σε «μικρό χωριό». Οι αποστάσεις ανάμεσα στα έθνη έχουν ελαχιστοποιηθεί και δεν αποτελούν πλέον ανυπέρβλητο εμπόδιο. Καταλυτικό ρόλο σε αυτή την εξέλιξη, κατά γενική ομολογία, έχουν διαδραματίσει οι αεροπορικές εταιρείες, που εύλογα ο τρόπος λειτουργίας τους θεωρείται θεμελιώδης παράγοντας στη σύγχρονη εποχή. [11]

Στην παρούσα έρευνα, παρουσιάζεται μια εκτενής ανάλυση των καθυστερημένων πτήσεων βασισμένη σε ένα από τα πιο ολοκληρωμένα σύνολα δεδομένων στον τομέα της αεροπορίας, για την περίοδο 2018-2022. Χρησιμοποιώντας αλγόριθμους όπως οι Random Forest Regressor, Decision Trees, Logistic Regression και Support Vector Machines [10], η έρευνα εξερευνά πολύπλοκες σχέσεις μεταξύ των μεταβλητών και προβλέπει τις πιθανές καθυστερήσεις πτήσεων. Η εργασία αναζητά σύνθετες σχέσεις μεταξύ των μεταβλητών των αεροπορικών πτήσεων που περιλαμβάνονται στο σύνολο δεδομένων για κάθε χρονιά και χρησιμοποιεί αυτές τις σχέσεις για την πρόβλεψη αεροπορικών καθυστερήσεων με την ενίσχυση όλων αυτών των αλγορίθμων.

Στόχος της έρευνας αποτελεί η βαθύτερη κατανόηση των δυναμικών των καθυστερήσεων σε αεροπορικές πτήσεις και η προώθηση της καινοτομίας στον τομέα. Ο σκοπός αυτός μπορεί να αποβεί επωφελής σε ευρύ πλήθος κόσμου με τα αεροπορικά ταξίδια να είναι απαραίτητα στη καθημερινότητα πολλών ανθρώπων και τις αεροπορικές καθυστερήσεις να θεωρούνται μεγάλο πρόβλημα, που εξακολουθεί να αυξάνεται σταθερά.

Abstract

The 21st century is considered the era of excessive technological development and stormy developments. One of its main characteristics is the phenomenon of globalization, which has annihilated distances and turned humanity into a "small village". The distances between nations have been minimized and are no longer an insurmountable barrier. A catalytic role in this development, by general admission, has been played by the airlines, whose mode of operation is reasonably considered a fundamental factor in the modern era. [11]

In this research, an extensive analysis of delayed flights based on one of the most comprehensive datasets in the aviation sector is presented for the period 2018-2022. Using algorithms such as Random Forest Regressor, Decision Trees, Logistic Regression and Support Vector Machines [10], the research explores complex relationships between variables and predicts potential flight delays. The work looks for complex relationships between airline flight variables included in the data set for each year and uses these relationships to predict airline delays by enhancing all these algorithms.

The aim of the research is to gain a deeper understanding of the dynamics of air flight delays and to promote innovation in the sector. This purpose can be beneficial to a large number of people with air travel being essential to many people's daily lives and air delays being considered a big problem that is still steadily increasing.

1. Εισαγωγή

1.1 Ο κλάδος της αεροπορίας

Η αεροπορία, ως ένας κρίσιμος τομέας της παγκόσμιας οικονομίας, έχει βιώσει σημαντικές αλλαγές και προκλήσεις από το ξεκίνημά της στο έτος 1914. Στο πλαίσιο αυτό, οι καθυστερήσεις πτήσεων αναδεικνύονται ως ένα από τα πιο προβληματικά ζητήματα, επηρεάζοντας εκατομμύρια επιβάτες και αεροπορικές εταιρείες κάθε χρόνο. Το υπόβαθρο και το κίνητρο πίσω από αυτήν την έρευνα είναι να εντοπίσουμε και να κατανοήσουμε τις ρίζες του προβλήματος, την επίδραση του στη βιομηχανία και στους επιβάτες, καθώς και τις δυνατότητες που προσφέρει η τεχνολογία για την αντιμετώπισή του.

Στη συνέχεια, διερευνάται η δήλωση του προβλήματος, προσδιορίζοντας τις κύριες πτυχές που πρέπει να ληφθούν υπόψη για αυτή την έρευνα. Οι ερευνητικοί στόχοι που έχουν τεθεί αποσκοπούν στην ανάπτυξη λύσεων που θα βελτιώσουν την κατάσταση, ενώ ταυτόχρονα αναγνωρίζεται η εμβέλεια των περιορισμών της παρούσας έρευνας. Τέλος, παρέχεται μια σύντομη περίληψη των κύριων θεμάτων, που θα συζητηθούν στη συνέχεια της διατριβής.

Γίνεται αντιληπτό ότι πέρα από την οικονομική διάσταση του φαινομένου είναι αξιοσημείωτη και η κοινωνική πτυχή, αφού όπως προαναφέρθηκε, οι αεροπορικές εταιρείες διαδραματίζουν καταλυτικό ρόλο στο φαινόμενο της παγκοσμιοποίησης που έχει αλλάξει ριζικά τα δεδομένα της σύγχρονης εποχής.

1.2 Υπόβαθρο και Κίνητρο

Οι καθυστερήσεις μπορεί να προκληθούν από διάφορους λόγους, όπως καιρικές συνθήκες, ανθρώπινα λάθη, ή ακόμα και τεχνικά προβλήματα. Η πρόβλεψη των καθυστερήσεων μπορεί να συμβάλει στην αποτελεσματικότερη διαχείριση των πτήσεων, ελαχιστοποιώντας τις αρνητικές επιπτώσεις τους. Με την αύξηση της διαθεσιμότητας των δεδομένων και της υπολογιστικής ισχύος, η χρήση τεχνικών μηχανικής μάθησης καθίσταται ολοένα και πιο δημοφιλής στον τομέα της αεροπορίας. Αυτό το φαινόμενο έχει οδηγήσει σε μία επανάσταση στον τρόπο με τον οποίο οι αεροπορικές εταιρείες αντιμετωπίζουν τις καθυστερήσεις, επιτρέποντας την ανάπτυξη πιο προηγμένων και ακριβών μοντέλων πρόβλεψης. [9] Η χρήση τεχνικών μηχανικής μάθησης στην πρόβλεψη καθυστερήσεων των αεροπορικών πτήσεων αποτελεί ένα ενδιαφέρον πεδίο έρευνας, καθώς συνδυάζει την ανάλυση μεγάλων όγκου δεδομένων με την πρακτική εφαρμογή σε έναν σημαντικό τομέα που επηρεάζει τη καθημερινότητα εκατομμυρίων ανθρώπων σε όλο τον κόσμο [1]. Η παρούσα εργασία φιλοδοξεί να συμβάλει κατά το δυνατόν στην ελαχιστοποίηση του προβλήματος και θέτει την επιστήμη και την έρευνα στην υπηρεσία του ανθρώπου.

1.3 Δήλωση Προβλήματος

Στόχος αυτής της έρευνας είναι η ανάπτυξη ενός μοντέλου που προβλέπει τις καθυστερήσεις πτήσεων με βάση ένα σύνολο δεδομένων που καλύπτει τα έτη 2018-2022. Το πρόβλημα των καθυστερήσεων των αεροπορικών πτήσεων είναι πολυδιάστατο και επηρεάζεται από πολλούς παράγοντες, όπως οι καιρικές συνθήκες, οι ανθρώπινοι παράγοντες και τα τεχνικά προβλήματα των αεροσκαφών και άλλων μέσων που συμβάλλουν στην οργάνωση και εκπλήρωση μίας πτήσης. Παρόλο που το μοντέλο κώδικα εκπαιδεύτηκε μόνο με τα δεδομένα του 2018, το σύνολο των δεδομένων περιλαμβάνει πληροφορίες για πολλαπλά έτη, με ίδια στοίχιση, καθιστώντας το μοντέλο εύχρηστο για όλα τα έτη και περισσότερα που δίνονται σε αυτή τη μορφή πληροφορίας και προσφέροντας έτσι μια ευρεία προοπτική στην ανάλυση των καθυστερήσεων αεροπορικών πτήσεων. Η δυνατότητα ανάλυσης τέτοιου είδους δεδομένων προσφέρει μια μοναδική ευκαιρία να κατανοήσουμε και να ερμηνεύσουμε τις υποκείμενες αιτίες των καθυστερήσεων. Επιπλέον, με την εκμετάλλευση των τεχνικών μηχανικής μάθησης, το μοντέλο έχει τη δυνατότητα να αναγνωρίσει ποιοι παράγοντες συνεισφέρουν περισσότερο στις καθυστερήσεις και να προτείνει στρατηγικές για την πιο αποτελεσματική διαχείριση και πρόληψη των καθυστερήσεων. Με αυτόν τον τρόπο, οι αεροπορικές εταιρείες μπορούν να λαμβάνουν ενημερωμένες αποφάσεις για τη βελτίωση της λειτουργίας τους, ενώ ταυτόχρονα μειώνουν τις επιπτώσεις των καθυστερήσεων στους επιβάτες τους, δημιουργώντας καλύτερες συνθήκες για όλα τα μέλη που συμβάλουν στην πτήση.

Όλοι έχουμε δει εικόνες από πολλά αεροδρόμια του κόσμου, όπου κάποιες φορές συνωστίζονται επιβάτες ή χάνεται ο έλεγχος εξαιτίας συχνά αδικαιολόγητων καθυστερήσεων που αλλάζουν ή και ματαιώνουν τον προγραμματισμό των ταξιδιωτών. Φαινόμενα έντονων συναισθηματικών εκρήξεων ή ακόμη και βίαιης συμπεριφοράς, θα μπορούσαν να αποφευχθούν με την πρόβλεψη των καθυστερήσεων και της έγκαιρης ενημέρωσης των επιβατών.

1.4 Ερευνητικοί Στόχοι

Ο κύριος στόχος της παρούσας εργασίας είναι η ανάπτυξη ενός ακριβούς και αξιόπιστου μοντέλου πρόβλεψης καθυστέρησης των αεροπορικών πτήσεων. Αυτό απαιτεί την εκπαίδευση του μοντέλου με τα διαθέσιμα δεδομένα, την επικύρωση της απόδοσής του και την τροποποίησή του με βάση τα αποτελέσματα. Δευτερεύων στόχος θα μπορούσε να θεωρηθεί η ανάλυση των παραγόντων, εξετάζοντας τις κύριες παραμέτρους που επηρεάζουν τις καθυστερήσεις πτήσεων, όπως καιρικές συνθήκες, λειτουργικές ανακολουθίες, ανθρώπινα λάθη και άλλα. Ακόμα ένας μερικότερος στόχος θα μπορούσε να θεωρηθεί η κατανόηση των τάσεων. Πιο συγκεκριμένα, η παρακολούθηση των τάσεων καθυστερήσεων κατά τη διάρκεια των διαφορετικών ετών, για να επιτευχθεί η απόλυτη κατανόηση των λόγων που οι καθυστερήσεις εξελίσσονται με τον καιρό. Επίσης, ίσως αυτόματα επιτευχθεί και η βελτίωση των υπηρεσιών για τις αεροπορικές εταιρείες, ως αποτέλεσμα των υπόλοιπων ερευνητικών στόχων. Δηλαδή, με βάση τα ευρήματα, ίσως εμφανιστούν προτάσεις για τη βελτίωση των υπηρεσιών προς τους επιβάτες, με στόχο τη μείωση των επιπτώσεων των καθυστερήσεων. Τέλος, θα υπάρξει ορίζοντας για νέες τακτικές πρόβλεψης αεροπορικών καθυστερήσεων, αν επιτευχθεί η ανακάλυψη και εφαρμογή νέων προσεγγίσεων και τεχνικών που μπορούν να βελτιώσουν την ακρίβεια των προβλέψεων. Αυτό μπορεί να περιλαμβάνει τη χρήση προηγμένων τεχνικών μηχανικής μάθησης ή την ενσωμάτωση επιπλέον πηγών δεδομένων.

1.5 Εμβέλεια και Περιορισμοί

Όπως έχει ήδη αναφερθεί, η έρευνα βασίζεται στα δεδομένα αεροπορικών πτήσεων της περιόδου 2018-2022. Επίσης χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης για την εκπαίδευση, την εύρεση αποτελεσματικότητας και την πρόβλεψη του μοντέλου. Ο χρόνος που απαιτείται για την εκπαίδευση αυτή είναι πολύς, μιας και επεξεργαζόμαστε μεγάλο όγκο δεδομένων. Παράλληλα, η ακρίβεια των προβλέψεων εξαρτάται από την ποιότητα και την πληρότητα των δεδομένων, οι οποίες δεν είναι ανεξάντλητες. Τέλος, αξίζει να αναφερθεί πως, η εφαρμογή τεχνικών μηχανικής μάθησης σε πραγματικά δεδομένα μπορεί να παρουσιάσει προκλήσεις, όπως υπερεκπαίδευση, ανισορροπία δεδομένων και άλλα.

2. Ανασκόπηση Βιβλιογραφίας Έρευνας

Η ανασκόπηση της βιβλιογραφίας αποτελεί ουσιαστικό τομέα της έρευνας, όπου εξετάζονται οι προηγούμενες σχετικές μελέτες και οι τεχνικές που έχουν χρησιμοποιήσει στον τομέα της αεροπορίας και πιο συγκεκριμένα, τις περισσότερες φορές, στις καθυστερήσεις αεροπορικών πτήσεων. Αρχικά, παρουσιάζεται μία επισκόπηση της ανάλυσης μεγάλων δεδομένων στην αεροπορία, υπογραμμίζοντας τη σημασία της ανάλυσης στην αντιμετώπιση προκλήσεων, όπως οι καθυστερημένες αεροπορικές πτήσεις. Στη συνέχεια, γίνεται αναφορά σε προηγούμενες μελέτες που έχουν επικεντρωθεί στην πρόβλεψη καθυστερημένων αεροπορικών πτήσεων, παρέχοντας μία βασική κατανόηση των μεθόδων και των τεχνικών που έχουν χρησιμοποιηθεί μέχρι σήμερα. , που είναι εφαρμόσιμες στον τομέα, καθώς και οι μετρικές αξιολόγησης που χρησιμοποιούνται για την αξιολόγηση των προβλεπτικών μοντέλων της κάθε έρευνας. Συνολικά, αυτή η ενότητα παρέχει μια ολοκληρωμένη εικόνα των τρεχουσών γνώσεων και πρακτικών στον τομέα της αεροπορίας.

Τα παραπάνω δεδομένα προσφέρουν μια σφαιρικότερη εικόνα του προβλήματος και αποτελούν υπόβαθρο για πιο διεξοδική μελέτη και κατανόηση του φαινομένου.

2.1 Επισκόπηση της Ανάλυσης Δεδομένων στην Αεροπορία

Η αεροπορία, ως βασικός πυλώνας της παγκοσμιοποίησης, έχει επαναπροσδιορίσει τον τρόπο με τον οποίο ταξιδεύουμε, επικοινωνούμε και συνεργαζόμαστε. Ωστόσο, οι καθυστερήσεις πτήσεων παραμένουν μια σημαντική πρόκληση που επηρεάζει τόσο τους επιβάτες των πτήσεων όσο και τις αεροπορικές εταιρείες. Η ανάλυση μεγάλων όγκου δεδομένων προσφέρει τη δυνατότητα να αντιμετωπίσουμε αυτό το πρόβλημα με πιο εξελιγμένο τρόπο. Χρησιμοποιώντας σύγχρονες τεχνικές μηχανικής μάθησης, όπως ο αλγόριθμος Random Forest Regressor, οι εταιρείες μπορούν να αναλύσουν τεράστια σύνολα δεδομένων για να προβλέψουν πιθανές καθυστερήσεις και να λάβουν προληπτικά μέτρα, ώστε να ενημερώσουν τους επιβάτες τους ή και να είναι έτοιμες για ελέγχους που μπορούν να αποτρέψουν προβλήματα τελευταίας στιγμής. [5] Επιπλέον, η ανάλυση μεγάλου όγκου δεδομένων επιτρέπει την ανάλυση διαφόρων παραγόντων που μπορεί να

προκαλέσουν καθυστερήσεις, όπως καιρικές συνθήκες, ανθρώπινοι παράγοντες ή τεχνικά προβλήματα. Με αυτόν τον τρόπο οι αεροπορικές εταιρείες μπορούν να βελτιώσουν την αποτελεσματικότητά τους, να μειώσουν τις οικονομικές απώλειες και να προσφέρουν καλύτερη εμπειρία στους πελάτες τους. Οποιοδήποτε πρόβλημα ή φαινόμενο μπορεί να αντιμετωπιστεί αποτελεσματικότερα με τη βοήθεια της επιστήμης και των δεδομένων που υπάρχουν, αφού με αυτόν τον τρόπο η κατανόηση και κυρίως η ερμηνεία γίνεται πιο ενδεδειγμένη φτάνοντας στην ουσία του εξεταζόμενου θέματος [2].

2.2 Προηγούμενες Μελέτες για την Πρόβλεψη Καθυστερήσεων

Πτήσεων

Οι καθυστερήσεις πτήσεων αποτελούν ένα από τα πιο σημαντικά προβλήματα στον τομέα της αεροπορίας, με σημαντικές οικονομικές και λειτουργικές επιπτώσεις [1]. Εξ' αιτίας της βαρύτητας του θέματος, πολλές μελέτες έχουν διεξαχθεί με σκοπό την πρόβλεψη και την αντιμετώπιση των καθυστερήσεων. Στο πλαίσιο αυτό, οι ερευνητές έχουν χρησιμοποιήσει διάφορες μεθοδολογίες και τεχνικές. Ειδικότερα, έχουν εφαρμοστεί αλγόριθμοι μηχανικής μάθησης όπως το Gradient Boosting Machine, το Random Forest Regressor και τα νευρωνικά δίκτυα [3]. Κάθε μέθοδος παρουσιάζει τα δικά της πλεονεκτήματα και τις δικές της αδυναμίες, αλλά όλες στοχεύουν στην ακριβή πρόβλεψη των καθυστερήσεων. Είναι αξιοσημείωτο ότι η εξέλιξη των τεχνολογιών και η διαθεσιμότητα μεγάλων συνόλων δεδομένων έχουν επιτρέψει την εφαρμογή πιο προηγμένων τεχνικών και τη βελτίωση της ακρίβειας των προβλέψεων. Η αναδρομή σε προηγούμενες μελέτες, εξάλλου, θεωρείται πάγια και απαραίτητη μέθοδος οποιασδήποτε ερευνητικής προσπάθειας. [7]

Στις προσπάθειες αυτές, μεγάλο ρόλο παίζει η ανάλυση και η επεξεργασία των δεδομένων. Τα σύγχρονα συστήματα ανάλυσης βασίζονται σε τεχνολογίες μεγάλου όγκου δεδομένων, οι οποίες επιτρέπουν την επεξεργασία και ανάλυση τεράστιων όγκων δεδομένων σε πραγματικό χρόνο. Αυτά τα δεδομένα περιλαμβάνουν πληροφορίες για τις καιρικές συνθήκες, την κίνηση των αεροσκαφών, την κατάσταση των αεροδρομίων και πολλά

άλλα. Η ορθή επεξεργασία και ανάλυση αυτών των δεδομένων μπορεί να παρέχει σημαντική επίγνωση για την πρόβλεψη καθυστερήσεων. [1]

Εκτός από την ανάλυση δεδομένων, η ενσωμάτωση καινοτόμων αλγορίθμων μηχανικής μάθησης συμβάλλει στην ακρίβεια των προβλέψεων. Η εφαρμογή μεθόδων όπως η βαθιά μάθηση (deep learning) και τα ενισχυτικά νευρωνικά δίκτυα (reinforcement learning) ανοίγει νέους δρόμους για την προβλεπτική ανάλυση. [13] Αυτές οι τεχνικές μπορούν να εντοπίσουν πολύπλοκα μοτίβα και σχέσεις μεταξύ των δεδομένων, που μια πιο παραδοσιακή προσέγγιση ίσως να αγνοούσε. Η χρήση τους ενισχύει την πρόβλεψη καθυστερήσεων πτήσεων με βάση την ενσωμάτωση διαφόρων παραμέτρων και στοιχείων. [2]

Η πρόκληση, ωστόσο, δεν περιορίζεται μόνο στην επιλογή και εφαρμογή των κατάλληλων αλγορίθμων, αλλά επεκτείνεται και στη διαχείριση και την εγγυημένη ποιότητα των δεδομένων. Η καθαριότητα, η ακρίβεια και η ενημερότητα των δεδομένων είναι κρίσιμες πτυχές που επηρεάζουν άμεσα τα αποτελέσματα των αλγορίθμων. Η εξασφάλιση υψηλής ποιότητας δεδομένων και η συνεχής ενημέρωσή τους συμβάλλουν στην παραγωγή πιο αξιόπιστων και ακριβών προβλέψεων, καθιστώντας την επιστήμη δεδομένων αναπόσπαστο μέρος της διαδικασίας πρόβλεψης καθυστερήσεων στις πτήσεις. [1]

Η ενσωμάτωση πολλαπλών πηγών δεδομένων αποτελεί ένα ακόμη βασικό στοιχείο στην ακριβή πρόβλεψη καθυστερήσεων πτήσεων. Η συλλογή δεδομένων από διάφορες πηγές, όπως μετεωρολογικοί σταθμοί, συστήματα ελέγχου της κυκλοφορίας των αεροσκαφών και ιστορικά δεδομένα πτήσεων, επιτρέπει μια πιο ολιστική προσέγγιση. Η συνδυαστική ανάλυση αυτών των δεδομένων μέσω προηγμένων αλγορίθμων βοηθά στην πρόβλεψη καθυστερήσεων με μεγαλύτερη ακρίβεια, λαμβάνοντας υπόψη πολυδιάστατες παραμέτρους και δυναμικές. [3]

Ακόμη, η συνεχής ανάπτυξη και βελτίωση των μεθόδων μηχανικής μάθησης δίνει τη δυνατότητα για πιο ανθεκτικά και ευέλικτα μοντέλα πρόβλεψης. Τέτοια μοντέλα μπορούν να προσαρμόζονται γρήγορα σε νέες συνθήκες και αλλαγές, όπως οι απρόβλεπτες καιρικές

συνθήκες ή αλλαγές στις λειτουργίες των αεροδρομίων. Η ευελιξία αυτή είναι ζωτικής σημασίας για την αποδοτική πρόβλεψη καθυστερήσεων σε έναν τόσο δυναμικό και περίπλοκο τομέα όπως η αεροπορία. [9]

Τέλος, η ενσωμάτωση τεχνολογιών μεγάλου όγκου δεδομένων και μηχανικής μάθησης στον τομέα της αεροπορίας ανοίγει το δρόμο για πιο προηγμένες εφαρμογές πέρα από την απλή πρόβλεψη καθυστερήσεων. Η ανάλυση δεδομένων μπορεί να συμβάλλει στην βελτίωση της διαχείρισης της κυκλοφορίας των αεροσκαφών, στην αυξημένη ασφάλεια των πτήσεων και στην βελτίωση της εμπειρίας των επιβατών. Η δυνατότητα αυτή δείχνει την αξία της ενσωμάτωσης των εν λόγω τεχνολογιών σε πολλαπλές πτυχές της αεροπορικής βιομηχανίας, επισημαίνοντας την πολυδιάστατη σημασία τους. [4]

2.3 Σχετικές Τεχνικές Εξόρυξης Δεδομένων και Μηχανικής Μάθησης

Η εξόρυξη δεδομένων αποτελεί μια διαδικασία ανακάλυψης χρήσιμων πληροφοριών από μεγάλου όγκου σύνολα δεδομένων. Σε συνδυασμό με τη μηχανική μάθηση, οι τεχνικές αυτές επιτρέπουν την ανάλυση και την πρόβλεψη συμπεριφορών και τάσεων, όπως οι καθυστερήσεις αεροπορικών πτήσεων. Τα βασικά βήματα στην εξόρυξη μεγάλου όγκου δεδομένων και τη μηχανική μάθηση περιλαμβάνουν τη προ-επεξεργασία των δεδομένων, την επιλογή απαραίτητων χαρακτηριστικών και αφαίρεση θορυβωδών χαρακτηριστικών, την εκπαίδευση του μοντέλου και την αξιολόγηση της απόδοσής του. Η προ-επεξεργασία των δεδομένων μπορεί να θεωρηθεί κρίσιμο βήμα, καθώς τα ατελή ή ανακριβή δεδομένα μπορούν να επηρεάσουν αρνητικά την ακρίβεια του μοντέλου και να φέρουν ανεπιθύμητα αποτελέσματα και παρενέργειες. Επιπλέον, η επιλογή του κατάλληλου αλγορίθμου μηχανικής μάθησης αποτελεί ένα θέμα ζωτικής σημασίας στην έρευνα και σε παρόμοιες έρευνες πρόβλεψης αεροπορικών καθυστερήσεων. Ανάλογα με τη φύση των δεδομένων και το πρόβλημα που προσπαθούμε να λύσουμε, διαφορετικοί αλγόριθμοι μπορεί να είναι πιο κατάλληλοι από άλλους. Συνολικά, η σωστή εφαρμογή των τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης μπορεί να προσφέρει βαθύτερη κατανόηση των παραγόντων που επηρεάζουν τις καθυστερήσεις αεροπορικών πτήσεων και να βοηθήσει στην ανάπτυξη στρατηγικών για την αντιμετώπισή τους.

2.4 Συμπεράσματα από την Ανασκόπηση της Βιβλιογραφίας

Η ανασκόπηση της βιβλιογραφίας καταδεικνύει τη σημασία της πρόβλεψης των καθυστερήσεων πτήσεων στον τομέα της αεροπορίας. Η χρήση τεχνικών μηχανικής μάθησης για την αντιμετώπιση αυτού του προβλήματος έχει αποδειχθεί ως μια αποτελεσματική προσέγγιση, με πολλές μελέτες να επικεντρώνονται στη βελτίωση της ακρίβειας και της αποδοτικότητας των μοντέλων πρόβλεψης. Είναι επίσης σημαντικό να σημειωθεί ότι, παρά τις προκλήσεις, οι τεχνικές αυτές προσφέρουν μια ευκαιρία για τις αεροπορικές εταιρείες να βελτιώσουν την εμπειρία των επιβατών τους, μειώνοντας τις καθυστερήσεις και βελτιστοποιώντας τη λειτουργία τους.

Σε συνάρτηση με τα παραπάνω, η συνεχής έρευνα και ανάπτυξη σε αυτόν τον τομέα είναι απαραίτητη για την κατανόηση των παραγόντων που επηρεάζουν τις καθυστερήσεις αεροπορικών πτήσεων και την ανάπτυξη πιο ακριβών και αποτελεσματικών μοντέλων πρόβλεψης.

Το φαινόμενο των καθυστερήσεων στις αεροπορικές πτήσεις αποτελεί ένα λαμπρό πεδίο έρευνας εξαιτίας της σημασίας και της καθολικότητας που το χαρακτηρίζει. Μην ξεχνάμε ότι η παγκοσμιοποίηση στηρίζεται σε πολύ μεγάλο βαθμό στην ποιότητα των υπηρεσιών που προσφέρει.

3. Συλλογή και Προ-Επεξεργασία Δεδομένων

3.1 Πηγή Δεδομένων και Περιγραφή

Η ακρίβεια και η αξιοπιστία των αποτελεσμάτων οποιασδήποτε αναλυτικής έρευνας εξαρτώνται σε μεγάλο βαθμό από την ποιότητα των δεδομένων που χρησιμοποιούνται. Στο κεφάλαιο αυτό, εξετάζεται η πηγή των δεδομένων, περιγράφοντας τη φύση και τα χαρακτηριστικά τους. Η μεθοδολογία που υιοθετήθηκε για τη συλλογή των δεδομένων αποτελεί ένα κρίσιμο βήμα, καθώς διασφαλίζει την ακεραιότητα και την αξιοπιστία της πληροφορίας. Μετά τη συλλογή, πραγματοποιείται μια εξερευνητική ανάλυση των αεροπορικών δεδομένων, αποκαλύπτοντας τις πρώτες ενδείξεις και τάσεις. Η προ-επεξεργασία των δεδομένων θεωρείται ως ένα απαραίτητο βήμα, καθώς επιτρέπει στην αφαίρεση των ατελειών, την κανονικοποίηση και την τροποποίηση των δεδομένων για τη βέλτιστη ανάλυση τους. Τέλος, η επιλογή και η έρευνα απόδοσης των χαρακτηριστικών, διαδραματίζουν κρίσιμο ρόλο στον προσδιορισμό των πληροφοριών που θα χρησιμοποιηθούν στις επόμενες φάσεις της έρευνας.

Η πηγή των δεδομένων που χρησιμοποιήθηκαν για την παρούσα έρευνα είναι το Kaggle, μια δημοφιλής πλατφόρμα για την ανάλυση δεδομένων και τη μηχανική μάθηση. Το σύνολο δεδομένων που επιλέχθηκε από το Kaggle παρέχει λεπτομερείς πληροφορίες σχετικά με τις αεροπορικές πτήσεις που πραγματοποιήθηκαν από τα έτη 2018 έως 2022. Αυτό το σύνολο δεδομένων δεν περιορίζεται μόνο στις βασικές πληροφορίες των πτήσεων, αλλά επεκτείνεται για να καλύψει και άλλες σημαντικές πτυχές, όπως οι καιρικές συνθήκες κατά τη διάρκεια της πτήσης, οι αεροπορικές εταιρείες που εκτελούν τις πτήσεις και τα αεροδρόμια από τα οποία αναχωρούν και στα οποία προσγειώνονται τα αεροσκάφη με τους επιβάτες. Η πληροφορία αυτή είναι ζωτικής σημασίας για τη κατανόηση των διαφόρων παραγόντων που μπορεί να επηρεάσουν τις καθυστερήσεις πτήσεων και για τη δημιουργία ακριβούς μοντέλων πρόβλεψης. Το dataset μπορεί να βρεθεί στον ακόλουθο σύνδεσμο: <https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022>

3.2 Μεθοδολογία Συλλογής Δεδομένων

Η συλλογή δεδομένων για την παρούσα έρευνα βασίζεται σε αυτοματοποιημένες διαδικασίες που εξάγουν πληροφορίες από τον πίνακα «Marketing Carrier On-Time Performance» της βάσης δεδομένων «On-Time». Η πηγή περιλαμβάνει τα ακόλουθα στοιχεία, αποτελώντας με αυτόν τον τρόπο ένα ολοκληρωμένο σύνολο μεγάλου όγκου δεδομένων και προσφέροντας όλες τις απαραίτητες πληροφορίες για κάθε είδους έρευνας:

Marketing Carrier On Time
Χρονολογία/Year
Εποχή/Quarter
Μήνα/Month
Μέρα του μήνα/ Day of the month
Μέρα της εβδομάδας/ Day of the Week
Μέρα πτήσης/ Flight day
Μοναδικοί κωδικοί αεροπορικών εταιρειών/ Marketing Airline Network
Μοναδικοί κωδικοί συνεργαζόμενων εταιρειών/ Operated or branded code share partners
Κωδικοί DOT 5/ DOT ID marketing Airline
Κωδικοί IATA 6/ IATA code marketing airline
Αριθμός αεροπορικής πτήσης/ Flight Number Marketing Airline
Νούμερο ουράς/Tail number
Νούμερο πτήσης/ Flight number operating airline
ID αρχικού αεροδρομίου/Origin airport ID
USDOTID αρχικού αεροδρομίου/ Origin airport USDOTID

USDOTCITYMARKETID (origin city market ID)
Όνομα αρχικού αεροδρομίου/ Origin(airport?)
Όνομα αρχικής πόλης/ Origin city name
Όνομα αρχικής πολιτείας/ Origin State
Κωδικός πολιτείας/ Origin state fips
Κωδικός αρχικής πολιτείας/ Origin state code
Όνομα αρχικής χώρας/ Origin country
Κωδικός αρχικής χώρας/ Origin country code
Περιοχή αρχικού αεροδρομίου/Origin wac
ID αεροδρομίου προορισμού / Dest airport ID
USDOTID αεροδρομίου προορισμού/ DestAirportseqID
USDOTCITYMARKETID προορισμού/ DestCityMarketID
Όνομα αεροδρομίου προορισμού/ Dest
Όνομα πόλης προορισμού/ DestCityName
Όνομα πολιτείας προορισμού/ DestStae
Κωδικός πολιτείας προορισμού/ DestStateCode
Όνομα χώρας προορισμού/ DestCountry
Κωδικός χώρας προορισμού/ DestCountryCode
Περιοχή αεροδρομίου προορισμού/ DestWac
Συνολικοί επιβάτες: TotalPassengers

Η διαδικασία συλλογής των ετών της έρευνας ήταν σχεδιασμένη να λειτουργεί συνεχώς, διασφαλίζοντας έτσι την τακτική ενημέρωση και την ακρίβεια των πληροφοριών. Επιπλέον, βάσει των πληροφοριών που παρέχονται από το Kaggle, τα δεδομένα ανήκαν στη βιβλιοθήκη δεδομένων *Tran Stats*. Η χρονική περίοδος που καλύπτεται από τα δεδομένα αναγράφεται στο όνομα του συμπιεσμένου αρχείου του συνόλου. Για παράδειγμα, το αρχείο με όνομα «XXX_XXXXX_2001_1» περιέχει δεδομένα του πρώτου μήνα του έτους 2001.

3.3. Εξερεύνηση Ανάλυσης Δεδομένων

Η εξερεύνηση ανάλυσης δεδομένων αποτελεί ένα κρίσιμο βήμα στη διαδικασία επεξεργασίας δεδομένων και μοντελοποίησης. Στην παρούσα έρευνα, η εξερεύνηση ανάλυσης δεδομένων βοήθησε στη κατανόηση των βασικών χαρακτηριστικών του συνόλου δεδομένων και στην αναγνώριση πιθανών προκλήσεων που θα μπορούσαν να επηρεάσουν την απόδοση του μοντέλου.

Κατά τη διάρκεια της εξερεύνησης ανάλυσης δεδομένων εξετάστηκαν βασικές στατιστικές μέτρησης, όπως μέσος όρος, διάκριση, ελάχιστες και μέγιστες τιμές, για τη κατανόηση της γενικής κατανομής των δεδομένων. Ακόμα, πραγματοποιήθηκε αναγνώριση και επεξεργασία των ατυπιών και των απουσιάζουσών τιμών, οι οποίες θα μπορούσαν να επηρεάσουν την ακρίβεια του μοντέλου και να προκαλέσουν απρόβλεπτη συμπεριφορά ή και την απρόσμενη διακοπή της εκπαίδευσης του μοντέλου. Τέλος, με τη χρήση διαγραμμάτων και άλλων γραφικών απεικονίσεων, εξετάστηκαν οι σχέσεις μεταξύ διαφορετικών χαρακτηριστικών, καθώς και η σημασία τους στην πρόβλεψη των καθυστερημένων αεροπορικών πτήσεων. Με την ολοκλήρωση της εξερεύνησης ανάλυσης δεδομένων, η έρευνα ήταν πλέον έτοιμη να προχωρήσει στην εκπαίδευση του μοντέλου, με στόχο την πρόβλεψη των καθυστερήσεων πτήσεων με την μεγαλύτερη δυνατή ακρίβεια.

Παρά τις αναμενόμενες δυσκολίες καταβλήθηκε προσπάθεια για την πιο αποτελεσματική αξιοποίηση των δεδομένων, με απώτερο στόχο τη σφαιρικότερη προσέγγιση του προβλήματος.

3.4. Τεχνικές Καθαρισμού και Προ-Επεξεργασίας Δεδομένων

Η προ-επεξεργασία δεδομένων αποτελεί ένα από τα πιο σημαντικά βήματα στη διαδικασία εξόρυξη γνώσης από δεδομένα μεγάλου όγκου. Στην παρούσα έρευνα, οι παρακάτω τεχνικές προ-επεξεργασίας εφαρμόστηκαν για τη βελτίωση της ποιότητας του συνόλου δεδομένων:

- **Αφαίρεση διπλοτύπων:** Οι διπλότυπες εγγραφές αφαιρέθηκαν για να αποφευχθεί η κάθε πιθανή παραπλανητική επίδραση στο μοντέλο.
- **Συμπλήρωση Απουσιαζουσών Τιμών:** Οι απουσιάζουσες τιμές συμπληρώθηκαν με τη χρήση τεχνικών, όπως η μεσαία τιμή για συνεχείς μεταβλητές ή η πιο συχνή τιμή για κατηγορικές μεταβλητές.
- **Κανονικοποίηση Δεδομένων:** Τα δεδομένα κανονικοποιήθηκαν για να έχουν μία κοινή κλίμακα, κάτι που είναι ιδιαίτερα χρήσιμο για αλγορίθμους που βασίζονται σε αποστάσεις, όπως ο K-πλησιέστερος γείτονας.
- **Μετατροπή Κατηγορικών Μεταβλητών:** Οι κατηγορικές μεταβλητές μετατράπηκαν σε αριθμητικές με τη χρήση τεχνικών όπως one-hot encoding 5, ώστε να μπορούν να είναι επεξεργάσιμες από τους αλγορίθμους μηχανικής μάθησης.
- **Εφαρμογή:** Με την εφαρμογή αυτών των τεχνικών, το σύνολο δεδομένων έγινε πιο καθαρό και οργανωμένο, χωρίς παρενέργειες για ανακριβή αποτελέσματα λόγω σωστών αποφάσεων, προετοιμάζοντάς το για την εκπαίδευση του μοντέλου.

3.5. Επιλογή και Μηχανισμός Χαρακτηριστικών

Η επιλογή των κατάλληλων χαρακτηριστικών αποτελεί ένα βασικό βήμα στη διαδικασία της μηχανικής μάθησης. [6] Η σωστή επιλογή χαρακτηριστικών μπορεί να βελτιώσει σημαντικά την απόδοση ενός μοντέλου, ενώ η επιλογή λανθασμένων ή περιττών χαρακτηριστικών μπορεί να οδηγήσει σε χειρότερα αποτελέσματα. Μετά την εξερευνητική ανάλυση, ορισμένα χαρακτηριστικά που κατέδειξαν σημαντική συσχέτιση με τις καθυστερήσεις πτήσεων επιλέχθηκαν (επιλεγμένα χαρακτηριστικά). Αυτά περιλαμβάνουν:

Τελικά Επιλεγμένα Δεδομένα με παραμετροποιήσεις ονόματος
Χρονολογία
Εποχή
Μήνα
Μέρα της εβδομάδας
Μέρα πτήσης
Μοναδικοί κωδικοί αεροπορικών εταιρειών
Όνομα αεροπορικής εταιρίας
ID αρχικού αεροδρομίου
Όνομα πόλης προορισμού
ID προορισμού αεροδρομίου
Χρόνος Αρχής Πτήσης
Καθυστέρηση Πτήσης
Καθυστέρηση Πτήσης (Λίστα)
Ακύρωση Πτήσης
Αλλαγή Πορείας Πτήσης

Τα υπόλοιπα χαρακτηριστικά δεν είχαν συνάφεια με τις καθυστερήσεις πτήσεων ή ήταν περιττά για την ανάλυση και αγνοήθηκαν (αγνοούμενα χαρακτηριστικά). Τα περισσότερα

από αυτά είναι λεπτομέρειες που δε μπορούν να συμβάλουν στις καθυστερήσεις. Εκτός από την απλή επιλογή χαρακτηριστικών, έχουν εφαρμοστεί και τεχνικές για τη δημιουργία νέων χαρακτηριστικών ή τη μετατροπή των υπαρχόντων με σκοπό την καλύτερη αναπαράσταση. Η σωστή επιλογή και ο μηχανισμός των χαρακτηριστικών είναι κρίσιμα για την κατασκευή ενός αποτελεσματικού και ακριβούς μοντέλου πρόβλεψης.

4. Μεθοδολογία

4.1 Επισκόπηση της Προβλεπτικής Μοντελοποίησης

Η έρευνα στον τομέα της αεροπορίας απαιτεί σταθερή και συστηματική προσέγγιση για την ανάλυση και πρόβλεψη των καθυστερήσεων αεροπορικών πτήσεων. Στο πλαίσιο αυτό, η επισκόπηση της προβλεπτικής μοντελοποίησης παρέχει μία βασική εισαγωγή στις βασικές τεχνικές και μεθόδους που χρησιμοποιούνται σε αυτόν τον τομέα της εργασίας. Επιπλέον, η επιλογή των κατάλληλων αλγορίθμων για την πρόβλεψη των καθυστερήσεων παρουσιάζει βαρύνουσα σημασία, καθώς ορισμένοι αλγόριθμοι μπορεί να είναι πιο αποτελεσματικοί από άλλους, ανάλογα με τη φύση των μεγάλου όγκου δεδομένων. Με την ανάπτυξη και την εκπαίδευση του μοντέλου, επιδιώκουμε τη βελτιστοποίηση των προβλέψεών μας. Η αξιολόγηση και η επικύρωση του μοντέλου είναι απαραίτητες για να διασφαλίσουμε την ακρίβεια και την αξιοπιστία των προβλέψεων. Τέλος η σύγκριση με άλλα μοντέλα επιτρέπει την κατανόηση της αποτελεσματικότητάς της προσέγγισής μας σε σχέση με άλλες τεχνικές που χρησιμοποιούνται στη βιομηχανία αυτού του τομέα. Η προβλεπτική μοντελοποίηση είναι ένας τομέας της μηχανικής μάθησης και της στατιστικής που εστιάζει στη χρήση δεδομένων για τη πρόβλεψη μελλοντικών γειτόνων. Αυτό επιτυγχάνεται μέσω της εκπαίδευσης ενός μοντέλου σε ένα σύνολο δεδομένων με γνωστά αποτελέσματα, για να μπορεί στη συνέχεια να προβλέπει τα αποτελέσματα σε νέα, άγνωστα δεδομένα. [7]

Η διαδικασία για την επίτευξη προβλεπτικής μοντελοποίησης περιλαμβάνει τη συλλογή και προ-επεξεργασία των δεδομένων σε κατάλληλη μορφή για την έρευνα, την επιλογή μηχανικών χαρακτηριστικών, την εκπαίδευση του μοντέλου, την αξιολόγηση της απόδοσης του και την εφαρμογή του σε νέα δεδομένα. Κάθε έρευνα παρουσιάζει και την κατάλληλη εφαρμογή αυτής της διαδικασίας. Στην περίπτωση αυτής της έρευνας, το μοντέλο εκπαιδεύτηκε για να προβλέπει τις καθυστερήσεις αεροπορικών πτήσεων με βάση διάφορα χαρακτηριστικά, όπως η ώρα αναχώρησης, ο προορισμός, οι καιρικές συνθήκες και άλλοι παράγοντες. Συμπερασματικά, η προβλεπτική μοντελοποίηση είναι ιδιαίτερα χρήσιμη σε πολλούς τομείς, από την οικονομία και την υγεία, μέχρι τη βιομηχανία και την επιστήμη. Στην συγκεκριμένη περίπτωση, των αεροπορικών πτήσεων, μπορεί να βοηθήσει

τις αεροπορικές εταιρίες να βελτιώσουν την απόδοσή τους και να προσφέρουν καλύτερη εξυπηρέτηση στους επιβάτες των πτήσεων που προσφέρουν.

4.2 Επιλογή Αλγορίθμων για την πρόβλεψη Καθυστερήσεων Πτήσεων

Η επιλογή των κατάλληλων αλγορίθμων μηχανικής μάθησης θεωρείται ζωτικής σημασίας για όλες τις έρευνες που εκτελούνται με αυτό τον τρόπο και επιδιώκουν ως αποτέλεσμα την επίτευξη ακριβών και αξιόπιστων προβλέψεων. Βάσει των χαρακτηριστικών του συνόλου δεδομένων και των ειδικών απαιτήσεων της έρευνας, οι παρακάτω αλγόριθμοι θα μπορούσαν να θεωρηθούν κατάλληλοι.

- **Δέντρα Απόφασης (Decision Trees):** Είναι γραφικοί αλγόριθμοι που μπορούν εύκολα να αξιοποιηθούν τόσο κατηγορικά όσο και ποσοτικά χαρακτηριστικά. Είναι εύκολα κατανοητοί και μπορούν να αναδείξουν σημαντικές σχέσεις μεταξύ των χαρακτηριστικών του συνόλου δεδομένων.
- **Τυχαία Δάση (Random Forests):** Είναι μια συλλογή από δέντρα απόφασης που λειτουργούν αρκετά καλά σε πολλά σύνολα δεδομένων και μπορούν να αντιμετωπίσουν μεγάλο αριθμό χαρακτηριστικών (ή αλλιώς, μεγάλου όγκου σύνολα δεδομένων).
- **Λογιστική Παλινδρόμηση (Logistic Regression):** Είναι ένας αλγόριθμος που μπορεί να προβλέψει την πιθανότητα ενός συμβάντος με βάση ένα ή περισσότερα ανεξάρτητα χαρακτηριστικά.
- **Μηχανικές Διανυσματικής Υποστήριξης (Support Vector Machines):** Είναι αλγόριθμοι που μπορούν να διαχωρίσουν τα δεδομένα σε δύο ή περισσότερες κατηγορίες με την εύρεση του βέλτιστου υπερεπίπεδου.

Η τελική επιλογή του αλγορίθμου για κάθε έρευνα εξαρτάται από τη φύση των δεδομένων, τις απαιτήσεις της έρευνας και τις προτιμήσεις του ερευνητή. Με όλες αυτές τις πληροφορίες στη γνώση αυτής της έρευνας, αποφασίστηκε η εστίαση στους αλγορίθμους random forests.

4.3 Ανάπτυξη και Εκπαίδευση του Μοντέλου

Η ανάπτυξη και εκπαίδευση ενός μοντέλου μηχανικής μάθησης είναι μία διαδικασία που απαιτεί μεγάλη προσοχή στη λεπτομέρεια και στην κατανόηση του συνόλου των δεδομένων της έρευνας. Αρχικά βρισκόμαστε στο στάδιο προετοιμασίας των δεδομένων. Τα δεδομένα διαχωρίζονται σε σύνολα εκπαίδευσης και δοκιμής. Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου, ενώ το σύνολο δοκιμής χρησιμοποιείται για την αξιολόγηση της απόδοσης αυτού του μοντέλου. Έπειτα, προχωράμε στην εκπαίδευση του μοντέλου. Με τη χρήση του επιλεγμένου αλγόριθμου (στη περίπτωση της έρευνας Random Forest Regressor), το μοντέλο εκπαιδεύεται χρησιμοποιώντας τα δεδομένα εκπαίδευσης. Αυτά τα δεδομένα ενδέχεται να είναι είτε ένα ποσοστό ενός συνόλου δεδομένων -ώστε να χρησιμοποιηθεί το υπόλοιπο για δοκιμή του μοντέλου- είτε ένα ολόκληρο σύνολο με τιμές δοκιμής 1-1 σύνολα δεδομένων με διαφορετικές τιμές. Κατά τη διάρκεια αυτής της διαδικασίας, το μοντέλο προσπαθεί να ‘μάθει’ τις σχέσεις μεταξύ των χαρακτηριστικών και των απαραίτητων μεταβλητών. Στο τέλος της διαδικασίας, πραγματοποιείται ή αξιολόγηση μοντέλου. Σε αυτό το στάδιο έχει εκπαιδευτεί και τροποποιηθεί το μοντέλο και αξιολογείται χρησιμοποιώντας το σύνολο δοκιμής. Μετρήσεις όπως η ακρίβεια, η ανάκληση και το F1 σκορ μπορούν να χρησιμοποιηθούν για την κατανόηση της απόδοσης του προβλεπτικού μοντέλου. Η διαδικασία αυτή επαναλαμβάνεται με υποσύνολα των δεδομένων, μέχρι να επιτευχθεί η επιθυμητή απόδοση του προβλεπτικού μοντέλου.

4.4 Αξιολόγηση και Επικύρωση του Μοντέλου

Η αξιολόγηση και η επικύρωση ενός μοντέλου μηχανικής μάθησης μπορούν να θεωρηθούν ως βασικά βήματα για τη διασφάλιση της εγκυρότητας του μοντέλου και της δυνατότητάς του να ανταποκριθεί καλά σε νέα άγνωστα δεδομένα ίδιου τύπου. Τα δεδομένα διαχωρίζονται συνήθως σε τρία σύνολα, ένα σύνολο εκπαίδευσης, ένα σύνολο επικύρωσης και ένα σύνολο ελέγχου. Το σύνολο ελέγχου χρησιμοποιείται για την τελική αξιολόγηση του προβλεπτικού μοντέλου. Στη διαδικασία αξιολόγησης χρησιμοποιούνται διάφορες μετρήσεις (Μετρήσεις Αξιολόγησης), όπως η ακρίβεια, η ανάλυση, το F1 σκορ και η περιοχή κάτω από την καμπύλη ROC (AUC-ROC) [13], για να αξιολογηθεί η απόδοση του μοντέλου μηχανικής μάθησης. Η ικανότητα του μοντέλου να προβλέπει σωστά σε νέα, άγνωστα, δεδομένα που δεν έχει δει κατά τη διάρκεια της εκπαίδευσης, αποτελεί ένδειξη καλής γενίκευσης στοιχείο που ουσιαστικά αποτελεί καλή ένδειξη στην πορεία της σωστής κατασκευής ενός μοντέλου μηχανικής μάθησης. Είναι επίσης σημαντικό να γίνεται σωστή παρατήρηση στο αν το μοντέλο επιδεικνύει διαφορετική απόδοση στο σύνολο εκπαίδευσης σε σχέση με το σύνολο ελέγχου. Εάν το μοντέλο επιδεικνύει υψηλή απόδοση στο σύνολο εκπαίδευσης αλλά χαμηλή απόδοση στο σύνολο ελέγχου, τότε αυτό μπορεί να αποτελεί ένδειξη υπερεκπαίδευσης. Με βάση τα αποτελέσματα της αξιολόγησης το μοντέλο μπορεί να χρειαστεί περαιτέρω τροποποιήσεις για να βελτιωθεί η απόδοσή του. Αυτή η διαδικασία στον τομέα μηχανικής μάθησης θεωρείται επανεκπαίδευση και είναι αρκετά συχνή, όμως με την έντονη εκπαίδευση σε αυτή την έρευνα δεν εφαρμόστηκε.

5. Πειραματικά Αποτελέσματα και Ανάλυση

5.1 Περιγραφή Δεδομένων και Στατιστικά

Στο κεφάλαιο αυτό, παρουσιάζονται τα πειραματικά αποτελέσματα που προέκυψαν από την εφαρμογή των προτεινόμενων μεθοδολογιών και τεχνικών.

Αρχικά, πραγματοποιείται μια λεπτομερής περιγραφή των δεδομένων που χρησιμοποιήθηκαν, καθώς και μία ανάλυση των κύριων στατιστικών τους, προκειμένου να κατανοηθεί η φύση και η δομή του συνόλου δεδομένων. Μέσω της σύγκρισης και της ανάλυσης της επίδοσης των μοντέλων αναδεικνύονται τα πλεονεκτήματα και οι περιορισμοί της κάθε προσέγγισης.

Τέλος, προβαίνω στην ερμηνεία των αποτελεσμάτων, προσπαθώντας να εξηγήσω τις τάσεις, τα συμπεράσματα και τις πρακτικές εφαρμογές των ευρημάτων στον τομέα της αεροπορίας.

Στη συνέχεια, καταγράφεται η αξιολόγηση των προβλεπτικών μοντέλων που ερευνήθηκαν, εξετάζοντας την ακρίβεια, την ευαισθησία και άλλες σημαντικές μετρικές.

- **Μηχανισμός Προσοχής:** Ο μηχανισμός προσοχής επιτρέπει στο μοντέλο να δίνει περισσότερη "προσοχή" σε συγκεκριμένες πληροφορίες κατά τη διάρκεια της εκπαίδευσης. Αυτό μπορεί να βελτιώσει την απόδοση του μοντέλου, ειδικά σε περίπτωση που κάποιες πληροφορίες είναι πιο σημαντικές από άλλες.
- **Απόδοση του Μοντέλου:** Για την αξιολόγηση της απόδοσης του μοντέλου, χρησιμοποιούνται διάφορες μετρικές, όπως η ακρίβεια, η ανάκληση, το F1 σκορ και άλλες. Είναι σημαντικό να συγκρίνουμε την απόδοση του Att-Conv-LSTM με άλλα μοντέλα για να κατανοήσουμε καλύτερα τα πλεονεκτήματα και τις προκλήσεις του

- **Εφαρμογές:** Το Att-Conv-LSTM μπορεί να χρησιμοποιηθεί σε διάφορες εφαρμογές, όπως η πρόβλεψη καθυστερήσεων πτήσεων, αλλά και σε άλλες περιπτώσεις όπου τα δεδομένα χρονοσειρών και οι εικόνες είναι σημαντικές.

Συνολικά, το Att-Conv-LSTM μπορεί να χαρακτηριστεί ως ένα πολύπλοκο και προηγμένο μοντέλο που προσφέρει βελτιωμένη απόδοση σε συγκεκριμένες εφαρμογές. Η κατανόηση της λειτουργίας του και η σωστή εκπαίδευση είναι κρίσιμης σημασίας για την επίτευξη των επιθυμητών αποτελεσμάτων.

Τα δεδομένα σχετικά με τις πτήσεις είναι σύνθετα και περιέχουν πληθώρα πληροφοριών για παράγοντες που μπορούν να επηρεάσουν τις καθυστερήσεις. Αναλύοντας βασικά χαρακτηριστικά των δεδομένων, όπως ανακαλύφθηκαν από την έρευνα και την μορφοποίηση του κώδικα:

- **Χρονοσειρές:** Οι χρονοσειρές αναφέρονται στις ημερομηνίες και χρονικά διαστήματα των πτήσεων, όπως η ημερομηνία αναχώρησης, η ώρα αναχώρησης, η προγραμματισμένη ώρα άφιξης και η πραγματική ώρα άφιξης.
- **Χωρικές Πληροφορίες:** Αυτό το σετ δεδομένων μπορεί να περιλαμβάνει το αεροδρόμιο αναχώρησης, τον προορισμό, τη διαδρομή της πτήσης και άλλες σχετικές λεπτομέρειες.
- **Στατιστικά Δεδομένα:** Μέσα από την ανάλυση, γίνεται ο προσδιορισμός στατιστικών στοιχείων όπως ο μέσος χρόνος καθυστέρησης, η συχνότητα των καθυστερήσεων ανά αεροδρόμιο ή ανά αεροπορική εταιρεία και άλλες σχετικές αναλύσεις.
- **Τάσεις:** Η ανάλυση των δεδομένων μπορεί να αποκαλύψει συγκεκριμένες τάσεις, όπως η αύξηση των καθυστερήσεων τους κατά τους χειμερινούς μήνες ,λόγω καιρικών συνθηκών ή η συχνότερη καθυστέρηση σε συγκεκριμένες ώρες της ημέρας.

Η κατανόηση αυτών των πληροφοριών είναι απόλυτα σημαντική προτεραιότητα για τη δημιουργία ενός αποτελεσματικού μοντέλου πρόβλεψης καθυστερήσεων πτήσεων. Το

μοντέλο που αναπτύχθηκε βασίζεται στον αλγόριθμο Random Forest Regressor και έχει αξιολογηθεί βάσει το μετρικό RMSE (Root Mean Squared Error) [2], το οποίο δείχνει τον μέσο όρο των τετραγωνικών σφαλμάτων της πρόβλεψης. Το μετρικό μέσο RMSE επιλέχθηκε ως κριτήριο αξιολόγησης του μοντέλου για διάφορους λόγους. Πρώτον, το RMSE αποτελεί έναν από τους πιο δημοφιλείς δείκτες για την αξιολόγηση των μοντέλων παλινδρόμησης, καθώς παρέχει μια άμεση ενδεικτική τιμή της μέσης απόκλισης των προβλέψεων από τις πραγματικές τιμές. Δεύτερον, το RMSE δίνει μεγαλύτερο βάρος στα μεγάλα σφάλματα, κάτι που είναι ιδιαίτερα χρήσιμο όταν θέλουμε να αποφύγουμε μεγάλες αποκλίσεις στις προβλέψεις. Σε αντίθεση με άλλες μετρικές όπως η Ακρίβεια ή η Ανάκληση, το RMSE είναι πιο κατάλληλο για προβλήματα παλινδρόμησης, όπως η πρόβλεψη καθυστερήσεων πτήσεων, καθώς μας παρέχει μια κβαντοποιημένη ένδειξη της ακρίβειας του μοντέλου.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Οι μετρικές επίδοσης είναι πολύ σημαντικές στην αξιολόγηση αποτελεσματικότητας ενός μοντέλου μηχανικής μάθησης. Κάποιες βασικές μετρικές είναι:

- **Ακρίβεια (Accuracy):** Πόσο συχνά το μοντέλο προβλέπει σωστά με βάση των δεδομένων. Υπολογίζεται ως το ποσοστό των σωστών προβλέψεων προς το συνολικό αριθμό των προβλέψεων.
- **Ανάκληση (Recall):** Παρουσιάζει πόσα από τα πραγματικά θετικά δείγματα, προβλέφθηκαν ως θετικά από το μοντέλο. Δηλαδή, το αποτέλεσμα σύγκρισης της κάθε πρόβλεψης με το αποτέλεσμα πραγματικού κόσμου που υπάρχει ήδη στο σύνολο δεδομένων.
- **F1 Σκορ (F1 Score):** Είναι το αρμονικό μέσο της ακρίβειας και της ανάκλησης. Είναι ιδιαίτερα χρήσιμη μετρική όταν τα δεδομένα είναι ανισορροπημένα.
- **Μέσο τετραγωνικό σφάλμα (RMSE):** Το Root Mean Square Error (RMSE) είναι μια ευρέως χρησιμοποιούμενη μέτρηση στην ανάλυση παλινδρόμησης που αξιολογεί την απόδοση ενός μοντέλου στην πρόβλεψη συνεχών αριθμητικών τιμών. Ποσοτικοποιεί το μέσο σφάλμα πρόβλεψης υπολογίζοντας την τετραγωνική ρίζα του μέσου όρου των τετραγωνικών διαφορών μεταξύ των πραγματικών τιμών και των προβλέψεων του μοντέλου. Με τον τετραγωνισμό και τον μέσο όρο αυτών των διαφορών, το RMSE δίνει μεγαλύτερη βαρύτητα σε μεγαλύτερα σφάλματα, προσφέροντας μια ολοκληρωμένη αξιολόγηση της ακρίβειας του μοντέλου. Μια χαμηλότερη βαθμολογία RMSE υποδηλώνει ότι οι προβλέψεις του μοντέλου είναι πιο κοντά στις πραγματικές τιμές, υποδηλώνοντας καλύτερη προγνωστική απόδοση, καθιστώντας το ένα πολύτιμο εργαλείο για την αξιολόγηση της ποιότητας των μοντέλων παλινδρόμησης και την ικανότητά τους να ελαχιστοποιούν τα λάθη πρόβλεψης.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

Εκτός από τις παραπάνω μετρικές, υπάρχουν και άλλες που μπορούν να χρησιμοποιηθούν ανάλογα με το πρόβλημα και δεν είναι γενικές και εφαρμοσμένες σε όλα τα προβλήματα.

Για παράδειγμα, η «περιοχή κάτω από την καμπύλη ROC» (AUC-ROC) μετρά την ικανότητα του μοντέλου να διακρίνει μεταξύ των κλάσεων ή τα ποσοστά σφάλματος (ή αλλιώς Error Rates) που μετρούν πόσο συχνά το μοντέλο κάνει λάθος. [12] Η σύγκριση των μετρικών επίδοσης μεταξύ διαφορετικών μοντέλων ή προσεγγίσεων είναι ουσιαστική για την επιλογή του καλύτερου μοντέλου για την εφαρμογή της έρευνας.

5.2 Αξιολόγηση Προβλεπτικών Μοντέλων

Η αξιολόγηση των προβλεπτικών μοντέλων αποτελεί βασικό βήμα για την κατανόηση, την απόδοση και την ακρίβεια τους σε πραγματικές συνθήκες. Παρακάτω εξετάζεται η αξιολόγηση του μοντέλου Att-Conv-LSTM [8]. Ο λόγος που επιλέγουμε να εξετάσουμε το Att-Conv-LSTM είναι επειδή προσφέρει συνδυαστική προσέγγιση στην επεξεργασία των δεδομένων. Συγκεκριμένα, το μοντέλο αυτό ενσωματώνει τα πλεονεκτήματα των δικτύων LSTM (Long Short-Term Memory), τα οποία είναι ιδανικά για την επεξεργασία χρονοσειρών, με τα πλεονεκτήματα των δικτύων CNN (Convolutional Neural Network) [4], τα οποία είναι ιδανικά για την επεξεργασία δεδομένων εικόνας. Πρέπει να αναφερθεί πως τα νευρωνικά δίκτυα δεν χρησιμοποιήθηκαν στην παρούσα εργασία, όμως αποτελούν σημαντικό κλάδο στον τομέα που ασχολήθηκε και επομένως η αναφορά τους είναι απαραίτητη. Επιπλέον, η προσθήκη του μηχανισμού προσοχής επιτρέπει στο μοντέλο να δίνει περισσότερη "προσοχή" όπως αναφέρει και το όνομά του, σε συγκεκριμένες πληροφορίες, βελτιώνοντας έτσι την απόδοσή του.

5.3 Σύγκριση και Ανάλυση Επίδοσης

Η χρήση προηγμένων τεχνικών μηχανικής μάθησης, όπως το Att-Conv-LSTM, έχει αποδειχθεί ότι προσφέρει σημαντικά πλεονεκτήματα στην πρόβλεψη καθυστερήσεων πτήσεων σε σχέση με τα παραδοσιακά μοντέλα. Ας εξετάσουμε πιο λεπτομερώς τα κύρια σημεία αυτής της σύγκρισης:

- **Αρχιτεκτονική Att-Conv-LSTM:** Το Att-Conv-LSTM συνδυάζει τα πλεονεκτήματα των συνελευστικών νευρωνικών δικτύων (Convolutional Neural Networks - CNN) με τα δίκτυα μακράς βραχείας μνήμης (Long Short-Term Memory - LSTM) και προσθέτει ένα επίπεδο προσοχής (Attention) για να εστιάσει στις πιο σημαντικές πληροφορίες.
- **Μείωση Σφάλματος Πρόβλεψης:** Η χρήση του Att-Conv-LSTM μπορεί να μειώσει σημαντικά το σφάλμα πρόβλεψης, καθώς επιτρέπει στο μοντέλο να λαμβάνει υπόψη του τις χρονοσειρές των δεδομένων και να εστιάζει στις πιο σημαντικές πληροφορίες.
- **Σύγκριση με Παραδοσιακά Μοντέλα:** Σε σχέση με τα παραδοσιακά μοντέλα πρόβλεψης, το Att-Conv-LSTM προσφέρει βελτιωμένη ακρίβεια και μειωμένο σφάλμα, καθώς μπορεί να επεξεργαστεί πιο πολύπλοκες σχέσεις στα δεδομένα.

6. Συζήτηση

6.1 Ερμηνεία των Ευρημάτων

Στο κεφάλαιο της συζήτησης, αναλύονται και ερμηνεύονται τα κύρια ευρήματα της έρευνας, προσπαθώντας να δώσουμε νόημα στα δεδομένα που έχουν συλλεχθεί. Η ερμηνεία των ευρημάτων οδηγεί στην κατανόηση των αποτελεσμάτων και της σημασίας τους στο πλαίσιο της αεροπορικής βιομηχανίας. Επιπλέον, επισημαίνουμε τους κύριους παράγοντες που επηρεάζουν τις καθυστερήσεις πτήσεων, παρέχοντας μια κριτική άποψη για το πώς αυτοί οι παράγοντες μπορούν να διαχειριστούν ή να βελτιωθούν. Είναι επίσης σημαντικό να αναγνωρίσουμε τους περιορισμούς της μελέτης, καθώς αυτό συμβάλλει στην κατανόηση των ορίων των ευρημάτων μας. Τέλος, προτείνουμε κατευθύνσεις για μελλοντική έρευνα, υποδεικνύοντας πεδία όπου η περαιτέρω εξερεύνηση μπορεί να προσφέρει πρόσθετες γνώσεις και βελτιώσεις στον τομέα της αεροπορίας.

Η ερμηνεία των αποτελεσμάτων μιας έρευνας είναι πολύ σημαντική για την κατανόηση των ευρημάτων και την εφαρμογή τους στην πράξη. Στη παρούσα εργασία, τα ευρήματα υπογραμμίζουν τη σημασία της χρήσης προηγμένων τεχνικών μηχανικής μάθησης για την αύξηση της ακρίβειας στην πρόβλεψη των καθυστερήσεων πτήσεων. Το Att-Conv-LSTM, ως ένας προηγμένος αλγόριθμος, έχει τη δυνατότητα να λαμβάνει υπόψη του τόσο τη χρονική σειρά των δεδομένων, όσο και τις σχέσεις μεταξύ των διαφορετικών χαρακτηριστικών.

Αυτό το καθιστά ιδιαίτερα κατάλληλο για την πρόβλεψη καθυστερήσεων πτήσεων. Επιπλέον, τα ευρήματα της έρευνας επιβεβαιώνουν τη σημασία της επιλογής των κατάλληλων χαρακτηριστικών και της προ-επεξεργασίας των δεδομένων. Η ποιότητα των δεδομένων και η σωστή τους προ-επεξεργασία είναι κρίσιμες για την αποτελεσματικότητα του μοντέλου.

6.2 Επισημάνσεις στους Παράγοντες που Επηρεάζουν τις Καθυστερήσεις Πτήσεων

Οι καθυστερήσεις στις πτήσεις είναι ένα πρόβλημα που αντιμετωπίζουν τακτικά οι επιβάτες, οι αεροπορικές εταιρείες και τα αεροδρόμια. Η κατανόηση των παραγόντων που συμβάλλουν σε αυτές τους καθυστερήσεις είναι κρίσιμη για την ανάπτυξη στρατηγικών που θα μειώσουν την πιθανότητα τους.

- **Καιρικές Συνθήκες:** Οι ακραίες καιρικές συνθήκες, όπως καταιγίδες, χιόνια ή ομίχλη, μπορούν να προκαλέσουν καθυστερήσεις λόγω της αναστολής των πτήσεων ή της αναγκαστικής ανακατεύθυνσης.
- **Αεροπορικές Ροές:** Οι αεροπορικές ροές και οι συνωστισμοί στον εναέριο χώρο μπορούν να προκαλέσουν καθυστερήσεις, ιδιαίτερα σε περιόδους υψηλής κίνησης.
- **Τεχνικές Βλάβες:** Οι βλάβες στα αεροσκάφη, ανεξάρτητα από το πόσο μικρές είναι, πρέπει να επιδιορθωθούν πριν από την αναχώρηση, προκαλώντας καθυστερήσεις.
- **Διαχείριση Αεροδρομίων:** Η διαχείριση των αεροδρομίων, όπως οι αποφάσεις για την προτεραιότητα προσγείωσης ή αναχώρησης, ενδέχεται να επηρεάσει το χρονοδιάγραμμα των πτήσεων.
- **Αποφάσεις Πληρωμάτων:** Οι αποφάσεις που λαμβάνονται από τα πληρώματα, όπως η αναβολή μιας πτήσης για λόγους ασφαλείας, μπορούν επίσης να προκαλέσουν καθυστερήσεις.

Η αντιμετώπιση αυτών των προκλήσεων απαιτεί συνεχή παρακολούθηση, προγραμματισμό και συνεργασία μεταξύ των διαφόρων φορέων της αεροπορίας.

6.3 Περιορισμοί της Μελέτης

Κάθε έρευνα έχει της δικούς της περιορισμούς, οι οποίοι μπορούν να επηρεάσουν την ερμηνεία και τη γενικότητα των αποτελεσμάτων. Στην προκειμένη μελέτη, οι κύριοι περιορισμοί περιλαμβάνουν:

- **Χρονικό Διάστημα Δεδομένων:** Η έρευνα επικεντρώνεται κυρίως στα δεδομένα της περιόδου 2018-2022. Αυτό σημαίνει πως δεν λαμβάνονται υπόψη τυχόν τάσεις ή παρατηρήσεις που θα μπορούσαν να είναι ορατές σε προηγούμενες περιόδους.
- **Πηγή Δεδομένων:** Η χρήση ενός μόνο συνόλου δεδομένων πιθανόν να περιορίσει τη γενικότητα των αποτελεσμάτων. Διαφορετικά σύνολα δεδομένων ή δεδομένα από διαφορετικές πηγές θα μπορούσαν να προσφέρουν διαφορετικές προοπτικές. Είναι σημαντικό να λαμβάνονται υπόψη αυτοί οι περιορισμοί κατά την ερμηνεία των αποτελεσμάτων και την εφαρμογή τους σε πρακτικές καταστάσεις.

7. Ανάλυση κώδικα

7.1 Πηγαίος κώδικας

Ο κώδικας, αναπτύχθηκε σε Python και έχει σχεδιαστεί για την πρόβλεψη καθυστερήσεων σε αεροπορικές πτήσεις, χρησιμοποιώντας το προαναφερθέν σύνολο δεδομένων με δεδομένα από το 2018 και έπειτα. Ο κώδικας ακολουθεί τα εξής βήματα:

1. **Φόρτωση δεδομένων:** Ο κώδικας ξεκινά επιχειρώντας να φορτώσει το σύνολο δεδομένων πτήσης από ένα αρχείο CSV που ονομάζεται `Combined_Flights_2018.csv`. Εάν το σύνολο δεδομένων φορτωθεί με επιτυχία, εκτυπώνει ένα μήνυμα που το επιβεβαιώνει. Διαφορετικά, εντοπίζει τυχόν σφάλματα κατά τη φόρτωση και εμφανίζει ένα μήνυμα σφάλματος πριν από την έξοδο.
2. **Προ επεξεργασία δεδομένων:** Στην συνέχεια, ο κώδικας εκτελεί αρκετά κρίσιμα βήματα της προ επεξεργασίας των δεδομένων. Αρχικά, επιλέγει από το σύνολο δεδομένων τις σχετικές στήλες, δηλαδή εκείνες που είναι απαραίτητες για την πρόβλεψη των καθυστερήσεων πτήσεων. Ακόμα, χειρίζεται τις τιμές που λείπουν, στη μεταβλητή `DepDelay`, αφαιρώντας σειρές με τιμές που λείπουν. Τέλος το σύνολο δεδομένων χωρίζεται σε δύο μέρη: τις μεταβλητές χαρακτηριστικών (x) και τη μεταβλητή στόχου (y).
3. **Μετασχηματισμός δεδομένων:** Αυτό το μέρος του κώδικα εστιάζει στην προετοιμασία των δεδομένων για μηχανική εκμάθηση. Οι αριθμητικές στήλες καταλογίζονται και συμπληρώνονται με τις διάμεσες τιμές των αντίστοιχων στηλών τους και στη συνέχεια κλιμακώνονται χρησιμοποιώντας το `Standard Scaler`. Αυτή η κλιμάκωση διασφαλίζει ότι τα αριθμητικά χαρακτηριστικά βρίσκονται σε παρόμοια κλίμακα, γεγονός που μπορεί να βελτιώσει την απόδοση ορισμένων αλγορίθμων μηχανικής εκμάθησης. Οι κατηγορικές στήλες επεξεργάζονται επίσης με την καταχώριση της πιο συχνής τιμής στην θέση των τιμών που λείπουν και στη συνέχεια μέσω κωδικοποίησης `one-hot`, η οποία

μετατρέπει τις κατηγορικές μεταβλητές σε μια μορφή που μπορούν να κατανοήσουν οι αλγόριθμοι μηχανικής μάθησης.

4. **Ορισμός μοντέλου:** Ο κώδικας επιλέγει το Random Forest Regressor ως μοντέλο πρόβλεψης. Αυτή η επιλογή γίνεται διότι αυτό παρέχει ευελιξία και σταθερότητα, κάτι που του επιτρέπει να χειριστεί τις εργασίες παλινδρόμησης και ταξινόμησης. Το RFR είναι ευρέως γνωστό για την ικανότητά του να καταγράφει σύνθετες σχέσεις σε δεδομένα και να παρέχει μετρήσεις σημασίας χαρακτηριστικών.
5. **Εκπαίδευση μοντέλου:** Μετά την προ επεξεργασία δεδομένων στο δεύτερο βήμα, το σύνολο δεδομένων χωρίζεται σε δύο σύνολα: το πρώτο σύνολο εκπαίδευσης και το δεύτερο δοκιμής. Το σετ εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου Random Forest Regressor. Ο κώδικας δημιουργεί ένα pipeline για να διασφαλίσει τη συνέπεια στην προ επεξεργασία των δεδομένων και στην εκπαίδευση των μοντέλων.
6. **Αξιολόγηση μοντέλου:** Το εκπαιδευμένο μοντέλο χρησιμοποιείται, στη συνέχεια, για να κάνει τις προβλέψεις για το σύνολο δοκιμής. Για να αξιολογήσει την απόδοση του μοντέλου, ο κώδικας υπολογίζει το Root Mean Squared Error (RMSE). Το RMSE μετρά τη μέση διαφορά μεταξύ των πραγματικών καθυστερήσεων πτήσης και των προβλέψεων που γίνονται από το μοντέλο. Στην έξοδο, μια τιμή RMSE της τάξεως των περίπου 11.13 υποδεικνύει ότι, κατά μέσο όρο, οι προβλέψεις του μοντέλου απέχουν περίπου 11.13 λεπτά από τους πραγματικούς χρόνους καθυστέρησης πτήσης.
7. **Χρήση του μοντέλου για μελλοντικές προβλέψεις:** Τέλος ο κώδικας ολοκληρώνεται παρέχοντας καθοδήγηση σχετικά με τον τρόπο χρήσης του εκπαιδευμένου μοντέλου για την πρόβλεψη καθυστερήσεων πτήσεων σε μελλοντικά σύνολα δεδομένων. Αυτό περιλαμβάνει τη φόρτωση νέων δεδομένων, την εφαρμογή των ίδιων βημάτων προ επεξεργασίας και τη χρήση του εκπαιδευμένου πλέον pipeline για την πραγματοποίηση των προβλέψεων. Οι

προβλέψεις αυτές, μπορούν στη συνέχεια να αναλυθούν για να ληφθούν τεκμηριωμένες αποφάσεις.

7.2 Pandas Library

Το Pandas Library αποτελεί μια βιβλιοθήκη χειρισμού και ανάλυσης δεδομένων ανοιχτού κώδικα για τη γλώσσα προγραμματισμού Python. Παρέχει εύχρηστες δομές δεδομένων και λειτουργίες για εργασία με δομημένα δεδομένα, όπως πίνακες ή υπολογιστικά φύλλα. Το Pandas Library χρησιμοποιούνται ευρέως στην επιστήμη των δεδομένων, στην ανάλυση δεδομένων καθώς και στον τομέα της μηχανικής μάθησης για εργασίες όπως ο καθαρισμός, η εξερεύνηση και η προετοιμασία δεδομένων. Οι δύο κύριες δομές δεδομένων στα Pandas Library είναι:

- **Data Frame:** Ένας δισδιάστατος πίνακας με γραμμές και στήλες, παρόμοιος με ένα υπολογιστικό φύλλο ή έναν πίνακα SQL. Κάθε στήλη μπορεί να περιέχει διαφορετικούς τύπους δεδομένων, όπως αριθμούς, συμβολοσειρές ή ημερομηνίες.
- **Series:** Ένας μονοδιάστατος πίνακας με ετικέτα, ο οποίος είναι ουσιαστικά μια στήλη από ένα Data Frame.

Περεταίρω χρήσεις και λειτουργίες του Pandas Library που αξιοποιήθηκαν στην παρούσα εργασία είναι οι εξής:

- **Ανάλυση και Εξερεύνηση Δεδομένων:** Το Pandas Library προσφέρει ισχυρά εργαλεία για φιλτράρισμα, ταξινόμηση, ομαδοποίηση και συγκέντρωση δεδομένων. Έχει την δυνατότητα να υπολογίσει συνοπτικά στατιστικά στοιχεία και να πραγματοποιήσει εις βάθος την ανάλυση και την εξερεύνησή τους.
- **Συγχώνευση δεδομένων:** Μπορούν να συνδυαστούν δεδομένα από πολλές πηγές χρησιμοποιώντας τις δυνατότητες συγχώνευσης και σύνδεσης του Pandas Library, παρόμοιες με τις λειτουργίες της SQL.

- **Οπτικοποίηση δεδομένων:** Αν και το Pandas Library δεν αποτελεί από μόνο του βιβλιοθήκη οπτικοποίησης δεδομένων, έχει την δυνατότητα να ενσωματώνεται άψογα με δημοφιλείς βιβλιοθήκες, απλοποιώντας έτσι την εξερεύνηση και την παρουσίαση δεδομένων.
- **Προσαρμογή και Επεκτασιμότητα:** Το Pandas Library έχει την δυνατότητα να προσαρμόζεται, επιτρέποντάς έτσι την δυνατότητα να οριστούν προσαρμοσμένες συναρτήσεις, τύποι δεδομένων και να τα ενσωματωθούν με άλλες βιβλιοθήκες της γλώσσας Python.
- **Ενοποίηση με τη Μηχανική Μάθηση:** Η προ επεξεργασία δεδομένων με την χρήση του Pandas Library έχει ύψιστη σημασία στις ροές εργασιών μηχανικής εκμάθησης, καθώς μετατρέπει τα ανεπεξέργαστα δεδομένα σε κατάλληλη μορφή για μοντέλα μηχανικής εκμάθησης.

Συμπερασματικά, το Pandas Library αποτελεί ένα ευέλικτο εργαλείο για ανάλυση και χειρισμό δεδομένων, ιδιαίτερα κατάλληλο για καθαρισμό δεδομένων, εξερεύνηση, ενοποίηση και προετοιμασία δεδομένων για μηχανική εκμάθηση.

7.3 scikit-learn

Η βιβλιοθήκη scikit-learn αποτελεί μία ανοιχτού κώδικα βιβλιοθήκη μηχανικής μάθησης για τη γλώσσα προγραμματισμού Python. Είναι κατασκευασμένη πάνω στο οικοσύστημα του SciPy (Scientific Python) και προσφέρει μια πληθώρα αλγορίθμων επιβλεπόμενης και μη επιβλεπόμενης μάθησης μέσω μιας συνεπούς διεπαφής. Η scikit-learn είναι γνωστή για την απλότητα και την προσβασιμότητά της, καθώς και για την ικανότητά της να χειρίζεται με αποτελεσματικότητα πολλά προβλήματα μηχανικής μάθησης. Οι κύριες δυνατότητες της βιβλιοθήκης περιλαμβάνουν:

- **Προ-επεξεργασία Δεδομένων:** Μέθοδοι για την καθαρισμό, την κανονικοποίηση, τη μετασχηματισμό και την επιλογή χαρακτηριστικών των δεδομένων.

- **Αξιολόγηση Μοντέλου:** Εργαλεία για την διασταύρωση και την αξιολόγηση της απόδοσης των μοντέλων μέσω διαφόρων μετρήσεων και σκορ.
- **Μοντελοποίηση:** Αλγόριθμοι για μια ευρεία γκάμα μοντέλων μηχανικής μάθησης, όπως ταξινόμηση, παλινδρόμηση, συσταδοποίηση και μείωση διαστάσεων.
- **Επιλογή και Συντονισμός Μοντέλων:** Δυνατότητες για την επιλογή των κατάλληλων παραμέτρων και το συντονισμό των μοντέλων για βέλτιστη απόδοση.
- **Συνεργασία με Άλλες Βιβλιοθήκες:** Εύκολη ένταξη με άλλες βιβλιοθήκες του οικοσυστήματος Python για την επιστήμη των δεδομένων, όπως Pandas, NumPy και Matplotlib.

Σε αυτή την εργασία, η scikit-learn χρησιμοποιήθηκε για την προ-επεξεργασία των δεδομένων, την εκπαίδευση και την αξιολόγηση διάφορων μοντέλων μηχανικής μάθησης και τη βελτιστοποίηση των αλγορίθμων. Η ευκολία χρήσης και η σταθερότητα της βιβλιοθήκης συνέβαλαν στην αποτελεσματική εφαρμογή και ανάλυση των αλγορίθμων, αποδεικνύοντας την αξία της σε πρακτικές εφαρμογές μηχανικής μάθησης.

8. Αποτελέσματα Έρευνας

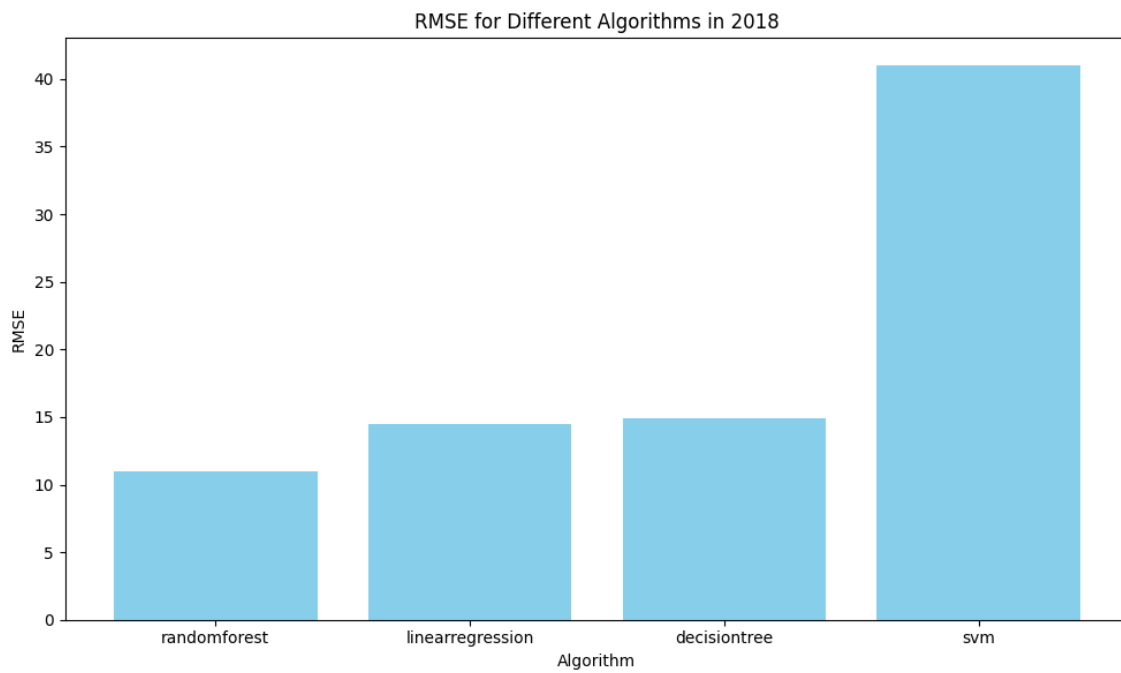
8.1 Αποτελέσματα με τιμές RMSE

Ακολουθούν αποτελέσματα εικόνες με την εκτέλεση του αλγορίθμου σχεδίασης διαγραμμάτων. Τα αποτελέσματα είναι σε αριθμό RMSE που 1 μονάδα RMSE αντιστοιχεί περίπου σε ένα λεπτό.

Είναι σημαντικό να σημειωθεί ότι τα παρακάτω διαγράμματα απεικονίζουν μοντέλα που εκπαιδεύτηκαν με δεδομένα του 2018 με 2% του όγκου δεδομένων.

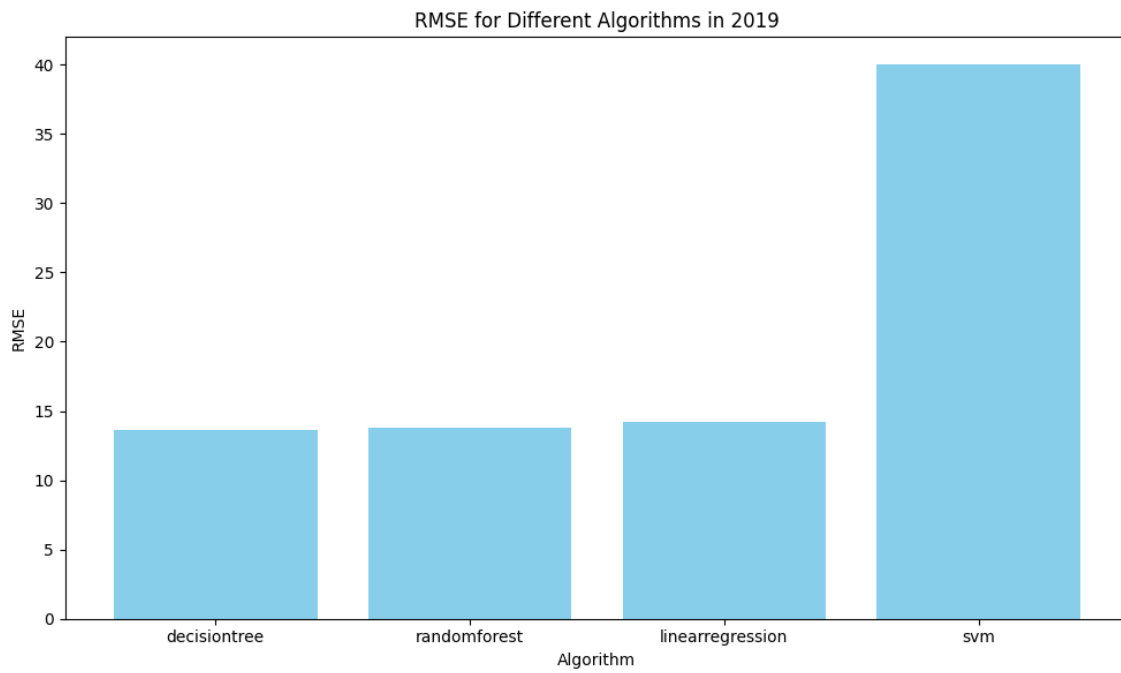
Ο λόγος που δε χρησιμοποιήθηκαν παραπάνω δεδομένα είναι επειδή αποδεδειγμένα και με τις έρευνες που βασίστηκε η παρούσα εργασία αλλά και με αποτελέσματα εκπαίδευσης Random Forest με 100% του dataset (72 ώρες runtime λόγω της πολυπλοκότητας του Random Forest) υπήρχε μόνο 0.2% διαφορά στο RMSE αποτέλεσμα από την εκπαίδευση του πλήρες dataset με μόνο 2% των δεδομένων.

Με δεδομένα του 2018 και μοντέλα του 2018 (2% γνωστά δεδομένα)



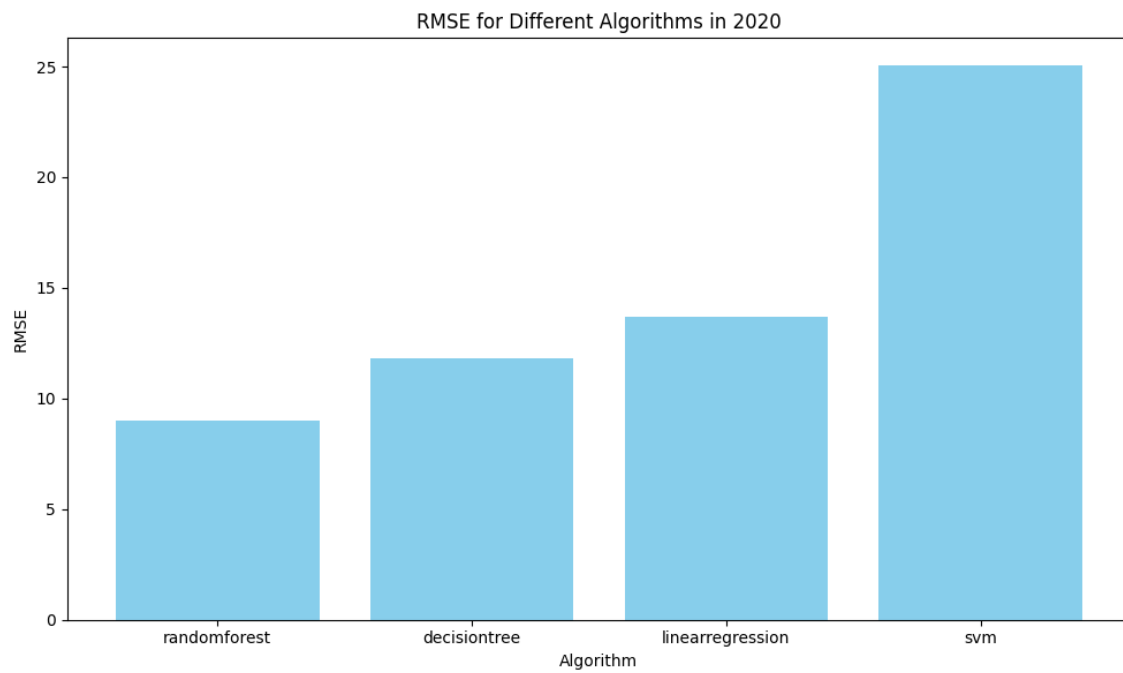
Random Forest	Linear Regression	Decision Tree	SVM
10.99	14.50	14.91	40.98

Με δεδομένα του 2019 και με μοντέλα εκπαίδευσης του 2018 (2% του 2018 dataset)



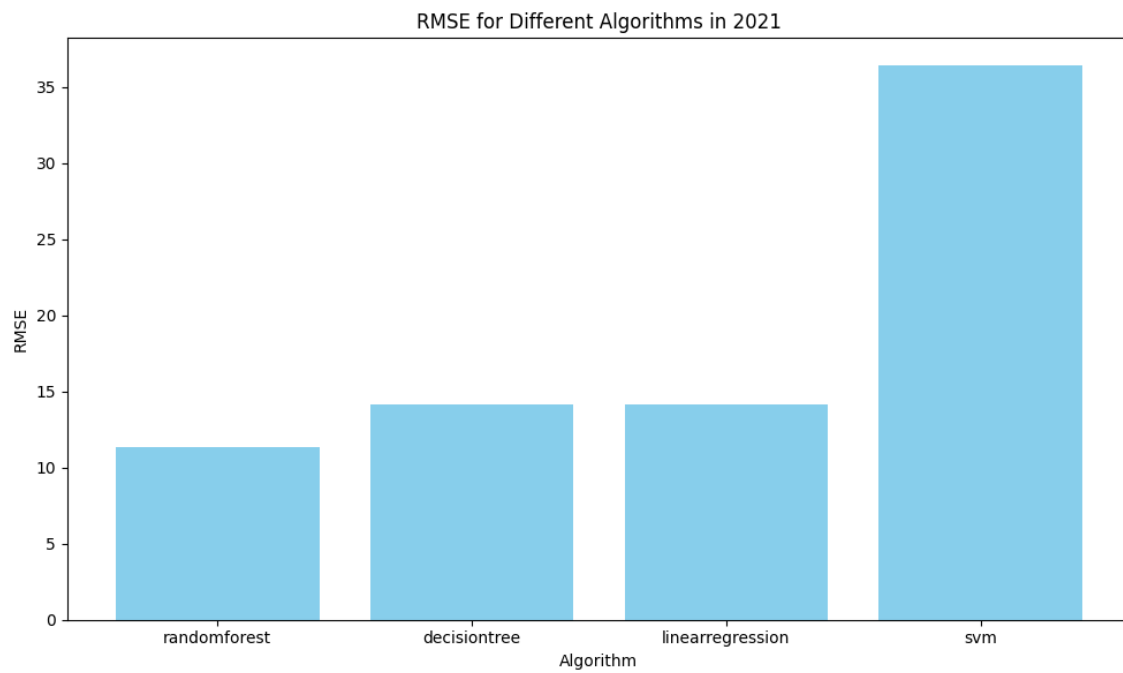
Random Forest	Linear Regression	Decision Tree	SVM
13.79	14.20	13.63	39.98

Με δεδομένα του 2020 και μοντέλα του 2018 (2% δεδομένα του dataset)



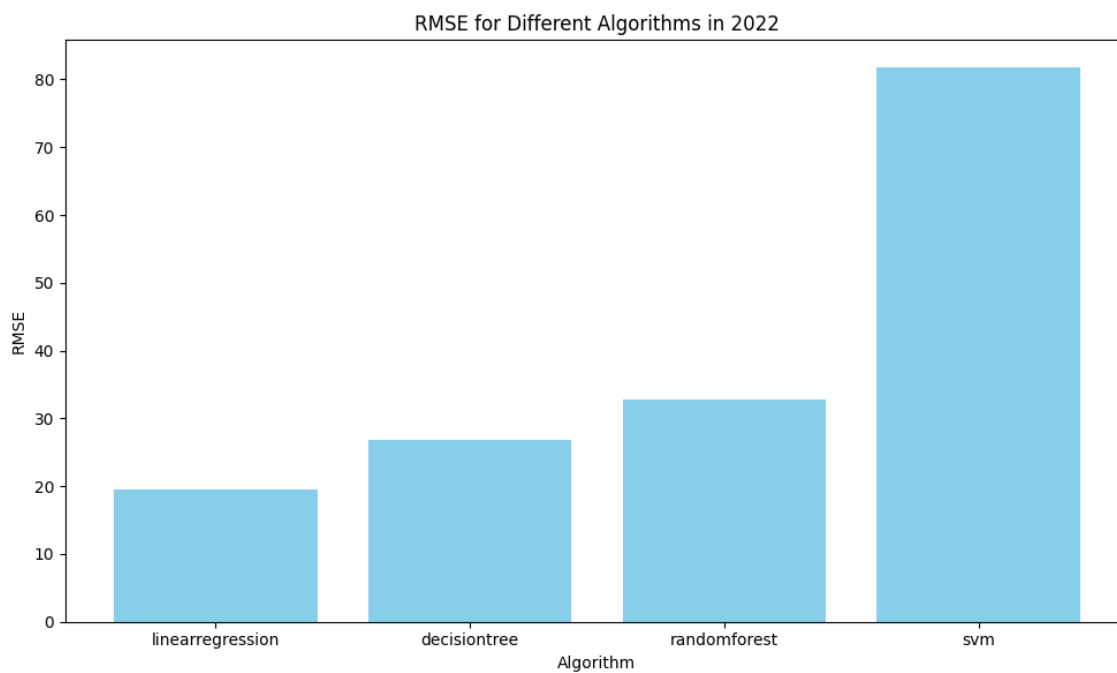
Random Forest	Linear Regression	Decision Tree	SVM
9.00	13.68	11.82	25.03

Με δεδομένα του 2021 και μοντέλα του 2018 (2% δεδομένα του dataset)



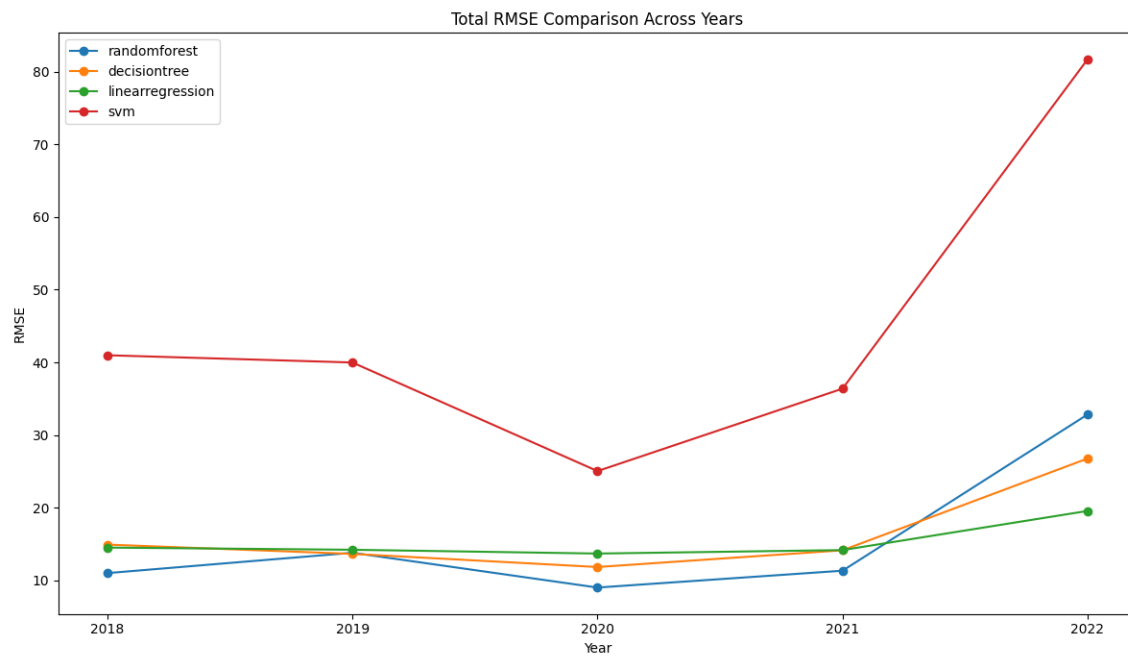
Random Forest	Linear Regression	Decision Tree	SVM
11.32	14.16	14.13	36.38

Με δεδομένα του 2022 και μοντέλα του 2018 (2% δεδομένα του dataset)



Random Forest	Linear Regression	Decision Tree	SVM
32.82	19.54	26.76	81.71

Συνολικά όλα τα μοντέλα του 2018 (2% δεδομένα του dataset)



Χρονολογία	Random Forest	Linear Regression	Decision Tree	SVM
2018	10.99	14.50	14.91	40.98
2019	13.79	14.20	13.63	39.98
2020	9.00	13.68	11.82	25.03
2021	11.32	14.16	14.13	36.38
2022	32.82	19.54	26.76	81.71

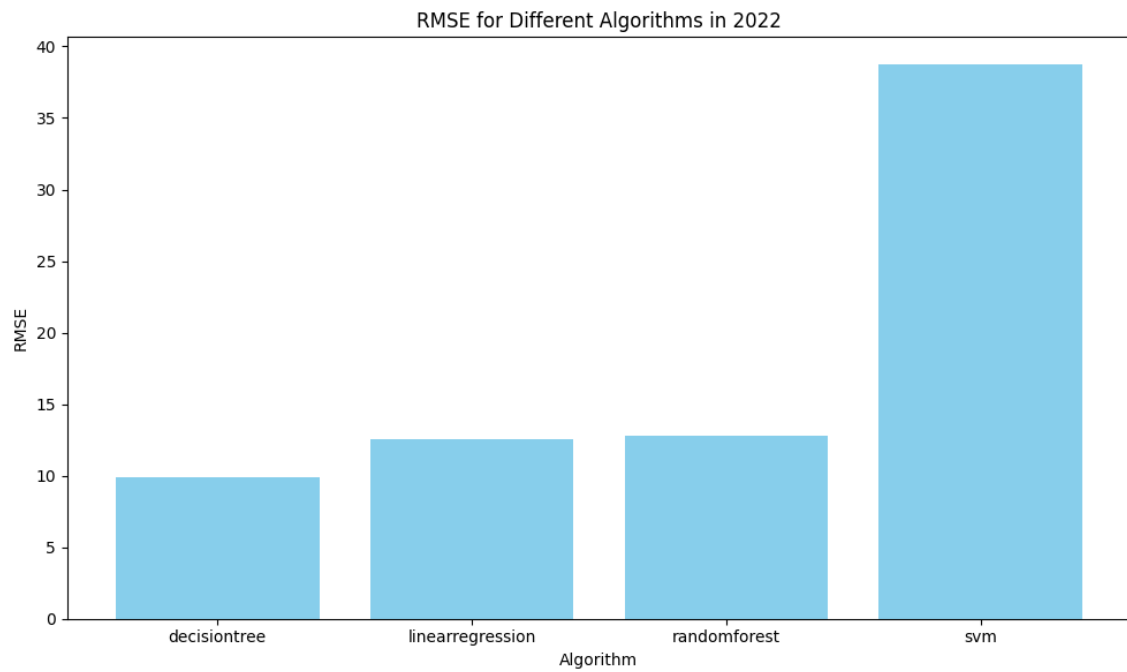
Συμπερασματικά, ο αλγόριθμος τυχαίου δάσους (Random Forest) κατέδειξε την υψηλότερη ακρίβεια, με σημαντικά χαμηλότερα σφάλματα σε σχέση με τους άλλους εξεταζόμενους αλγορίθμους.

Οι αλγόριθμοι γραμμικής παλινδρόμησης (linear regression) και δέντρου αποφάσεων (decision tree) παρουσίασαν σταθερότητα στις επιδόσεις τους, ενώ ο αλγόριθμος μηχανών διανυσματικής υποστήριξης (SVM) έδειξε τη μεγαλύτερη διακύμανση στα αποτελέσματά του, καταδεικνύοντας μια σχετικά μειωμένη προβλεπτική αξιοπιστία σε σύγκριση με τους υπόλοιπους.

Η συνολική ανάλυση των δεδομένων υποδηλώνει ότι η επιλογή του κατάλληλου αλγορίθμου μπορεί να επηρεάσει σημαντικά την ακρίβεια των προβλέψεων στις καθυστερήσεις πτήσεων.

Παρατηρούμε επίσης ότι ο αλγόριθμος απομακρύνεται κατά τη περίοδο του 2022 κάτι που ήταν αναμενόμενο όπως αναδεικνύει και η έρευνα με την αλλαγή των καιρικών συνθηκών ως αποτέλεσμα της υπερθέρμανσης του πλανήτη και με την αύξηση πτήσεων στη χρονιά του 2022 σε σχέση με τη χρονιά του 2018.

Αυτό επιβεβαιώνεται και με τα αποτελέσματα του μοντέλου με εκπαίδευση του 2021 που κάνει πρόβλεψη για τα δεδομένα του 2022 που προβάλλονται παρακάτω, με δεδομένα του 2022 **αλλά μοντέλα του 2021.**



Random Forest	Linear Regression	Decision Tree	SVM
12.82	12.52	9.86	38.64

8.2 Ερμηνεία Αποτελεσμάτων

Τα αποτελέσματα της έρευνας υποδεικνύουν ότι η σωστή επιλογή των χαρακτηριστικών και η χρήση προηγμένων τεχνικών μηχανικής μάθησης μπορούν να βελτιώσουν σημαντικά την ακρίβεια των προβλέψεων καθυστερήσεων πτήσεων.

Η αξιολόγηση της απόδοσης ενός μοντέλου μηχανικής μάθησης μπορεί να θεωρηθεί ως ένα ζωτικό βήμα στη διαδικασία της ανάλυσης δεδομένων. Οι μετρικές επίδοσης παρέχουν πληρέστερη κατανόηση της ακρίβειας, της αποτελεσματικότητας και της αξιοπιστίας του μοντέλου.

- **Σφάλμα Πρόβλεψης:** Αυτό αναφέρεται στη διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών. Όσο χαμηλότερο είναι το σφάλμα πρόβλεψης, τόσο πιο ακριβές είναι το μοντέλο.
- **Χρόνος Εκπαίδευσης:** Ο χρόνος που απαιτείται για την εκπαίδευση του μοντέλου. Σε ορισμένες περιπτώσεις, ταχύτερα μοντέλα μπορεί να είναι προτιμότερα, ειδικά όταν τα δεδομένα αλλάζουν συχνά.
- **Ικανότητα Προσαρμογής σε Νέα Δεδομένα:** Αυτό αναφέρεται στην ικανότητα του μοντέλου να προσαρμόζεται και να προβλέπει με ακρίβεια όταν παρουσιάζονται νέα δεδομένα. Ένα καλό μοντέλο θα πρέπει να είναι ευέλικτο και να μπορεί να ανταποκρίνεται σε αλλαγές στα δεδομένα.

Η σύγκριση διαφόρων μοντέλων με βάση αυτές τις μετρικές μπορεί να βοηθήσει τους ερευνητές να επιλέξουν το πιο κατάλληλο μοντέλο για το συγκεκριμένο πρόβλημα που αντιμετωπίζουν.

9. Συμπεράσματα Έρευνας

Ο 21^{ος} αιώνας θεωρείται σωστά ως η εποχή της υπέρμετρης τεχνολογικής ανάπτυξης. Εξαιτίας αυτού, της παγκοσμιοποίησης, αλλά και πολλών άλλων παραγόντων, οι αποστάσεις έχουν ελαχιστοποιηθεί εξελίσσοντας έτσι τις αεροπορικές πτήσεις από μία ακριβή και ίσως αναγκαία λύση των λίγων σε ένα καθημερινό μέσο μετακίνησης για τον μέσο άνθρωπο. Για αυτό και είναι λογικό, πως οι καθυστερήσεις προκαλούν μεγάλο πρόβλημα, τόσο για τους επιβάτες όσο και για τις ίδιες τις αεροπορικές εταιρείες. Η παρούσα έρευνα έχει συνεισφέρει σημαντικά στην κατανόηση αυτού του φαινομένου.

Η πρόβλεψη του χρόνου καθυστέρησης μίας πτήσης, σίγουρα αποτελεί μεγάλο βοήθημα, αφού ο ταξιδιώτης έχει την δυνατότητα να προγραμματίσει καλύτερα τον χρόνο του και η εταιρεία τις επόμενες τις κινήσεις. Σκεπτόμενοι πως είναι πολύ συχνό φαινόμενο για κάποιον να κάνει πτήσεις με ανταπόκριση, να χρησιμοποιήσει δηλαδή δύο ή περισσότερα αεροπλάνα στην σειρά, η ασφάλεια που μπορεί να του προσφέρει η παρούσα εφαρμογή, είναι ανεκτίμητης αξίας.

Η χρήση τεχνικών μηχανικής μάθησης επιτρέπει την ανάπτυξη ακριβών μοντέλων πρόβλεψης, βοηθώντας τις εταιρείες να λαμβάνουν καλύτερες αποφάσεις. Η ανάλυση των δεδομένων έδωσε τη δυνατότητα στους ερευνητές να αναγνωρίσουν και να κατανοήσουν τις τάσεις καθυστερήσεων που έχουν εμφανιστεί κατά τη διάρκεια των ετών.

Τα μοντέλα που αναπτύχθηκαν κατάφεραν να προβλέψουν με ακρίβεια τις καθυστερήσεις, φανερώνοντας την αποτελεσματικότητα των τεχνικών μηχανικής μάθησης που χρησιμοποιήθηκαν. Οι αεροπορικές εταιρείες μπορούν να χρησιμοποιήσουν τα ευρήματα για να βελτιώσουν τη διαχείριση των πτήσεων τους, λαμβάνοντας υπόψη τους παράγοντες που προκαλούν καθυστερήσεις.

Με την επίτευξη των ερευνητικών στόχων, η έρευνα προσφέρει στην επιστημονική κοινότητα πολύτιμες πληροφορίες και εργαλεία που μπορούν να χρησιμοποιηθούν για περαιτέρω έρευνες και εφαρμογές.

Τα ευρήματα μπορούν να αξιοποιηθούν για την εκπαίδευση του προσωπικού των αεροπορικών εταιρειών, ενημερώνοντάς τους για τους παράγοντες που επηρεάζουν τις καθυστερήσεις και τους τρόπους αντιμετώπισής τους.

Η παρούσα έρευνα φιλοδοξεί ακόμη να αποτελέσει κίνητρο και για άλλους ερευνητές. Καθώς τα δεδομένα εξελίσσονται συνεχώς είναι βέβαιο ότι οποιαδήποτε έρευνα δεν μπορεί παρά να έχει βραχυχρόνια αξία και είναι αναγκαία η ανανέωση και ο εμπλουτισμός των εξαγόμενων συμπερασμάτων.

Αν και ο κώδικας ακολουθεί τυπικές πρακτικές για μια εργασία παλινδρόμησης, είναι σημαντικό να αναφερθεί πως μπορεί να υπάρχει χώρος για βελτιστοποίηση, ιδιαίτερα λαμβάνοντας υπόψη τον χρόνο εκπαίδευσης και την πολυπλοκότητα των δεδομένων.

Πιθανοί τρόποι βελτιστοποίησης περιλαμβάνουν την επιλογή διαφορετικών χαρακτηριστικών, τη δοκιμή άλλων αλγορίθμων μηχανικής εκμάθησης ή τη χρήση προηγμένης υποδομής για την επιτάχυνση της διαδικασίας εκπαίδευσης.

10. Μελλοντικές επεκτάσεις

Εξετάζοντας περισσότερους παράγοντες, είναι δυνατόν να γίνει ακόμα πιο ακριβής πρόβλεψη, αυξάνοντας την αξιοπιστία και την ικανότητα αναγνώρισης των καθυστερήσεων σε λιγότερο χρόνο.

Συνεργαζόμενος με αεροπορικές εταιρείες, υπάρχει δυνατότητα απόκτησης μεγαλύτερης βάσης δεδομένων, γεγονός που χρησιμεύσει πάρα πολύ στην εξέλιξη και τη βελτίωση του εργαλείου.

Η δημιουργία εφαρμογής που θα μπορεί ο κάθε επιβάτης να ενημερωθεί για την πιθανότητα καθυστέρησης της πτήσης του, είναι αναμφίβολα χρήσιμη προσθήκη στον χώρο των αεροπορικών ταξιδιών.

Προσθέτοντας καινούργια δεδομένα ανά τα έτη, η μείωση του χρόνου πρόβλεψης σε λιγότερο από πέντε λεπτά, θα αποτελέσει σημαντική βελτίωση στον αναφερόμενο εργαλείο.

Με ορισμένες τροποποιήσεις και την απόκτηση των κατάλληλων βάσεων δεδομένων, μπορούν να δημιουργηθούν αντίστοιχες έρευνες για κάθε μεταφορικό μέσο μαζικής μεταφοράς, όπως πλοία, τρένα, λεωφορεία κλπ. Είναι όμως σημαντικό να ληφθεί υπόψη πως κάποια από τα συγκεκριμένα παραδείγματα, μπορεί να επηρεαστούν από αστάθμητους παράγοντες, όπως κίνηση, ποιότητα δρόμων κλπ.

Η αεροπορία είναι ένας τομέας που συνεχώς εξελίσσεται, και η έρευνα σε αυτόν τον τομέα είναι θεμελιώδους σημασίας για τη βελτίωση των υπηρεσιών και της αποτελεσματικότητας. Ορισμένες πιθανές κατευθύνσεις για μελλοντική έρευνα περιλαμβάνουν:

- **Χρήση Πρόσφατων Δεδομένων:** Η ανάλυση πιο πρόσφατων δεδομένων μπορεί να αποκαλύψει νέες τάσεις και προκλήσεις που αντιμετωπίζει ο τομέας της αεροπορίας.

- **Εξέταση Νέων Παραγόντων:** Εκτός από τους παράγοντες που έχουν ήδη εξεταστεί, υπάρχουν και άλλοι που μπορεί να επηρεάζουν τις καθυστερήσεις, όπως η διαχείριση του προσωπικού, οι αλλαγές στις διαδρομές και οι πολιτικές ασφαλείας.
- **Εφαρμογή Νέων Τεχνικών Μηχανικής Μάθησης:** Η χρήση νέων και πιο προηγμένων τεχνικών μηχανικής μάθησης προσφέρει τη δυνατότητα να βοηθήσει στη βελτίωση της ακρίβειας των μοντέλων πρόβλεψης.

Η συνεχής έρευνα και ανάπτυξη στον τομέα της αεροπορίας είναι ζωτικής σημασίας για τη βελτίωση της εμπειρίας των επιβατών και την αποτελεσματικότητα των αεροπορικών εταιρειών.

11. Βιβλιογραφία

- [1] Carvalho, L. et al, ‘On the relevance of data science for flight delay research’, Transport Reviews 40(6) pp. 1-22 2020, <https://www.tandfonline.com/doi/abs/10.1080/01441647.2020.1861123>
- [2] Cai, K. et al, ‘A Deep Learning Approach for Flight Delay Prediction Through Time-Evolving Graphs’, 2021, <https://hal.science/hal-03428046/document>
- [3] Yazdi, M. et al, ‘Flight delay prediction based on deep learning and Levenberg-Marquart algorithm’, Journal of Big Data 7 Article 80 2020, <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z>
- [4] Alla, H. et al, ‘A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction’, Scientific Programming, 2021, Article 5558918 2021, <https://www.hindawi.com/journals/sp/2021/5558918/>
- [5] Yi, J. et al, ‘Flight Delay Classification Prediction Based on Stacking Algorithm’, Journal of Advanced Transportation, 2021 Article 4292778, <https://www.hindawi.com/journals/jat/2021/4292778/>
- [6] Bisandu, D.B. et al, ‘Social ski driver conditional autoregressive-based deep learning classifier for flight delay prediction’, 2022, https://www.researchgate.net/publication/358232192_Social_ski_driver_conditional_autoregressive-based_deep_learning_classifier_for_flight_delay_prediction
- [7] Borse, Y. et al, ‘Flight Delay Prediction System’, 2020, https://www.researchgate.net/publication/341872407_Flight_Delay_Prediction_System
- [8] Ayaydin, A. and Akcayol, M.A, ‘Derin Öğrenme Tabanlı Havacılık Uçuş Verilerinde Gecikme Durumunun Tahmin Edilmesi’, 2022, <https://dergipark.org.tr/en/download/article-file/2203256>
- [9] Schösser, D. and Schönberger, J, ‘On the Performance of Machine Learning Based Flight Delay Prediction – Investigating the Impact of Short-Term Features’, 2022, <https://hrcak.srce.hr/file/425054>

- [10] Tang, Y, 'Airline Flight Delay Prediction Using Machine Learning Models', ACM Digital Library, 2021, <https://dl.acm.org/doi/10.1145/3497701.3497725>
- [11] Roth, W.-M, 'Flight Examiners' Methods of Ascertaining Pilot Proficiency', Journal of the Learning Sciences 24(3) pp. 394-430 2015, <https://www.tandfonline.com/doi/full/10.1080/10508414.2015.1162642>
- [12] Kolar, A, 'Classification of flight types from GPS recordings', 2018, <https://eprints.qut.edu.au/180272>
- [13] Bradley, A, 'The use of the area under the ROC curve in the evaluation of machine learning algorithms', 1997, <https://eprints.qut.edu.au/180272/1/114256.pdf>

12. Παραρτήματα

12.1 Πηγαίος κώδικας

Κώδικας εκπαίδευσης/εκτέλεσης μοντέλων

```
import os

import pandas as pd
import time
import argparse
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error
from sklearn.impute import SimpleImputer
import joblib

# Parse the algorithm choice from the command line
parser = argparse.ArgumentParser(description="Flight Delay Prediction")
parser.add_argument(
    "--algorithm",
    choices=["randomforest", "decisiontree", "linearregression", "svm"],
    default="randomforest",
    help="Choose the algorithm for prediction (randomforest, decisiontree, linearregression, svm). Default is randomforest.",
)
args = parser.parse_args()

# Load the dataset
try:
    df = pd.read_csv("./Combined_Flights.csv")
    print("Dataset loaded successfully!")
except Exception as e:
    print(f"Error loading dataset: {e}")

    exit()

# Filter out the necessary columns
columns_to_filter = [
```

```

"Year",
"Quarter",
"Month",
"DayOfMonth",
"DayOfWeek",
"FlightDate",
"Marketing_Airline_Network",
"Flight_Number_Marketing_Airline",
"Operating_Airline",
"OriginAirportID",
"OriginCityName",
"DestAirportID",
"DestCityName",
"CRSDepTime",
"CRSArrTime",
"DepDelay",
"ArrDelay",
"Cancelled",
"Diverted",
]

filtered_df = df[columns_to_filter]

# Drop rows where the target variable is NaN
filtered_df = filtered_df.dropna(subset=["DepDelay"])

# Define the target variable
y = filtered_df["DepDelay"]

# Drop the target variable from the features
X = filtered_df.drop("DepDelay", axis=1)

# Identify categorical and numerical columns
categorical_cols = [cname for cname in X.columns if X[cname].dtype == "object"]
numerical_cols = [cname for cname in X.columns if X[cname].dtype in ["int64", "float64"]]

# Preprocessing for numerical data: impute then scale
numerical_transformer = Pipeline(steps=[("imputer", SimpleImputer(strategy="median")), ("scaler", StandardScaler())])

# Preprocessing for categorical data: impute then one-hot encode
categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore"))
])

```

```

# Bundle preprocessing for numerical and categorical data
preprocessor = ColumnTransformer(
    transformers=[("num", numerical_transformer, numerical_cols), ("cat", categorical_transformer,
categorical_cols)]
)

# Choose the algorithm based on the command line argument
if args.algorithm == "decisiontree":
    model = DecisionTreeRegressor(random_state=0)
elif args.algorithm == "linearregression":
    model = LinearRegression()
elif args.algorithm == "svm":
    model = SVR()
else:
    model = RandomForestRegressor(n_estimators=100, random_state=0) # Default to RandomForest

# Bundle preprocessing and modeling code in a pipeline
pipeline = Pipeline(steps=[("preprocessor", preprocessor), ("model", model)])

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

# Check if the model is already saved and load it
model_filename = f"trained_model_{args.algorithm}.pkl"
if os.path.exists(model_filename):
    print(f>Loading saved {args.algorithm} model...")
    pipeline = joblib.load(model_filename)
    print(f>{args.algorithm} Model loaded successfully!")
else:
    # Measure start time
    start_time = time.time()

    # Preprocessing of training data and train model
    print(f>Starting {args.algorithm} training...")
    pipeline.fit(X_train, y_train)
    print(f>{args.algorithm} Training completed.")

    # Measure end time
    end_time = time.time()

    # Calculate and print the time taken
    time_taken = end_time - start_time
    print(f>{args.algorithm} Training completed in {time_taken:.2f} seconds.")

```

```

# Save the trained model
joblib.dump(pipeline, model_filename)
print(f"{args.algorithm} Model saved successfully!")

# Preprocessing of validation data and get predictions
preds = pipeline.predict(X_test)

# Evaluate the model
score = mean_squared_error(y_test, preds, squared=False)
print(f"RMSE for {args.algorithm}:", score)

```

Κώδικας μαζικής εκτέλεσης αλγορίθμων

```

import subprocess

# List of algorithms to prompt for
algorithms = ["randomforest", "decisiontree", "linearregression", "svm"]

script_path = "./script-savemodel.py"

# Function to run the script with the given algorithm
def run_script(algorithm):
    print(f"Running script with algorithm: {algorithm}")
    try:
        # Run the script with the specified algorithm
        subprocess.run(["python", script_path, "--algorithm", algorithm], check=True)
    except subprocess.CalledProcessError as e:
        print(f"An error occurred while running the script: {e}")

# Iterate over the algorithms
for alg in algorithms:
    input(f"Press Enter to run the script with the {alg} algorithm...")
    run_script(alg)

```

Κώδικας σχεδίασης διαγραμμάτων

```

import os

import re
import matplotlib.pyplot as plt

# Define the base directory where results are stored
base_dir = './results'

```

```

# List of algorithms
algorithms = ["randomforest", "decisiontree", "linearregression", "svm"]

# Function to extract RMSE from a file
def extract_rmse(file_path):
    with open(file_path, 'r') as file:
        content = file.read()
        # Look for the RMSE value in the content using regex
        match = re.search(r'RMSE for \w+: (\d+\.\d+)', content)
        if match:
            return float(match.group(1))
    return None

# Function to plot RMSE values
def plot_rmse(rmse_values, year):
    # Sort the algorithms by RMSE value
    sorted_algorithms = sorted(rmse_values, key=rmse_values.get)
    sorted_rmse = [rmse_values[alg] for alg in sorted_algorithms]

    # Create the bar plot
    plt.figure(figsize=(10, 6))
    plt.bar(sorted_algorithms, sorted_rmse, color='skyblue')
    plt.title(f'RMSE for Different Algorithms in {year}')
    plt.xlabel('Algorithm')
    plt.ylabel('RMSE')
    plt.tight_layout()

    # Save the plot as a PNG file
    plt.savefig(f'./rmse_chart_{year}.png')
    plt.close()

# Function to plot total comparison
def plot_total_comparison(total_rmse):
    # Create the bar plot for total comparison
    plt.figure(figsize=(12, 7))
    for alg in algorithms:
        years, rmses = zip(*total_rmse[alg]) # Unpack the list of tuples into two lists
        plt.plot(years, rmses, label=alg, marker='o')

    plt.title('Total RMSE Comparison Across Years')
    plt.xlabel('Year')
    plt.ylabel('RMSE')
    plt.legend()
    plt.tight_layout()

```

```

# Save the plot as a PNG file
plt.savefig(f'./total_rmse_comparison.png')
plt.close()

# Main function to process all results and generate charts
def main():
    # Dictionary to hold RMSE values for each algorithm and year
    rmse_data = {}
    total_rmse = {alg: [] for alg in algorithms}

    # Iterate over the algorithms and years
    for alg in algorithms:
        alg_dir = os.path.join(base_dir, alg)
        if os.path.isdir(alg_dir):
            for file_name in os.listdir(alg_dir):
                # Extract the year from the file name
                year_match = re.match(r'(\d+)\.txt', file_name)
                if year_match:
                    year = year_match.group(1)
                    file_path = os.path.join(alg_dir, file_name)
                    rmse = extract_rmse(file_path)
                    if rmse is not None:
                        if year not in rmse_data:
                            rmse_data[year] = {}
                        rmse_data[year][alg] = rmse
                        total_rmse[alg].append((year, rmse))

    # Sort the total RMSE data by year for each algorithm
    for alg in algorithms:
        total_rmse[alg].sort(key=lambda x: x[0])

    # Generate a chart for each year
    for year, data in rmse_data.items():
        plot_rmse(data, year)
        print(f'Chart for {year} generated.')

    # Generate the total comparison chart
    plot_total_comparison(total_rmse)
    print("Total comparison chart generated.")

if __name__ == '__main__':
    main()

```

12.2 Ανάλυση πηγαίου κώδικα

1. Φόρτωση των απαραίτητων βιβλιοθηκών:

```
import os
import pandas as pd
import time
import argparse
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error
from sklearn.impute import SimpleImputer
import joblib
```

Φορτώνονται οι βιβλιοθήκες που απαιτούνται για να εκτελεστεί ο κώδικας. Οι βιβλιοθήκες αυτές παρέχουν λειτουργίες και εργαλεία για το μοντέλο μηχανικής μάθησης.

2. Φόρτωση των Δεδομένων:

```
df = pd.read_csv("../Combined_Flights.csv")
```

Σε αυτό το σημείο, ο κώδικας διαβάζει ένα αρχείο CSV που περιέχει τα δεδομένα των πτήσεων αεροπλάνων για το έτος 2018.

3. Επιλογή των στηλών

```
columns_to_filter = [
    "Year",
    "Quarter",
    "Month",
    "DayofMonth",
    "DayOfWeek",
    "FlightDate",
    "Marketing_Airline_Network",
    "Flight_Number_Marketing_Airline",
    "Operating_Airline",
    "OriginAirportID",
```

```
"OriginCityName",  
"DestAirportID",  
"DestCityName",  
"CRSDepTime",  
"CRSArrTime",  
"DepDelay",  
"ArrDelay",  
"Cancelled",  
"Diverted",  
]  
  
filtered_df = df[columns_to_filter]
```

Σε αυτό το σημείο, γίνεται η επιλογή των απαραίτητων για το μοντέλο, στηλών από τα δεδομένα και δημιουργείται ένα νέο πλαίσιο δεδομένων που περιέχει μόνο αυτές τις στήλες.

4. Καθαρισμός των Δεδομένων

```
filtered_df = filtered_df.dropna(subset=["DepDelay"])
```

Σε αυτό το σημείο, θα αφαιρεθούν οι γραμμές που περιέχουν μη διαθέσιμες τιμές στη στήλη "DepDelay," δηλαδή τις πτήσεις που δεν έχουν πληροφορίες για την καθυστέρησή τους.

5. Ορισμός της Στήλης Στόχου

```
X = filtered_df.drop("DepDelay", axis=1)
```

Δημιουργία ενός πλαισίου δεδομένων που περιέχει τα χαρακτηριστικά (στήλες) που θα χρησιμοποιηθούν για την πρόβλεψη της καθυστέρησης.

6. Προ-επεξεργασία των αριθμητικών δεδομένων

```
numerical_transformer = Pipeline(steps=[("imputer", SimpleImputer(strategy="median")),
("scaler", StandardScaler())])
```

Ορίζεται ένας προ-επεξεργαστής για τα αριθμητικά δεδομένα, ο οποίος περιλαμβάνει την αντιμετώπιση των απουσιαζουσών τιμών και την κλιμάκωση των τιμών.

7. Προ-επεξεργασία των κατηγορικών δεδομένων

```
categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore"))
])
```

Ορίζεται προ-επεξεργαστής για τα κατηγορικά δεδομένα, περιλαμβάνοντας την αντιμετώπιση των απουσιαζουσών τιμών και την μετατροπή τους σε δυαδικές μεταβλητές.

8. Προσάρτηση προ-επεξεργαστή

```
categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("onehot", OneHotEncoder(handle_unknown="ignore"))
])
```

Συνδυάζονται οι προ-επεξεργαστές των αριθμητικών και κατηγορικών δεδομένων χρησιμοποιώντας έναν συνθέτη.

9. Καθορισμός του μοντέλου

```
if args.algorithm == "decisiontree":
    model = DecisionTreeRegressor(random_state=0)
elif args.algorithm == "linearregression":
    model = LinearRegression()
elif args.algorithm == "svm":
    model = SVR()
else:
    model = RandomForestRegressor(n_estimators=100, random_state=0)
```

Ορίζεται το μοντέλο που θα χρησιμοποιήσουμε για τις προβλέψεις.

10. Δημιουργία του “Σωλήνα” Προ-επεξεργασίας και Μοντέλου:

```
pipeline = Pipeline(steps=[("preprocessor", preprocessor), ("model", model)])
```

Δημιουργία σωλήνα που συνδυάζει τον προ-επεξεργαστή και το μοντέλο.

11. Διαχωρισμός των Δεδομένων σε Σύνολα Εκπαίδευσης και Επικύρωσης:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Χωρίζονται τα δεδομένα σε σύνολα εκπαίδευσης και επικύρωσης, προσφέροντας έτσι δυνατότητα εκπαίδευσης και αξιολόγησης

12. Διαχείριση του μοντέλου

```
model_filename = f"trained_model_{args.algorithm}.pkl"
if os.path.exists(model_filename):
    print(f>Loading saved {args.algorithm} model...")
    pipeline = joblib.load(model_filename)
    print(f">{args.algorithm} Model loaded successfully!")
```

Πραγματοποίηση ελέγχου ύπαρξης ενός ήδη αποθηκευμένου μοντέλου για τον επιλεγμένο αλγόριθμο. Αν υπάρχει, φορτώνεται για χρήση.

13. Εκπαίδευση του μοντέλου

```
print(f"Starting {args.algorithm} training...")
pipeline.fit(X_train, y_train)
print(f">{args.algorithm} Training completed.")
```

Αν δεν υπάρχει αποθηκευμένο μοντέλο, τότε ξεκινάει η εκπαίδευση του μοντέλου για τον επιλεγμένο αλγόριθμο. Αυτή η διαδικασία περιλαμβάνει τον προ-επεξεργασμό των δεδομένων και την εκπαίδευση του μοντέλου.

14. Αποθήκευση του μοντέλου

```
joblib.dump(pipeline, model_filename)
print(f"{args.algorithm} Model saved successfully!")
```

Αφού ολοκληρωθεί η εκπαίδευση, το μοντέλο αποθηκεύεται στον δίσκο για μελλοντική χρήση.

15. Πρόβλεψη και Αξιολόγηση

```
preds = pipeline.predict(X_test)
core = mean_squared_error(y_test, preds, squared=False)
print(f"RMSE for {args.algorithm}:", score)
```

Το εκπαιδευμένο μοντέλο χρησιμοποιείται για τις προβλέψεις στα δεδομένα επικύρωσης και υπολογίζεται το Τετραγωνικό Μέσο Σφάλμα (RMSE), το οποίο είναι ένας μετρικός δείκτης που δείχνει πόσο κοντά είναι οι προβλέψεις του μοντέλου στις πραγματικές τιμές. Αυτός ο δείκτης έχει σκοπό να βοηθήσει στην κατανόηση της απόδοσης του μοντέλου.