Πανεπιστήμιο Μακεδονίας

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Διπλωματική Εργασία:

«Προβλεπτική Αναλυτική για Καθυστερήσεις Πτήσεων στην Εφοδιαστική Αλυσίδα»

Πασχαλέρης Γεώργιος

Επιβλέπων: Κωνσταντάρας Ιωάννης

Θεσσαλονίκη,

Αύγουστος 2023

Περίληψη

Σκοπός: Η παρούσα εργασία αποσκοπεί στη μελέτη των επιπτώσεων των καθυστερήσεων που παρουσιάζονται σε πτήσεις, στην εφοδιαστική αλυσίδα και τους επιβάτες. Επίσης επιχειρείται η ανάπτυξη ενός μοντέλου Μηχανικής Μάθησης που θα είναι σε θέσει να προβλέψει παρόμοιες καθυστερήσεις

Μεθοδολογία: Για τη βιβλιογραφική ανασκόπηση χρησιμοποιήθηκαν βιβλία και επιστημονικά άρθρα της παγκόσμιας βιβλιογραφίας. Για την ανάπτυξη του μοντέλου Μηχανικής Μάθησης, χρησιμοποιήθηκε ένα σύνολο δεδομένων με πτήσεις του 2015 από τις Ηνωμένες Πολιτείες Αμερικής από το Kaggle.com. Επίσης έγινε χρήση Python και βιβλιοθηκών που βοήθησαν στη μοντελοποίηση.

Αποτελέσματα: Η επίπτωση των καθυστερήσεων είναι αδιαμφισβήτητη και η πρόβλεψή τους απαραίτητη. Το μοντέλο που δημιουργήθηκε είναι σε θέση να προβλέψει τις καθυστερήσεις αλλά αυτό γίνεται με τη συνεισφορά μιας μεταβλητής που υπάρχει πολύ μεγάλη συσχέτιση.

Λέξεις κλειδιά: Μηχανική Μάθηση, Python, Supply Chain, Πτήσεις, Καθυστέρηση Πτήσεων, Predictive Analytics

Abstract

Purpose: The present study aims to investigate the effects of delays observed in flights, on the supply chain and passengers. Additionally, there is an attempt to develop a Machine Learning model capable of predicting similar delays.

Methodology: For literature review, books and scientific articles from global literature were used. For the development of the Machine Learning model, a dataset with flights from 2015 from the United States of America from Kaggle.com was utilized. Python and libraries that assisted in modeling were also used.

Results: The impact of delays is undeniable, and their prediction is essential. The model that was created can predict delays, but this is done with the contribution of a variable that has a very strong correlation.

Keywords: Machine Learning, Python, Supply Chain, Flights, Flight Delay, Predictive Analytics

Περιεχόμενα

1 Εισαγωγή

Στην εποχή της τεχνολογίας και της παγκοσμιοποίησης, η αεροπορική βιομηχανία έχει αναδειχτεί ως ένας από τους πιο ζωτικούς τομείς της παγκόσμιας οικονομίας. Προσφέροντας μια μοναδική συνδυαστική δυνατότητα ταχύτητας και εμβέλειας, η αεροπορία έχει επαναπροσδιορίσει τον τρόπο με τον οποίο αντιλαμβανόμαστε τον χρόνο και τον χώρο.

Παρά τα οφέλη της, όμως, η αεροπορική βιομηχανία έρχεται αντιμέτωπη με πολλές προκλήσεις. Οι καθυστερήσεις και οι ακυρώσεις πτήσεων αποτελούν δύο από τις πιο εμφανείς προκλήσεις της, με εκατομμύρια επιβάτες παγκοσμίως να αντιμετωπίζουν τα άσχημα αποτελέσματα τέτοιων περιστατικών ετησίως (Hassan, Santas, & Vink, 2021).

Στο πλαίσιο του ευρύτερου επιχειρηματικού περιβάλλοντος, η αξιοπιστία των αεροπορικών μεταφορών έχει αναδειχτεί ως σημαντικός παράγοντας για την επιτυχία ή την αποτυχία των διεθνών επιχειρήσεων. Η άμεση πρόσβαση σε προϊόντα και υπηρεσίες, η ταχεία μεταφορά προσωπικού και το διεθνές εμπόριο εξαρτώνται ολοένα και περισσότερο από τη σταθερότητα και την αποτελεσματικότητα του αεροπορικού συστήματος.

Η Τεχνητή Νοημοσύνη (AI) και η Μηχανική Μάθηση είναι τεχνολογίες που αναπτύσσονται ταχύτατα και έχουν το δυναμικό να προσφέρουν λύσεις σε αυτές τις προκλήσεις. Μέσω της ανάλυσης μεγάλων όγκων δεδομένων, αυτές οι τεχνολογίες μπορούν να προβλέπουν και να διαχειρίζονται πιθανές καθυστερήσεις και ακυρώσεις, βελτιώνοντας έτσι την αποτελεσματικότητα της αλυσίδας εφοδιασμού (Li, 2021).

Επιπλέον, οι τεχνολογίες αυτές μπορούν να ενσωματωθούν στον πληροφοριακό εξοπλισμό των αεροπορικών εταιρειών, δημιουργώντας ένα πιο σύνθετο, αλλά ταυτόχρονα πιο αποδοτικό σύστημα διαχείρισης των πτήσεων. Το αποτέλεσμα θα είναι μια πιο σταθερή και αποτελεσματική

βιομηχανία που θα μπορεί να ανταποκρίνεται στις συνεχώς αυξανόμενες απαιτήσεις των επιβατών και των πελατών.

Βέβαια, η εισαγωγή των τεχνολογιών αυτών δεν είναι χωρίς προκλήσεις. Η εκπαίδευση του προσωπικού, η ενσωμάτωση των τεχνολογικών λύσεων στα υφιστάμενα συστήματα και η διασφάλιση της ιδιωτικότητας των δεδομένων των επιβατών είναι κάποιες από τις προκλήσεις που πρέπει να αντιμετωπιστούν (Badea, Zamfiroiu, & Boncea, 2018).

Στο πλαίσιο αυτό, είναι σημαντικό να εξετάσουμε πώς η αεροπορική βιομηχανία μπορεί να ωφεληθεί από την εφαρμογή της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης, αλλά και ποιες είναι οι προκλήσεις που πρέπει να υπερκεραστούν.

2 Βιβλιογραφική Ανασκόπηση

2.1 Ιστορική Αναδρομή στην Ανάλυση Καθυστερήσεων και Ακυρώσεων στην Αεροπορία

Η αεροπορική βιομηχανία, από τις πρώτες της μέρες, έχει υποστεί σημαντικές αλλαγές στην λειτουργία και διαχείρισή της. Παρότι οι πρώτες πτήσεις ήταν γεμάτες αβεβαιότητα και κινδύνους, οι τεχνολογικές και οργανωτικές εξελίξεις έχουν μετατρέψει τον τομέα σε έναν από τους πιο αξιόπιστους τρόπους μεταφοράς.

Παρόλα αυτά, καθυστερήσεις και ακυρώσεις πάντα υπήρχαν και θα υπάρχουν. Στις αρχές του 20ου αιώνα, οι κύριοι λόγοι για τέτοια περιστατικά ήταν κυρίως τεχνικά προβλήματα και καιρικές συνθήκες. Χωρίς την ύπαρξη της σύγχρονης τεχνολογίας πρόγνωσης του καιρού ή τα προηγμένα διαγνωστικά εργαλεία, οι πιλότοι και οι αεροπορικές εταιρείες ήταν πολύ περισσότερο εκτεθειμένοι σε απρόβλεπτα γεγονότα (Hassan, Santas, & Vink, 2021).

Με την εξέλιξη της τεχνολογίας, και ιδιαίτερα μετά τον Δεύτερο Παγκόσμιο Πόλεμο, τα αεροσκάφη έγιναν πιο αξιόπιστα. Ωστόσο, η διαχείριση των ροών επιβατών, τα ζητήματα με τα logistics και η εμπορική πτυχή των αεροπορικών εταιρειών έγιναν πιο πολύπλοκες, δημιουργώντας νέες προκλήσεις.

Στα τέλη του 20ου αιώνα και στις αρχές του 21ου, οι τεχνολογικές καινοτομίες, όπως τα συστήματα διαχείρισης πτήσεων και οι εφαρμογές πρόγνωσης καιρού, έχουν επιτρέψει στις αεροπορικές εταιρείες να διαχειρίζονται πιο αποτελεσματικά τις καθυστερήσεις και ακυρώσεις. Όμως, η αυξανόμενη κίνηση, τα όλο και περισσότερα γεμάτα αεροδρόμια και οι συνεχείς αλλαγές στο διεθνές περιβάλλον, συμπεριλαμβανομένων των πολιτικών και οικονομικών κρίσεων, συνεχίζουν να προκαλούν προβλήματα
.

Η ανάγκη για βελτίωση στη διαχείριση των καθυστερήσεων και ακυρώσεων έχει οδηγήσει στην αναζήτηση νέων λύσεων. Η Μηχανική Μάθηση και η Τεχνητή Νοημοσύνη, με την ικανότητά τους να αναλύουν μεγάλα σύνολα δεδομένων και να προβλέπουν συμπεριφορές, προσφέρουν μια ελπιδοφόρα προοπτική για την αντιμετώπιση των προκλήσεων αυτών (Seelhorst, 2014).

Η έρευνα σε αυτόν τον τομέα είναι ακόμα σε εξέλιξη, και οι δυνατότητες που προσφέρονται με τη χρήση των νέων τεχνολογικών λύσεων είναι πολλές. Είναι σημαντικό για την αεροπορική βιομηχανία να αναπτύξει στρατηγικές που θα ενσωματώνουν αυτές τις τεχνολογίες με τρόπο που θα βελτιώνει την εμπειρία των επιβατών, ενώ ταυτόχρονα θα μειώνει το κόστος λειτουργίας.

2.2 Ιστορία και σημασία της αεροπορικής μεταφοράς στις αλυσίδες εφοδιασμού

Η αεροπορική μεταφορά έχει αναπτυχθεί σε ένα απαραίτητο στοιχείο των παγκοσμίων αλυσίδων εφοδιασμού, παρέχοντας ταχύτητα, αποτελεσματικότητα και ευελιξία. Ας εξετάσουμε την εξέλιξή της και τη σημασία της για τις σύγχρονες επιχειρηματικές δραστηριότητες.

Κατά τη διάρκεια του πρώτου μισού του 20ου αιώνα, η αεροπορική μεταφορά ήταν κυρίως στρατιωτικής φύσης ή περιορισμένη στη μεταφορά πολιτών. Ωστόσο, μετά τον Δεύτερο Παγκόσμιο Πόλεμο, ο ερχομός της παγκοσμιοποίησης και η ανάγκη για γρήγορες μεταφορές εμπορευμάτων σε παγκόσμιο επίπεδο έθεσαν την αεροπορία στο προσκήνιο του εφοδιασμού.

Οι πρώτες εμπορευματικές αεροπορικές γραμμές παρείχαν μεταφορά για επείγοντα και πολύτιμα αντικείμενα, όπως φάρμακα, τεχνολογικό εξοπλισμό και ταχυδρομεία. Εντούτοις, καθώς οι τεχνολογίες πτήσης βελτιώθηκαν και τα αεροσκάφη μεγάλωσαν σε μέγεθος, η χωρητικότητα για τη μεταφορά εμπορευμάτων αυξήθηκε σημαντικά (Yuan, Low, & Tang, 2010).

Η ταχύτητα και η ικανότητα να παρακάμπτουν τα γεωγραφικά εμπόδια έκαναν την αεροπορία ιδανική για τη μεταφορά προϊόντων σε απομακρυσμένες αγορές με περιορισμένο χρόνο. Σύντομα, η αεροπορία

έγινε απαραίτητη για την επιτυχία των διεθνών επιχειρήσεων, ειδικά σε τομείς όπως η τεχνολογία, η φαρμακευτική βιομηχανία και όπου διαχειρίζονται πολυτελή εμπορεύματα.

Καθώς η παγκόσμια οικονομία συνέχισε να αναπτύσσεται, ο ρόλος της αεροπορικής μεταφοράς εξελίχθηκε. Οι εταιρείες χρειάζονταν ταχείες μεταφορές για να διατηρούν χαμηλά τα αποθέματα, να ανταποκρίνονται γρήγορα στις αλλαγές της αγοράς και να μειώνουν τον κίνδυνο σε περιπτώσεις αναταραχών ή κρίσεων.

Αυτό που είναι σαφές είναι ότι η αεροπορική μεταφορά δεν είναι απλά ένα μέσο μεταφοράς - είναι ένα ουσιαστικό εργαλείο για τη διασφάλιση της απρόσκοπτης λειτουργίας των παγκόσμιων αλυσίδων εφοδιασμού, παρέχοντας την ταχύτητα, την αξιοπιστία και την αποτελεσματικότητα που χρειάζονται για να παραμείνουν ανταγωνιστικές στον σύγχρονο κόσμο.

2.3 Οι συνέπειες των καθυστερήσεων και των ακυρώσεων πτήσεων στις αλυσίδες εφοδιασμού

Η αεροπορική μεταφορά είναι ένα ζωτικής σημασίας κομμάτι των παγκόσμιων αλυσίδων εφοδιασμού. Παρ' όλο που παρέχει ταχύτητα και αποτελεσματικότητα, είναι επίσης ευάλωτη σε καθυστερήσεις και ακυρώσεις. Αυτά τα προβλήματα έχουν βαθιές και πολύπλοκες συνέπειες για τις αλυσίδες εφοδιασμού (Liu, Yin, & Hansen, Economic costs of air cargo flight delays related to late package deliveries, 2019).

- Άμεσο Κόστος: Οι καθυστερήσεις και οι ακυρώσεις συχνά επιφέρουν άμεσα κόστη για τις εταιρείες. Από την ανάγκη για αναδιάταξη πτήσεων, μέχρι τις επιβαρυντικές χρεώσεις και τα πρόσθετα λειτουργικά έξοδα, το οικονομικό βάρος μπορεί να είναι σημαντικό (Zamkova, Rojik, Prokop, & Stolin, 2022).
- Διαταραχές στην Παραγωγή: Όταν τα ανταλλακτικά, τα υλικά ή τα τελικά προϊόντα καθυστερούν, μπορεί να δημιουργηθούν διαταραχές στις γραμμές

παραγωγής. Αυτό μπορεί να έχει ως αποτέλεσμα την περαιτέρω καθυστέρηση της παραγωγής ή ακόμη και τη διακοπή της.

- Επιπτώσεις στην Αξία της Επωνυμίας: Οι καταναλωτές και οι εταιρικοί πελάτες αναμένουν αξιοπιστία. Συχνές καθυστερήσεις και ακυρώσεις μπορούν να αποδυναμώσουν την εμπιστοσύνη του πελάτη και να βλάψουν τη φήμη της επιχείρησης.
- Πρόσθετες Διαδικασίες Εφοδιασμού και Διαχείρισης: Κάθε ακύρωση ή καθυστέρηση απαιτεί πρόσθετη δουλειά σε ό,τι αφορά την παρακολούθηση, την αναδιάταξη και την επικοινωνία. Αυτό αυξάνει την εργασιακή φόρτιση των ομάδων logistics και εφοδιασμού (Liu, 2019).
- Κόστος Αποθήκευσης: Η ανάγκη για πρόσθετη αποθήκευση λόγω καθυστερήσεων ή ακυρώσεων μπορεί να οδηγήσει σε υψηλότερα κόστη αποθήκευσης, ειδικά αν τα προϊόντα χρειάζονται ειδικές συνθήκες.
- Συμβατικές Ποινές: Σε πολλές περιπτώσεις, οι εταιρείες έχουν υπογράψει συμβάσεις που προβλέπουν ποινές σε περίπτωση καθυστερήσεων, καθιστώντας τις οικονομικές συνέπειες ακόμη πιο βαριές.
- Επιπτώσεις στο Περιβάλλον: Καθυστερήσεις μπορούν να οδηγήσουν σε μεγαλύτερη κατανάλωση καυσίμων λόγω της ανάγκης για επανεκκίνηση πτήσεων ή περισσότερο χρόνο στον αέρα. Αυτό μπορεί να έχει αρνητικές επιπτώσεις για το περιβάλλον.

Σε συνάρτηση με τα παραπάνω, είναι σαφές ότι οι καθυστερήσεις και οι ακυρώσεις πτήσεων έχουν πολυεπίπεδες συνέπειες για τις αλυσίδες εφοδιασμού. Οι εταιρείες πρέπει να είναι έτοιμες να αντιμετωπίσουν αυτές τις προκλήσεις, να σχεδιάζουν εκ των προτέρων και να εφαρμόζουν στρατηγικές για την ελαχιστοποίηση των επιπτώσεων τους.

2.4 Παραδοσιακές μέθοδοι πρόβλεψης και διαχείρισης καθυστερήσεων και ακυρώσεων πτήσεων

Οι καθυστερήσεις και ακυρώσεις πτήσεων αποτελούν σημαντικό πρόβλημα για την αεροπορική βιομηχανία, καθώς έχουν σημαντικές οικονομικές και λειτουργικές επιπτώσεις. Εδώ και δεκαετίες, οι αεροπορικές εταιρείες εφαρμόζουν διάφορες μεθόδους για την πρόβλεψη και διαχείριση τέτοιων προβλημάτων (Cavusoglou & Macario, 2021).

1. Ιστορικά Δεδομένα: Πολλές αεροπορικές εταιρείες χρησιμοποιούν ιστορικά δεδομένα για την πρόβλεψη καθυστερήσεων. Μελετώντας τις τάσεις του παρελθόντος, προσπαθούν να προβλέψουν τυχόν μελλοντικές καθυστερήσεις.

2. Μοντέλα Πρόβλεψης: Τα στατιστικά μοντέλα, όπως η γραμμική παλινδρόμηση, χρησιμοποιούνται ευρέως για την πρόβλεψη καθυστερήσεων, βασισμένες σε παράγοντες όπως ο καιρός, οι εποχικές διακυμάνσεις και άλλες λειτουργικές παράμετροι (Li, 2021). Επίσης χρησιμοποιούνται μοντέλα κατηγοριοποίησης όπως τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα, και τα τυχαία δάση (random forests) τα οποία μπορούν να κατηγοριοποιήσουν τις πτήσεις σε "πιθανές για καθυστέρηση" ή "όχι πιθανές για καθυστέρηση" με βάση συγκεκριμένα χαρακτηριστικά. Αυτά τα μοντέλα, όταν συνδυάζονται με τα δεδομένα και την τεχνολογία που είναι διαθέσιμα σήμερα, μπορούν να προσφέρουν πολύτιμες πληροφορίες στους φορείς λειτουργίας των αεροπορικών εταιρειών, βοηθώντας τους να λαμβάνουν πιο ενημερωμένες αποφάσεις για την καλύτερη διαχείριση των πτήσεών τους. Συνολικά, η πρόβλεψη και η κατηγοριοποίηση των καθυστερήσεων μέσω στατιστικών και αλγοριθμικών μεθόδων αποτελεί έναν ουσιαστικό τομέα εργασίας στην αεροπορική βιομηχανία, με τεράστιες επιπτώσεις στην αποτελεσματικότητα των λειτουργιών και την οικονομική τους αποδοτικότητα (Gui, et al., 2020).

3. Συστήματα Διαχείρισης: Τα συστήματα διαχείρισης της κίνησης στον αέρα και στο έδαφος είναι κρίσιμα για τον συντονισμό και τη διαχείριση των

πτήσεων. Αυτά τα συστήματα παρακολουθούν την κίνηση των αεροσκαφών και διαχειρίζονται τυχόν καθυστερήσεις ή ακυρώσεις.

4. Επικοινωνία: Η άμεση και συνεχής επικοινωνία μεταξύ του πληρώματος, των πυργίσκων ελέγχου κίνησης και των λειτουργικών κέντρων των αεροπορικών εταιρειών είναι ζωτικής σημασίας για την αποτελεσματική διαχείριση των καθυστερήσεων.

5. Διαδικασίες Έκτακτης Ανάγκης: Σε περίπτωση μεγάλων καθυστερήσεων ή ακυρώσεων, οι αεροπορικές εταιρείες εφαρμόζουν προκαθορισμένες διαδικασίες για τη φροντίδα και την εξυπηρέτηση των επιβατών, όπως την παροχή τροφής, διαμονής ή εναλλακτικών μέσων μεταφοράς.

6. Συνεργασία με Τρίτους: Οι αεροπορικές εταιρείες συχνά συνεργάζονται με άλλες εταιρείες για την παροχή υπηρεσιών στους επιβάτες τους σε περίπτωση καθυστερήσεων ή ακυρώσεων.

7. Εκπαίδευση Προσωπικού: Το εκπαιδευμένο προσωπικό μπορεί να αντιμετωπίσει καλύτερα τις καθυστερήσεις, ενημερώνοντας τους επιβάτες και λαμβάνοντας αποφάσεις για τη βέλτιστη λειτουργία.

8. Επαναπρογραμματισμός: Όταν οι πτήσεις καθυστερούν ή ακυρώνονται, ο επαναπρογραμματισμός είναι ένας τρόπος για να βρεθεί η καλύτερη δυνατή λύση, ώστε να μειωθούν οι επιπτώσεις στους επιβάτες και τη λειτουργία της εταιρείας.

9. Διαχείριση Επιβατικής Ροής: Οι αεροπορικές εταιρείες εφαρμόζουν στρατηγικές για να διασφαλίσουν ότι οι επιβάτες που πλήττονται από καθυστερήσεις ή ακυρώσεις μπορούν να προχωρήσουν ομαλά στον επόμενο προορισμό τους.

10. Τεχνολογικές Λύσεις: Η τεχνολογία επιτρέπει στις αεροπορικές εταιρείες να προβλέπουν και να διαχειρίζονται καλύτερα τις καθυστερήσεις, μέσω συστημάτων πληροφορικής, προγραμματισμού και άλλων τεχνολογικών εργαλείων.

Αναλύοντας τα παραπάνω, είναι σαφές ότι η διαχείριση καθυστερήσεων και ακυρώσεων πτήσεων αποτελεί πολύπλοκο ζήτημα, το οποίο απαιτεί συνεχή

προσαρμογή στις συνθήκες. Οι παραδοσιακές μέθοδοι παρέχουν τα βασικά εργαλεία για την αντιμετώπιση αυτού του προβλήματος, αλλά όπως και με όλες τις παραδοσιακές μεθόδους, η συνεχής αναθεώρηση, εκπαίδευση και προσαρμογή είναι απαραίτητες για να διασφαλίσουν την αποτελεσματικότητά τους.

Επιπλέον, η κοινότητα της αεροπορίας κατανοεί πλέον τη σημασία της ενσωμάτωσης νέων τεχνολογιών και μεθόδων για τη βελτίωση της ακρίβειας των προβλέψεων και της διαχείρισης των καθυστερήσεων (Cavusoglou & Macario, 2021).

Τέλος, είναι σημαντικό να ληφθεί υπόψη ότι η διαχείριση καθυστερήσεων και ακυρώσεων δεν επηρεάζει μόνο τους επιβάτες και τις αεροπορικές εταιρείες, αλλά επηρεάζει επίσης τον τουριστικό κλάδο, τους προμηθευτές, τα αεροδρόμια και πολλούς άλλους στελέχη της βιομηχανίας. Για αυτό το λόγο, η συνεχής βελτίωση των μεθόδων διαχείρισης αυτών των ζητημάτων είναι κρίσιμη για την προώθηση της βιωσιμότητας και της ανθεκτικότητας της αεροπορικής βιομηχανίας.

2.5 Οι κύριοι παράγοντες που προκαλούν καθυστερήσεις στις πτήσεις

Οι καθυστερήσεις στις πτήσεις είναι ένα ζήτημα που αντιμετωπίζουν τακτικά τόσο οι επιβάτες όσο και οι αεροπορικές εταιρείες. Υπάρχουν πολλοί παράγοντες που μπορούν να προκαλέσουν καθυστερήσεις, ενώ ορισμένοι από αυτούς είναι αναπόφευκτοι, άλλοι μπορεί να αντιμετωπιστούν με καλύτερο σχεδιασμό και διαχείριση (Asfe, Zehi, Tash, & Yaghoubi, 2014).

- Καιρικές Συνθήκες: Οι κακές καιρικές συνθήκες είναι ίσως ο πιο συχνός παράγοντας που προκαλεί καθυστερήσεις. Θυελλώδεις ανέμοι, καταιγίδες, χιονόπτωση ή ομίχλη μπορούν να καταστήσουν την απογείωση, την προσγείωση ή ακόμη και την πτήση επικίνδυνη.

- Τεχνικά Προβλήματα: Τα αεροσκάφη είναι πολύπλοκες μηχανές, και όπως όλες οι μηχανές, μπορεί να αντιμετωπίσουν τεχνικά προβλήματα. Από απλές βλάβες μέχρι σοβαρότερα ζητήματα, αυτά τα προβλήματα απαιτούν συχνά άμεση επισκευή πριν την πτήση.
- Προβλήματα στην Κυκλοφορία: Σε πολυσύχναστα αεροδρόμια, οι καθυστερήσεις μπορούν να προκληθούν λόγω της υπερφόρτωσης των διαδρόμων απογείωσης και προσγείωσης ή των ουρών στον εναέριο χώρο.
- Λειτουργικά Προβλήματα: Μερικές φορές, οι καθυστερήσεις προκαλούνται από ζητήματα στη λειτουργία της αεροπορικής εταιρείας, όπως προβλήματα στο προσωπικό, απεργίες ή άλλες εσωτερικές διαταραχές.
- Ασφάλεια και Έλεγχοι: Αυξημένοι έλεγχοι ασφαλείας, αναζητήσεις ή άλλα προβλήματα σχετικά με τους ελέγχους ασφαλείας μπορεί να προκαλέσουν καθυστερήσεις στις πτήσεις.
- Καθυστερήσεις στην Επίγεια Υπηρεσία: Προβλήματα με τη φόρτωση των αποσκευών, τον ανεφοδιασμό ή τα οχήματα που χρησιμοποιούνται για τη μεταφορά των επιβατών στο αεροσκάφος μπορούν να προκαλέσουν επίσης καθυστερήσεις.
- Ανθρώπινος Παράγοντας: Μερικές φορές, οι καθυστερήσεις μπορεί να προκληθούν από ανθρώπινα λάθη, όπως λάθη στο σχεδιασμό πτήσεων, εκπαίδευση πληρώματος ή ακόμη και επιβάτες που δεν συμμορφώνονται με τους κανονισμούς.
- Πολιτικές ή Κοινωνικές Συνθήκες: Καταστάσεις όπως πολιτική αστάθεια, τρομοκρατικές απειλές ή άλλες κρίσεις μπορεί να οδηγήσουν σε καθυστερήσεις ή ακυρώσεις πτήσεων.
- Περιορισμοί Αεροδρομίων: Ορισμένα αεροδρόμια έχουν περιορισμούς στις ώρες λειτουργίας ή στον αριθμό των πτήσεων, προκαλώντας έτσι πιθανές καθυστερήσεις.
- Εποχικότητα: Κατά τις περιόδους υψηλής εποχικότητας, όπως οι διακοπές, τα αεροδρόμια και οι αεροπορικές εταιρείες μπορούν να υπερφορτωθούν, προκαλώντας περισσότερες καθυστερήσεις.

2.6 Φυσικά φαινόμενα και κλιματολογικές συνθήκες ως κύριοι παράγοντες που προκαλούν καθυστερήσεις στις πτήσεις

Η αεροπλοΐα είναι μια από τις πλέον ευαίσθητες στον καιρό βιομηχανίες. Φυσικά φαινόμενα και κλιματολογικές συνθήκες παίζουν καίριο ρόλο στη συνέπεια των πτήσεων και την ασφάλεια των επιβατών. Τα φυσικά φαινόμενα και οι κλιματολογικές συνθήκες είναι ένας από τους πιο σημαντικούς παράγοντες που επηρεάζουν την αεροπλοΐα (Chen & Wange, 2019). Η πρόκληση για τις αεροπορικές εταιρείες είναι να βρουν τρόπους να αντιμετωπίζουν αυτές τις συνθήκες με τρόπο που εξασφαλίζει την ασφάλεια των επιβατών και του πληρώματος, ενώ ταυτόχρονα ελαχιστοποιεί τις καθυστερήσεις (Schultz, Lorenz, Schmitz, & Delgado, 2019).

- Καταιγίδες και Κεραυνοί: Οι καταιγίδες, ιδιαίτερα όταν συνοδεύονται από δυνατούς κεραυνούς, μπορούν να αποτελέσουν σοβαρό κίνδυνο για τα αεροσκάφη, κυρίως κατά την απογείωση και την προσγείωση. Οι αεροπορικές εταιρείες συχνά αναβάλλουν ή μεταθέτουν τις πτήσεις για να αποφύγουν τις συνθήκες αυτές (Brosky & Unterberger, 2019).
- Ομίχλη: Η πυκνή ομίχλη μπορεί να μειώσει την ορατότητα στο αεροδρόμιο σε επίπεδα που δεν επιτρέπουν την ασφαλή απογείωση ή προσγείωση. Πολλά αεροδρόμια διαθέτουν συστήματα που επιτρέπουν προσγειώσεις σε συνθήκες ομίχλης, αλλά ακόμη και με τέτοια συστήματα, μπορεί να υπάρξουν καθυστερήσεις.
- Χιονόπτωση και πάγος: Η χιονόπτωση και η παγωνιά μπορούν να προκαλέσουν πολλαπλά προβλήματα. Πέρα από τις συνθήκες που δυσκολεύουν την προσγείωση, το πάγωμα των φτερών και άλλων εξωτερικών επιφανειών του αεροσκάφους μπορεί να επηρεάσει την αεροδυναμική του. Επιπλέον, τα αεροδρόμια μπορεί να κλείσουν για να επιτραπεί η απομάκρυνση του χιονιού από τις πίστες (Zamkova, Rojik, Prokop, & Stolin, 2022).
- Θερμοκρασία: Οι εξαιρετικά υψηλές ή χαμηλές θερμοκρασίες μπορούν να επηρεάσουν τη λειτουργία του αεροσκάφους. Υψηλές θερμοκρασίες

μπορούν να μειώσουν την αεροδυναμική απόδοση, αλλάζοντας την πυκνότητα του αέρα.

- Ανεμοι: Δυνατοί ανέμοι ή πλευρικοί στην πίστα προσγείωσης μπορούν να καθιστούν τις προσγειώσεις δύσκολες ή αδύνατες.
- Ανωμαλίες στην Ατμόσφαιρα: Φαινόμενα όπως οι τυρβώδεις ροές αέρα μπορεί να είναι επικίνδυνα για τα αεροσκάφη, ιδιαίτερα σε μεγάλο υψόμετρο.
- Ηφαιστειακές Τέφρες: Η έκρηξη ηφαιστείων και η τέφρα που απελευθερώνεται μπορεί να καταστήσει τον αέρα επικίνδυνο για την πτήση, καθώς η τέφρα μπορεί να προκαλέσει βλάβη στους κινητήρες και άλλα κρίσιμα συστήματα του αεροσκάφους.

2.7 Δεδομένα που δημιουργούνται από ένα αεροδρόμιο και τις πτήσεις

Τα αεροδρόμια και οι πτήσεις είναι πηγές πλούσιων και πολύπλοκων δεδομένων που μπορούν να αξιοποιηθούν για μια σειρά από εφαρμογές, ιδιαίτερα στο πεδίο της μηχανικής μάθησης. Ένας αναλυτής δεδομένων που εστιάζει σε αυτόν τον τομέα έχει τη δυνατότητα να ανακαλύψει σημαντικές πληροφορίες και να δημιουργήσει εργαλεία πρόβλεψης, βελτιστοποίησης και ανάλυσης (Badea, Zamfiroiu, & Boncea, 2018).

1. Δεδομένα επιβατών: Τα αεροδρόμια καταγράφουν λεπτομερείς πληροφορίες για κάθε επιβάτη, όπως τον αριθμό της κράτησης, τον τύπο εισιτηρίου, την κατηγορία θέσης, τον προορισμό και τα σημεία αναχώρησης και άφιξης. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν για να αναλύσουν τις προτιμήσεις των επιβατών, να προβλέψουν τις κορυφαίες ώρες κίνησης ή να προσαρμόσουν τα δρομολόγια των πτήσεων.
2. Δεδομένα λειτουργίας των πτήσεων: Τα δεδομένα αυτά περιλαμβάνουν πληροφορίες για τις ώρες αναχώρησης και άφιξης, καθυστερήσεις, τύπο αεροσκάφους, διάρκεια πτήσης, κατάσταση του καιρού και πολλά άλλα. Χρησιμοποιώντας τεχνικές μηχανικής μάθησης, μπορούμε να προβλέψουμε

καθυστερήσεις πτήσεων, να αναλύσουμε τους λόγους τους ή ακόμη και να βελτιστοποιήσουμε τα δρομολόγια.

3. Δεδομένα λειτουργίας αεροδρομίου: Τα στοιχεία αυτά μπορεί να περιλαμβάνουν τον αριθμό των αεροσκαφών που βρίσκονται στον αερολιμένα, τον χρόνο αναμονής στις ουρές, την κατανομή των πτήσεων ανά ώρα ή ακόμα και την κατανάλωση καυσίμου. Ένας αναλυτής δεδομένων μπορεί να χρησιμοποιήσει αυτές τις πληροφορίες για να βελτιστοποιήσει τις λειτουργίες του αεροδρομίου.

4. Δεδομένα εμπορευματικών πτήσεων: Πολλά αεροδρόμια διαθέτουν επίσης εμπορευματικές πτήσεις, γεγονός που παράγει δεδομένα όπως το βάρος και τον τύπο των εμπορευμάτων, τον προορισμό και τον χρόνο παράδοσης. Η ανάλυση αυτών των δεδομένων μπορεί να βοηθήσει στη βελτίωση της αποδοτικότητας και της ταχύτητας της εμπορευματικής λειτουργίας.

5. Δεδομένα επικοινωνίας και ασφάλειας: Η παρακολούθηση των επικοινωνιακών δεδομένων ανάμεσα στον πύργο ελέγχου και τα αεροσκάφη, καθώς και δεδομένα ασφαλείας, όπως τα συστήματα παρακολούθησης, μπορεί να προσφέρει σημαντικές πληροφορίες για τη βελτίωση της ασφάλειας του αεροδρομίου.

6. Κοινωνικά Δίκτυα και Σχόλια Πελατών: Σε μια εποχή όπου οι επιβάτες εκφράζουν τις απόψεις τους μέσω κοινωνικών δικτύων, τα δεδομένα αυτά μπορούν να αξιοποιηθούν για τη βελτίωση των υπηρεσιών και της εμπειρίας του πελάτη.

Ως αποτέλεσμα, τα αεροδρόμια και οι πτήσεις είναι πλούσιες πηγές δεδομένων που, όταν αναλύονται σωστά, μπορούν να προσφέρουν πολύτιμες πληροφορίες για τη βελτίωση των υπηρεσιών, της αποδοτικότητας και της ασφάλειας. Ο ρόλος του αναλυτή δεδομένων σε αυτό το πεδίο είναι ζωτικής σημασίας, καθώς η ικανότητα να "διαβάζει" και να ερμηνεύει τα δεδομένα μπορεί να οδηγήσει σε σημαντικές βελτιώσεις στη λειτουργία των αεροδρομίων και των πτήσεων.

2.8 Εισαγωγή στη Μηχανική Μάθηση

Η μηχανική μάθηση (Machine Learning - ML) αποτελεί έναν τομέα της τεχνητής νοημοσύνης (Artificial Intelligence - AI) που ασχολείται με την ανάπτυξη αλγορίθμων ικανών να "μαθαίνουν" από δεδομένα. Στην πράξη, αυτό σημαίνει ότι ένα σύστημα μηχανικής μάθησης μπορεί να βελτιώσει την απόδοσή του στον χρόνο, βασιζόμενο σε προηγούμενες εμπειρίες ή νέα δεδομένα.

2.8.1 Κύριες Κατηγορίες της Μηχανικής Μάθησης:

- Επιβλεπόμενη Μάθηση (Supervised Learning): Η επιβλεπόμενη μάθηση είναι η πιο διαδεδομένη τεχνική στον τομέα της μηχανικής μάθησης. Σε αυτήν την προσέγγιση, το σύστημα εκπαιδεύεται χρησιμοποιώντας ένα σύνολο δεδομένων που περιλαμβάνει τα δεδομένα εισόδου καθώς και τις αντίστοιχες επιθυμητές εξόδους (δείκτες). Το κυρίαρχο χαρακτηριστικό είναι η ύπαρξη των ετικετών, τα οποία λειτουργούν ως "δάσκαλος", δίνοντας στον αλγόριθμο τη σωστή απάντηση για κάθε δείγμα εισόδου. Στόχος είναι το σύστημα να μάθει τον κανόνα που συνδέει τα δείγματα με τις ετικέτες τους, ώστε να μπορεί να προβλέψει την ετικέτα για νέα δεδομένα εισόδου. Η επιβλεπόμενη μάθηση χρησιμοποιείται σε εφαρμογές όπως η αναγνώριση εικόνων, η πρόβλεψη χρηματοοικονομικών δεδομένων και πολλές άλλες (Motchell, 1997).
- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Σε αντίθεση με την επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση δεν χρησιμοποιεί ετικέτες. Αντί για αυτό, το σύστημα προσπαθεί να ανακαλύψει τις υποκείμενες δομές ή τα πρότυπα στα δεδομένα. Το κυριότερο εργαλείο σε αυτήν την προσέγγιση είναι οι αλγόριθμοι συσταδοποίησης, όπως η K-μέση τιμή, που προσπαθούν να διακρίνουν ομάδες ή "συστάδες" παρόμοιων δεδομένων. Άλλες εφαρμογές περιλαμβάνουν τη μείωση της διαστατικότητας και την ανάλυση συσσωματώσεων (Burkov, 2019).
- Ενισχυτική Μάθηση (Reinforcement Learning): Στην ενισχυτική μάθηση, ο αλγόριθμος λειτουργεί ως ένας παίκτης σε ένα παιχνίδι, προσπαθώντας να

προσδιορίσει τις καλύτερες ενέργειες σε ένα δεδομένο περιβάλλον για να μεγιστοποιήσει κάποια έννοια ανταμοιβής. Εδώ, η "εκπαίδευση" είναι περισσότερο διαδικαστική, καθώς ο αλγόριθμος δοκιμάζει διάφορες ενέργειες, παρατηρεί τα αποτελέσματα και προσαρμόζει τις στρατηγικές του ανάλογα. Ένα κλασικό παράδειγμα είναι το παιχνίδι του σκακιού, όπου ο αλγόριθμος προσπαθεί να "μάθει" την καλύτερη κίνηση βάσει των προηγούμενων παιχνιδιών και των αποτελεσμάτων τους (Goodfellow, Yoshua, & Courville, 2016).

Εν κατακλείδι, τα τρία αυτά παραδείγματα είναι αντιπροσωπευτικά των βασικών προσεγγίσεων στον τομέα της μηχανικής μάθησης. Κάθε μέθοδος έχει τα δικά της πλεονεκτήματα και μειονεκτήματα και επιλέγεται ανάλογα με το πρόβλημα που πρέπει να λυθεί. Το κοινό τους χαρακτηριστικό είναι η προσπάθεια να ανακαλύψουν, να προσαρμοστούν και να προβλέψουν από τα δεδομένα, ενώ προσπαθούν να μεγιστοποιήσουν την απόδοση και την ακρίβεια.

2.8.2 Εφαρμογές της Μηχανικής Μάθησης στην Ανάλυση Δεδομένων:

Η μηχανική μάθηση έχει διεισδύσει σε κάθε τομέα της τεχνολογίας, μετατρέποντας τον τρόπο με τον οποίο αντιλαμβανόμαστε και αναλύουμε τα δεδομένα. Από τη βελτιστοποίηση των επιχειρησιακών διαδικασιών μέχρι την ανακάλυψη νέων ιατρικών θεραπειών, οι εφαρμογές της είναι αμέτρητες (Najafabadi, et al., 2015).

- Αναγνώριση Φωνής και Δεδομένων: Ψηφιακοί βοηθοί όπως η Siri και η Alexa βασίζονται στην αναγνώριση φωνής, αλλά το παρασκήνιο περιλαμβάνει την ανάλυση δεδομένων προκειμένου να παράγουν συνεκτικές απαντήσεις.
- Αναγνώριση Εικόνων και Επεξεργασία: Αναλυτικά εργαλεία που κάνουν χρήση deep learning μπορούν να αναγνωρίζουν πρότυπα σε φωτογραφίες και βίντεο, να επεξεργάζονται μεγάλους όγκους δεδομένων για αναγνώριση προσώπων, αντικειμένων και περισσότερα.

- Ανάλυση Κειμένου: Οι αλγόριθμοι Μηχανικής Μάθησης μπορούν να "διαβάσουν" και να κατανοήσουν το περιεχόμενο των κειμένων, εκτελώντας εργασίες όπως η αναγνώριση συναισθημάτων, ταξινόμηση και άλλες.
- Προβλέψεις και Διαγνωστικές Αναλύσεις: Στην αγορά μετοχών, η μηχανική μάθηση μπορεί να προβλέψει τις μελλοντικές τιμές βασισμένη σε ιστορικά δεδομένα, ενώ στην ιατρική, μπορεί να βοηθήσει στην ανίχνευση ασθενειών από εικόνες ή γενετικά δεδομένα.
- Προτεινόμενα Συστήματα: Ανάλυση των προτιμήσεων των χρηστών με βάση την ιστορική τους συμπεριφορά, για να προτείνουν προϊόντα ή υπηρεσίες.
- Ανάλυση κατηγοριοποίησης: Η κατηγοριοποίηση των δεδομένων σε ομάδες βασιζόμενη σε παρατηρούμενα χαρακτηριστικά.

Η μηχανική μάθηση και η ανάλυση δεδομένων είναι δύο τομείς που περιλαμβάνουν μια σειρά εργαλείων, τεχνικών και μεθοδολογιών που συνεργάζονται για τη βελτίωση της λήψης αποφάσεων και της προσαρμογής σε έναν κόσμο που είναι γεμάτος με πληροφορίες.

2.8.3 Προκλήσεις στη Μηχανική Μάθηση

Η τεχνολογία της μηχανικής μάθησης, παρά την εντυπωσιακή της εξέλιξη και τις εφαρμογές της σε διάφορους τομείς, αντιμετωπίζει αρκετές προκλήσεις. Η διαδικασία εκπαίδευσης ενός αλγορίθμου μηχανικής μάθησης είναι συχνά πολύπλοκη και μπορεί να παρουσιάζει διάφορες προκλήσεις (L'heureux, Grolinger, Elyamany, & Capretz, 2017).

- Υπερεκπαίδευση (Overfitting): Όταν ένα μοντέλο "μαθαίνει" τα δεδομένα εκπαίδευσης σε βαθμό που λαμβάνει υπόψη του ακόμα και τον θόρυβο ή τις ατυχίες, τότε έχουμε το φαινόμενο της υπερεκπαίδευσης. Αυτό σημαίνει ότι το μοντέλο μπορεί να έχει άριστη απόδοση στα δεδομένα εκπαίδευσης, αλλά όταν το δοκιμάζουμε σε νέα, ανεξάρτητα δεδομένα, η απόδοσή του πέφτει δραματικά.
- Υποεκπαίδευση (Underfitting): Το ακριβώς αντίθετο της υπερεκπαίδευσης. Όταν ένα μοντέλο είναι υποεκπαιδευμένο, δεν έχει "μάθει"

επαρκώς τα χαρακτηριστικά των δεδομένων, οπότε έχει χαμηλή απόδοση τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα ελέγχου.

- Ανισορροπία Δεδομένων (Data Imbalance): Σε πολλές περιπτώσεις, τα δεδομένα που έχουμε δεν είναι ισορροπημένα. Για παράδειγμα, στην ανίχνευση απάτης, τα περιστατικά πραγματικής απάτης είναι συχνά λίγα σε σχέση με τα περιστατικά όπου δεν υπάρχει απάτη. Τα μοντέλα που εκπαιδεύονται σε αυτά τα ανισόρροπα σύνολα δεδομένων είναι πιθανό να παραβλέπουν την πιο σπάνια κλάση.
- Προκαταλήψεις και Διακρίσεις (Bias and Discrimination): Αν τα δεδομένα που χρησιμοποιούμε για την εκπαίδευση των μοντέλων περιλαμβάνουν προκαταλήψεις, τότε τα μοντέλα θα "μάθουν" αυτές τις προκαταλήψεις και θα τις αναπαράγουν στις προβλέψεις τους.
- Πρόκληση της Διαφάνειας (Model Transparency): Ορισμένα προηγμένα μοντέλα, όπως τα νευρωνικά δίκτυα, μπορεί να έχουν εξαιρετική απόδοση, αλλά είναι "μαύρα κουτιά" στο βαθμό που είναι δύσκολο να κατανοήσουμε πώς λαμβάνουν τις αποφάσεις τους.
- Εξωτερικοί Παράγοντες (External Factors): Τα μοντέλα μηχανικής μάθησης εκπαιδεύονται σε ιστορικά δεδομένα. Αυτό σημαίνει ότι αν υπάρξουν αλλαγές στο περιβάλλον, τα μοντέλα μπορεί να μην ανταποκρίνονται επαρκώς.
- Διαχείριση Μεγάλων Όγκων Δεδομένων (Big Data Management): Οι πηγές δεδομένων συνεχώς αυξάνονται. Η εκπαίδευση σε τεράστια σύνολα δεδομένων απαιτεί σημαντικούς πόρους και υπολογιστική ισχύ.

Συνοψίζοντας, ενώ η μηχανική μάθηση προσφέρει τεράστιες δυνατότητες για ποικίλες εφαρμογές, οι προκλήσεις που παρουσιάζει είναι πολύπλοκες και απαιτούν προσεκτική διαχείριση και ανάλυση.

2.8.4 Εργαλεία και Πλατφόρμες στη Μηχανική Μάθηση

Στον κόσμο της μηχανικής μάθησης, η επιλογή του σωστού εργαλείου ή της πλατφόρμας είναι κρίσιμη για την αποτελεσματικότητα και την απόδοση των μοντέλων. Δύο κορυφαίες γλώσσες προγραμματισμού που χρησιμοποιούνται ευρέως στον τομέα αυτόν είναι η Python και η R. Ενώ η Python προσφέρει

ευελιξία και είναι ευρέως υποστηριζόμενη, η R είναι εξειδικευμένη για στατιστική ανάλυση και δεδομένα (Gollapudi, 2016).

TensorFlow:

Περιγραφή: Η TensorFlow είναι μια ανοιχτού κώδικα βιβλιοθήκη μηχανικής μάθησης που αναπτύχθηκε από την Google και είναι γραμμένη σε Python. Σχεδιάστηκε για να παρέχει ένα ευέλικτο πλαίσιο για την ανάπτυξη και την εκτέλεση μεγάλων δεδομένων και νευρωνικών δικτύων.

Πλεονεκτήματα:

- Δυνατότητα εκτέλεσης σε πολλαπλές πλατφόρμες, συμπεριλαμβάνοντας GPUs.
- Οπτικοποίηση με TensorBoard, που παρέχει διαγράμματα και αναπαραστάσεις.
- Καλά τεκμηριωμένη και με μεγάλη κοινότητα.

Keras:

Περιγραφή: Η Keras είναι μια υψηλού επιπέδου βιβλιοθήκη νευρωνικών δικτύων, γραμμένη σε Python. Σχεδιάστηκε για να είναι ευέλικτη, χρησιμοποιώντας TensorFlow, Theano ή Microsoft Cognitive Toolkit ως backend.

Πλεονεκτήματα:

- Διασυνδεδεμένη και ευέλικτη δομή, καθιστώντας την ανάπτυξη και την εκπαίδευση νευρωνικών δικτύων πιο ευχάριστη.
- Κατάλληλη για ερευνητές και προγραμματιστές που θέλουν να πειραματιστούν με διαφορετικές αρχιτεκτονικές δικτύων.

Scikit-learn:

Περιγραφή: Η Scikit-learn είναι μια ανοιχτού κώδικα βιβλιοθήκη για την Μηχανική Μάθηση στην Python. Παρέχει απλές και αποτελεσματικές εργαλειοτεχνικές εργασίες για την ανάλυση δεδομένων (Bisong, 2019).

Πλεονεκτήματα:

- Παρέχει μεγάλη γκάμα αλγορίθμων, από ταξινόμηση και παλινδρόμηση μέχρι clustering και προεπεξεργασία δεδομένων.
- Εύκολος συνδυασμός με άλλες βιβλιοθήκες Python όπως NumPy και Pandas.
- Καλά τεκμηριωμένη με πλούσια παραδείγματα και tutorials.

Καταληκτικά, η επιλογή της σωστής πλατφόρμας ή βιβλιοθήκης εξαρτάται από τη φύση του προβλήματος, τις ανάγκες του χρήστη και τις προτιμήσεις του στον προγραμματισμό. Ενώ η TensorFlow και η Keras είναι ιδανικές για νευρωνικά δίκτυα και είναι γραμμένες στη γλώσσα Python, η Scikit-learn είναι περισσότερο γενικής χρήσης και καλύπτει πολλούς αλγορίθμους μηχανικής μάθησης, επίσης σε Python. Η R, από την άλλη πλευρά, προσφέρει πλούσια πακέτα για στατιστική ανάλυση και μηχανική μάθηση.

Σε κάθε περίπτωση, η δυνατότητα της κοινότητας να αναπτύσσει και να υποστηρίζει αυτά τα εργαλεία καθιστά τη μηχανική μάθηση πιο προσβάσιμη και ευέλικτη για τους ερευνητές, τους προγραμματιστές και τους επαγγελματίες της βιομηχανίας.

2.9 Προηγούμενες έρευνες που χρησιμοποιούν μηχανική μάθηση για την πρόβλεψη καθυστερήσεων πτήσεων

- Στην έρευνα που πρότειναν οι (Jiang, Miao, Zhang, & Le, 2020), χρησιμοποιήθηκε η μέθοδος του sparse autoencoder deep learning για τη δημιουργία και πρόβλεψη των αεροπορικών κυκλοφοριακών διατάξεων. Στο μοντέλο πρόβλεψης, χρησιμοποιήθηκε το μοντέλο της λογιστικής παλινδρόμησης. Η ακρίβεια αυτού του μοντέλου κυμαινόταν από 85% έως 90%. Ωστόσο, τα αποτελέσματα επετεύχθησαν μόνο με τη χρήση μόνο της λογιστικής παλινδρόμησης για τη δημιουργία του μοντέλου, η ακρίβεια μπορεί να βελτιωθεί με τη χρήση πιο ισχυρών μοντέλων.
- Μια μέθοδος βαθιάς μάθησης χρησιμοποιήθηκε στην έρευνα που πρότειναν οι (Pamplona, Weigang, de Barros, Shiguemori, & Alves, 2018).

Χρησιμοποιώντας ένα τεχνητό νευρωνικό δίκτυο, δημιουργήθηκε ένα προγνωστικό μοντέλο με ακρίβεια 90%. Σύμφωνα με αυτή την έρευνα, οι τρεις κρίσιμες συνιστώσες σε αυτό το πείραμα είναι η ημέρα της εβδομάδας, η ώρα πτήσης και η αεροπορική εταιρεία. Το μοντέλο μπορεί να βελτιωθεί αν χρησιμοποιηθούν μετεωρολογικά δεδομένα στην εκπαίδευση.

- Στην έρευνα που πρότειναν οι (Kim, Choi, Briceno, & Mavris, 2016) δημιουργήθηκε μέθοδος recurrent neural network (RNN) με πολλαπλά μοντέλα συνδυασμένα. Η ακρίβεια του μοντέλου κυμαινόταν από 85% έως 87%.
- Εώς τώρα, συζητήθηκαν τα προγνωστικά μοντέλα της εποπτευόμενης μηχανικής μάθησης και της βαθιάς μάθησης. Αλλά στην έρευνα που πρότειναν οι (Xing & Tang, 2016), οι ερευνητές χρησιμοποίησαν έναν γενετικό αλγόριθμο για την βελτιστοποίηση της κυκλοφορίας του εναέριου χώρου. Κάνοντας αυτό, οι καθυστερημένες πτήσεις διανέμονται έτσι ώστε ο συνολικός χρόνος καθυστέρησης να μειώνεται για όλες τις πτήσεις. Αυτό το μοντέλο εξετάστηκε σε δεδομένα που συλλέχθηκαν από ένα αεροδρόμιο στη Κίνα. Κάποιοι παράγοντες που προκαλούν καθυστερήσεις στις πτήσεις είναι σημαντικοί, ενώ άλλοι είναι δευτερεύοντες.
- Στην έρευνα που πρότειναν οι (Dhanawade, Deo, Khanna, & Deolekar, 2019), αναλύθηκαν τέτοιοι παράγοντες. Το πείραμα έγινε σε εγχώριες και διεθνείς πτήσεις στην Ινδία, και τα δεδομένα περιλαμβάνουν 14 διαφορετικές αεροπορικές εταιρείες. Στην Ινδία, παρατηρήθηκε αύξηση των εγχώριων επιβατών κατά 10,76% και των διεθνών επιβατών κατά 8,32% κατά την περίοδο 2007-08 έως 2017-18. Η κρίσιμη ανάλυση δημιούργησε ερμηνείες όπως, για ορισμένες αεροπορικές εταιρείες, οι καθυστερήσεις των πτήσεων είναι λιγότερες, ενώ για άλλες είναι περισσότερες. Έτσι, σύμφωνα με αυτή την έρευνα, η αεροπορική εταιρεία είναι ουσιαστική για τη δημιουργία ενός προγνωστικού μοντέλου.
- Ένα άλλο προγνωστικό μοντέλο δημιουργήθηκε χρησιμοποιώντας το δίκτυο Μακροπρόθεσμης Μνήμης Βραχυπρόθεσμης Διάρκειας (Long short-term memory LSTM) στην έρευνα που πρότειναν οι (Jiang, Miao, Zhang, & Le, 2020). Χρησιμοποιήθηκαν τα Tensorflow και Keras για τη δημιουργία αυτού του μοντέλου. Το σφάλμα του μοντέλου είναι 6,76% και μπορεί να

βελτιωθεί αν χρησιμοποιηθούν περισσότερα δεδομένα για την εκπαίδευση. Ο καιρός, η εναέρια κίνηση κ.ά. είναι οι κύριοι παράγοντες που επηρεάζουν τις καθυστερήσεις των πτήσεων.

- Στην έρευνα που πρότειναν οι (Demir & Demir, 2017), τα δεδομένα συλλέχθηκαν με τη βοήθεια αισθητήρων που είχαν τοποθετηθεί στα αεροδρόμια. Επίσης, χρησιμοποιήθηκαν πληροφορίες για το ωράριο πτήσεων μαζί με τα δεδομένα από τους αισθητήρες. Ένα τεχνητό νευρωνικό δίκτυο χρησιμοποιήθηκε για τη δημιουργία ενός προγνωστικού μοντέλου και η ακρίβεια του μοντέλου είναι 96%. Επίσης, παρατηρήθηκε ότι η καθυστέρηση σε μια πτήση προκαλεί καθυστερήσεις στις πτήσεις που είναι προγραμματισμένες μετά από αυτή.
- Οι ερευνητές ανέλυσαν τους κυριότερους παράγοντες που επηρεάζουν τις καθυστερήσεις πτήσεων στην έρευνα που πρότειναν οι (Gao, Huyan, & Ju, 2015). Για αυτή την έρευνα, τα δεδομένα συλλέχθηκαν από το αεροδρόμιο του Πεκίνου και εκπαιδεύτηκε ένα νευρωνικό δίκτυο. Το σφάλμα αυτού του μοντέλου είναι 25%, το οποίο αντιστοιχεί σε σφάλμα 10 λεπτών. Επιπροσθέτως, αυτό το μοντέλο μπορεί να δοκιμαστεί σε δεδομένα που συλλέγονται από άλλα αεροδρόμια.

3 Μεθοδολογία

Το πρόβλημα που καλείται να επιλύσει το παρόν έργο είναι η ακριβής πρόβλεψη των καθυστερήσεων πτήσεων όταν έχουμε στη διάθεσή μας ορισμένα χαρακτηριστικά της πτήσης, όπως οι αεροπορικές εταιρείες που τις εκτελούν, η απόσταση που πρέπει να καλυφθεί, το αεροδρόμιο αναχώρησης, το αεροδρόμιο άφιξης, οι ώρες αναχώρησης και άλλα. Η δυνατότητα ακριβούς πρόβλεψης των καθυστερήσεων πτήσεων μπορεί να βοηθήσει τους επιβάτες να γνωρίζουν ποιες καθυστερήσεις πρέπει να είναι έτοιμοι να αντιμετωπίσουν, ανάλογα με το πού πετούν από και τις αεροπορικές εταιρείες που επιλέγουν. Αυτό μπορεί να τους επιτρέψει να λάβουν ένα περιθώριο, έτσι ώστε να μην χάσουν συνδέσεις πτήσεων ή συναντήσεις. Ο στόχος του έργου

θα είναι η εκτέλεση εμβριθούς ανάλυσης των δεδομένων και η εξέταση των χαρακτηριστικών εισόδου για να δούμε πώς αλλάζει η ακρίβεια της πρόβλεψης. Η ανάπτυξη μοντέλων πρόβλεψης που είναι αποδοτικά είναι δύσκολη λόγω της πολύπλοκης φύσης της αεροπορικής μεταφοράς.

Ο στόχος είναι να προβλέψουμε αρχικά την καθυστέρηση άφιξης που θα υποστεί η πτήση, συμπεριλαμβάνοντας την καθυστέρηση αναχώρησης ως μία από τις χαρακτηριστικές μεταβλητές. Στη συνέχεια, η χαρακτηριστική μεταβλητή της καθυστέρησης αναχώρησης θα αφαιρεθεί και θα προσπαθήσουμε να την προβλέψουμε χρησιμοποιώντας τις υπόλοιπες μεταβλητές και να αξιολογήσουμε πόσο καλά αποδίδει το μοντέλο μας. Επιπλέον, το σχέδιο είναι να τρέξουμε ορισμένα μοντέλα ταξινόμησης για να δούμε πώς αποδίδουν στο σύνολο των δεδομένων μας.

3.1 Τα Δεδομένα

Το Γραφείο Στατιστικών Μεταφορών αποτελεί μία από τις κύριες ομοσπονδιακές στατιστικές υπηρεσίες. Είναι υπεύθυνο για την παρακολούθηση της επίδοσης της έγκαιρης άφιξης των εσωτερικών πτήσεων που εκτελούνται από διάφορες αεροπορικές εταιρείες στις Ηνωμένες Πολιτείες. Τα δεδομένα που χρησιμοποιούνται για αυτήν τη μελέτη προέρχονται από το Kaggle.com (https://www.kaggle.com/datasets/usdot/flight-delays). Υπάρχουν τρία σύνολα δεδομένων: airlines.csv, air-ports.csv και flights.csv. Το αρχείο airlines.csv περιλαμβάνει τους αριθμούς IATA (Διεθνούς Συνδέσμου Αερομεταφοράς) όλων των αεροπορικών εταιρειών που λειτούργησαν το 2015. Στα δεδομένα περιλαμβάνονται συνολικά 14 αεροπορικές εταιρείες. Το αρχείο airports.csv περιλαμβάνει τα δεδομένα όλων των αεροδρομίων που λειτούργησαν το 2015 στις ΗΠΑ, καθώς και τις γεωγραφικές συντεταγμενες των αεροδρομίων. Με τη χρήση αυτών των δεδομένων και της βιβλιοθήκης Pandas της Python, μπορούν να γίνουν πραγματικές τοποθεσίες στο χάρτη των Ηνωμένων Πολιτειών. Στο επόμενο τμήμα ανάλυσης, θα πραγματοποιηθούν αυτές οι ενέργειες. Το αρχείο flights.csv περιλαμβάνει τα

κρίσιμα δεδομένα για αυτήν τη μελέτη, περιλαμβανομένων των δεδομένων όλων των πτήσεων που εκτελέστηκαν ή ακυρώθηκαν το 2015. Αυτά τα δεδομένα, μαζί με τα δεδομένα του airports.csv, θα χρησιμοποιηθούν για τη δημιουργία αναλύσεων και προβλεπτικών μοντέλων σε αυτήν την έρευνα. Περιέχει πληροφορίες για πτήσεις από το 2015 και περιλαμβάνει τις ακόλουθες πληροφορίες για τις πτήσεις:

- Year
- Month
- Day
- Day Of Week
- Airline
- Flight Number
- Tail Number
- Origin Airport
- Destination Airport
- Scheduled Departure
- Departure Time
- Departure Delay
- Taxi Out
- Wheels Off
- Scheduled Time
- Elapsed Time
- Air Time
- Distance
- Wheels On
- Taxi In
- Scheduled Arrival
- Arrival Time
- Arrival Delay
- Diverted
- Cancelled
- Cancellation Reason

- Air System Delay
- Security Delay
- Airline Delay
- Late Aircraft Delay
- Weather Delay
- Unnamed
- Country

3.2 Διερευνητική Ανάλυση των Δεδομένων

3.2.1 Αριθμητικές στήλες πτήσεων

Statistical Summary of Numerical Columns

	SCHEDULED_DEPARTURE	DEPARTURE_TIME	DEPARTURE_DELAY	SCHEDULED_ARRIVAL	ARRIVAL_TIME	ARRIVAL_DELAY	AIR_SYSTEM_DELAY	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
count	581908.00	573300.00	573300.00	581908.00	572673.00	571448.00	105977.00	105977.00	105977.00	105977.00	105977.00
mean	1329.66	1335.21	9.37	1494.34	1476.12	4.42	13.54	0.08	19.01	23.67	2.93
std	483.90	496.59	37.22	506.91	526.49	39.46	28.39	2.17	47.96	43.81	20.69
min	1.00	1.00	-55.00	1.00	1.00	-81.00	0.00	0.00	0.00	0.00	0.00
25%	917.00	921.00	-5.00	1110.00	1059.00	-13.00	0.00	0.00	0.00	0.00	0.00
50%	1325.00	1330.00	-2.00	1520.00	1512.00	-5.00	2.00	0.00	2.00	3.00	0.00
75%	1730.00	1740.00	7.00	1918.00	1916.00	8.00	18.00	0.00	19.00	29.00	0.00
max	2359.00	2400.00	1670.00	2359.00	2400.00	1665.00	855.00	241.00	1665.00	1294.00	1152.00

Πίνακας 1 Αριθμητικές στήλες flights.csv

1. **Χρόνοι Αναχώρησης και Άφιξης:**

- Οι προγραμματισμένοι χρόνοι αναχώρησης κυμαίνονται από 1 έως 23:59, με μεσαίο χρόνο στις 13:25 (που είναι περίπου 1:25 μ.μ.).

- Οι πραγματικοί χρόνοι αναχώρησης εμφανίζουν μια μικρή καθυστέρηση κατά μέσο όρο, με μέσο όρο χρόνο 13:35.

- Ομοίως, οι προγραμματισμένοι χρόνοι άφιξης κυμαίνονται από 1 έως 23:59, με μεσαίο χρόνο στις 15:20 (που είναι περίπου 3:20 μ.μ.).

- Οι πραγματικοί χρόνοι άφιξης έχουν ένα μέσο όρο 14:76, πράγμα που υποδηλώνει ότι, κατά μέσο όρο, ίσως τα αεροπλάνα να φτάνουν νωρίτερα από το προγραμματισμένο.

2. **Καθυστερήσεις:**

- Η μέση καθυστέρηση αναχώρησης είναι περίπου 9,37 λεπτά.

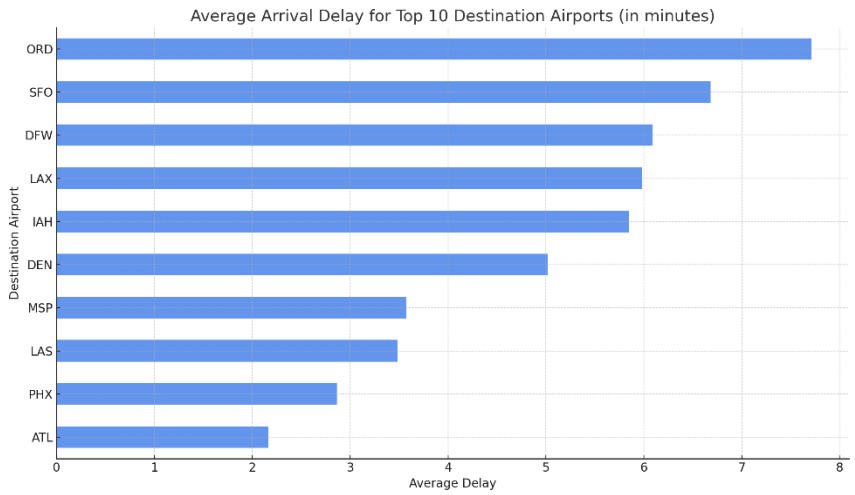
- Η μέση καθυστέρηση άφιξης είναι περίπου 4,42 λεπτά.

- Οι καθυστερήσεις λόγω του αεροπορικού συστήματος έχουν μέσο όρο περίπου 13,54 λεπτά, ενώ οι καθυστερήσεις των αεροπορικών εταιρειών και οι καθυστερήσεις λόγω καθυστέρησης αεροσκαφών έχουν μέσο όρο περίπου 19 και 23,67 λεπτά αντίστοιχα.

- Οι καθυστερήσεις λόγω ασφαλείας είναι πολύ ελάχιστες, με μέσο όρο περίπου 0,075 λεπτά.

- Οι καθυστερήσεις λόγω καιρού, παρόλο που είναι σπάνιες, μπορεί να είναι σημαντικές όταν συμβαίνουν, με μέσο όρο περίπου 2,93 λεπτά.

3.2.2 Καθυστέρηση στα Αεροδρόμια

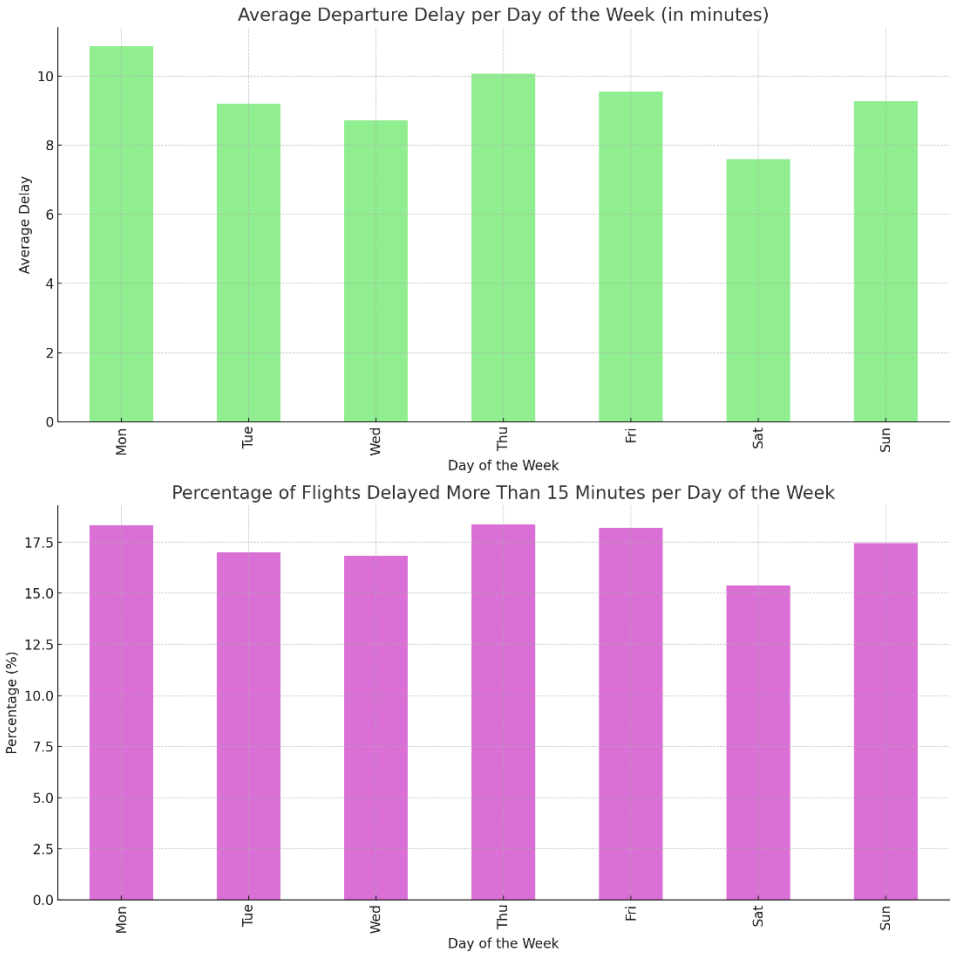


Γράφημα 1 Καθυστέρηση Αφίξεων για τα 10 πολυσύχναστα αεροδρόμια

- Από τα δέκα κορυφαία αεροδρόμια προορισμού, το αεροδρόμιο ORD έχει την υψηλότερη μέση καθυστέρηση άφιξης.
- Τα αεροδρόμια ATL και PHX έχουν σχετικά χαμηλές μέσες καθυστερήσεις σε σύγκριση με τα υπόλοιπα αεροδρόμια στη λίστα.

Μπορεί να υπάρχουν πολλοί παράγοντες που επηρεάζουν τις καθυστερήσεις σε ένα συγκεκριμένο αεροδρόμιο, όπως οι καιρικές συνθήκες, η κυκλοφοριακή ροή των αεροπλάνων, οι εγκαταστάσεις του αεροδρομίου και άλλοι λογιστικοί παράγοντες.

3.2.3 Καθυστέρηση ανα ημέρες



Γράφημα 2 Ανάλυση καθυστέρησης για καθε ημέρα

3. **Μέση Καθυστέρηση Αναχώρησης ανά Ημέρα της Εβδομάδας**:

- Το Σάββατο φαίνεται να έχει τη χαμηλότερη μέση καθυστέρηση αναχώρησης.

- Η Τετάρτη και η Παρασκευή έχουν τις υψηλότερες μέσες καθυστερήσεις αναχώρησης.

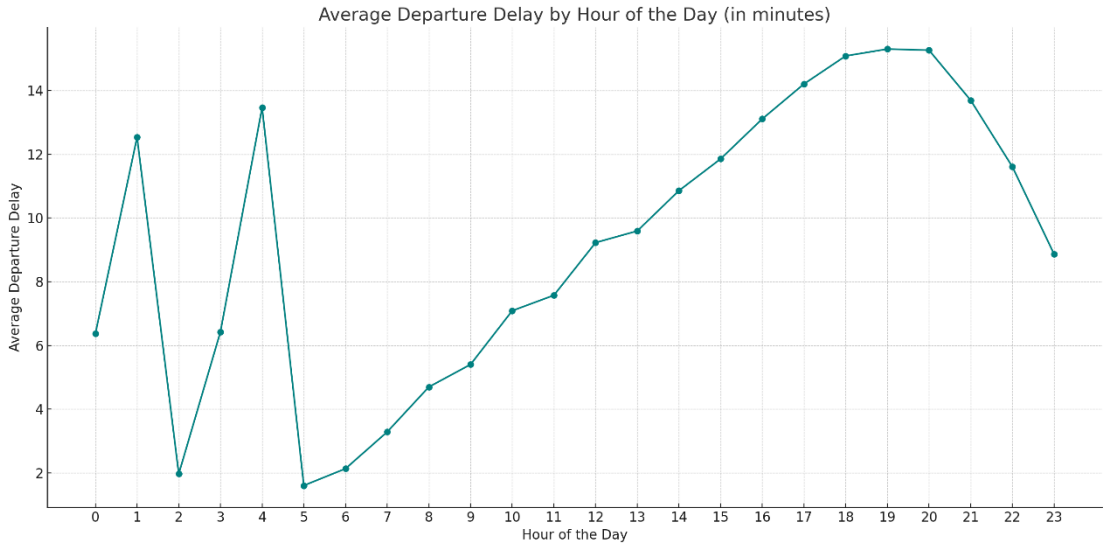
4. **Ποσοστό Πτήσεων με Καθυστέρηση πάνω από 15 Λεπτά ανά Ημέρα της Εβδομάδας**:

- Η Τετάρτη και η Παρασκευή έχουν επίσης το υψηλότερο ποσοστό πτήσεων που καθυστέρησαν πάνω από 15 λεπτά.

- Το Σάββατο έχει το χαμηλότερο ποσοστό καθυστερημένων πτήσεων, συμβαδίζοντας με τα ευρήματα για τις μέσες καθυστερήσεις.

Είναι σημαντικό να σημειώσουμε ότι το Σάββατο φαίνεται να είναι η καλύτερη ημέρα για να ταξιδέψει κανείς από πλευράς καθυστερήσεων, ενώ η Τετάρτη και η Παρασκευή φαίνεται να είναι οι χειρότερες

3.2.4 Καθυστέρηση ανά ώρα

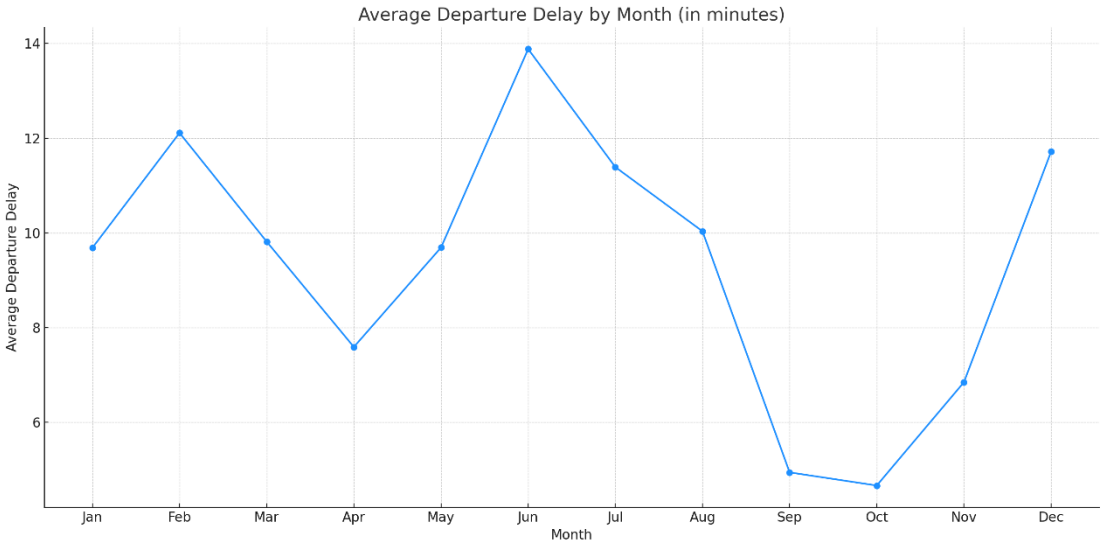


Γράφημα 3 Καθυστέρηση ανά ώρα

- Οι μέσες καθυστερήσεις αναχώρησης αυξάνονται σημαντικά τις νυχτερινές ώρες, ειδικά από τις 15:00 έως τις 22:00.
- Υπάρχει ένα σχετικά χαμηλό επίπεδο καθυστερήσεων από τις 5:00 έως τις 11:00, μετά από το οποίο οι καθυστερήσεις αρχίζουν να αυξάνονται εκ νέου.
- Η υψηλότερη μέση καθυστέρηση σημειώνεται γύρω στις 19:00.

Αυτό μπορεί να οφείλεται σε πολλούς παράγοντες, όπως η αυξημένη κυκλοφορία των αεροπλάνων τις νυχτερινές ώρες, οι καθυστερήσεις που συσσωρεύονται καθ' όλη τη διάρκεια της ημέρας και επηρεάζουν τις νυχτερινές πτήσεις, ή άλλες λογιστικές προκλήσεις που εμφανίζονται τη νύχτα.

3.2.5 Καθυστέρηση ανά μήνα



Γράφημα 4 Καθυστέρηση ανά μήνα

- Οι μήνες Φεβρουάριο, Ιούνιος, Ιούλιος και Δεκέμβριος έχουν τις υψηλότερες μέσες καθυστερήσεις αναχώρησης. Αυτό μπορεί να οφείλεται στο ότι είναι μήνες διακοπών, με αυξημένη κίνηση αεροπλάνων και περισσότερους επιβάτες.
- Ο Οκτώβριος έχει τη χαμηλότερη μέση καθυστέρηση αναχώρησης. Αυτό μπορεί να οφείλεται σε λιγότερη κίνηση αεροπλάνων ή λιγότερες καιρικές παρεμβολές.

Τα στοιχεία αυτά μπορεί να βοηθήσουν τους ταξιδιώτες να σχεδιάσουν καλύτερα τα ταξίδια τους, γνωρίζοντας τους μήνες που είναι πιο πιθανό να αντιμετωπίσουν καθυστερήσεις.

3.3 Προεπεξεργασία Δεδομένων

Πριν αρχίσουμε τη διαδικασία εκπαίδευσης των μοντέλων, είναι απαραίτητο να πραγματοποιήσουμε ορισμένα βήματα προεπεξεργασίας. Οι τεχνικές και

μεθοδολογίες που χρησιμοποιούνται για την προεπεξεργασία αναφέρονται παρακάτω

3.3.1 Χειρισμός των απουσών τιμών

Μεταβλητή	Ποσοστό Κενών Τιμών
YEAR	0.00%
MONTH	0.00%
DAY	0.00%
DAY_OF_WEEK	0.00%
AIRLINE	0.00%
FLIGHT_NUMBER	0.00%
TAIL_NUMBER	0.25%
ORIGIN_AIRPORT	0.00%
DESTINATION_AIRPORT	0.00%
SCHEDULED_DEPARTURE	0.00%
DEPARTURE_TIME	1.48%
DEPARTURE_DELAY	1.48%
TAXI_OUT	1.53%
WHEELS_OFF	1.53%
SCHEDULED_TIME	0.00%
ELAPSED_TIME	1.81%
AIR_TIME	1.81%
DISTANCE	0.00%
WHEELS_ON	1.59%
TAXI_IN	1.59%
SCHEDULED_ARRIVAL	0.00%
ARRIVAL_TIME	1.59%
ARRIVAL_DELAY	1.81%
DIVERTED	0.00%
CANCELLED	0.00%
CANCELLATION_REASON	98.46%
AIR_SYSTEM_DELAY	81.72%
SECURITY_DELAY	81.72%
AIRLINE_DELAY	81.72%
LATE_AIRCRAFT_DELAY	81.72%
WEATHER_DELAY	81.72%

Πίνακας 2 Ποσοστό Κενών Τιμών

Οι μεταβλητές όπως "security delay," "weather delay," "late aircraft delay," "Air system delay" και "Elapsed time" έχουν κενές τιμές για περισσότερο από 80% των εγγραφών. Επομένως, η καλύτερη λύση είναι να αφαιρεθούν αυτές οι μεταβλητές. Μετά την αφαίρεση των προαναφερθέντων μεταβλητών, οι κενές τιμές στις μεταβλητές αντιμετωπίζονται αφαιρώντας αυτές τις εγγραφές καθώς αφορούν λιγότερο από το 1.5% για την εκάστοτε στήλη. Μετά την αφαίρεση όλων των κενών εγγραφών, τα δεδομένα είναι καθαρά και έτοιμα για περαιτέρω διαδικασίες.

3.3.2 Μορφοποίηση χρόνων

Αρχικά, οι χρόνοι στα δεδομένα μας είναι σε μορφή 4ψήφιων αριθμών, οι οποίοι δεν είναι πολύ χρήσιμοι. Έτσι, μετατρέπονται σε μορφή ΩΩ:ΛΛ. Επομένως, δημιουργούνται νέες στήλες με τους μορφοποιημένους χρόνους για τις μεταβλητές Departure time, Scheduled arrival, Scheduled departure και arrival time.

3.3.3 Επιλογή χαρακτηριστικών

Οι μεταβλητές "Unnamed" και "flight number" έχουν μοναδικές τιμές για κάθε εγγραφή. Τέτοιου είδους στήλες δεν διδάσκουν το μοντέλο σχετικά με την καθορισμένη μεταβλητή. Αυτές οι στήλες δεν επηρεάζουν την καθορισμένη μεταβλητή, και ως εκ τούτου δεν είναι πολύ χρήσιμες για την κατασκευή προβλεπτικού μοντέλου. Επομένως, αφαιρούμε τις στήλες στο αρχικό στάδιο. Εκτός από αυτές τις στήλες, η μεταβλητή "year" έχει την ίδια τιμή για όλες τις εγγραφές, το 2015, και η μεταβλητή "country" έχει την ίδια τιμή, την "U.S.A.", καθώς τα δεδομένα αφορούν πτήσεις που πραγματοποιήθηκαν στις Ηνωμένες Πολιτείες κατά το έτος 2015. Γι' αυτό, αυτές οι δύο στήλες αφαιρούνται επίσης καθώς δεν προσθέτουν καμία αξία στο να δημιουργηθεί ένα προβλεπτικό μοντέλο.

Επίσης πρέπει να μελετήσουμε τις τρεις μεταβλητές, "scheduled departure," "departure time" και "departure delay." Η μεταβλητή "scheduled departure"

είναι ο προγραμματισμένος χρόνος αναχώρησης των πτήσεων, και η μεταβλητή "departure time" περιέχει τα δεδομένα της ώρας που αναχώρησαν οι πτήσεις. Η μεταβλητή "departure delay" είναι η διαφορά μεταξύ των μεταβλητών "scheduled departure" και "departure time." Επομένως, μπορούμε να αφαιρέσουμε τη μεταβλητή "departure time", καθώς είναι περιττή πληροφορία.

Επιπλέον, η μεταβλητή "diverted" αναπαριστά την τιμή των πτήσεων που ανακατευθύνθηκαν κατά τη λειτουργία. Στο σύνολο των δεδομένων, μόνο 128 πτήσεις ανακατευθύνθηκαν από πάνω από 50 χιλιάδες πτήσεις, που σημαίνει λιγότερο από το 0,5%. Επομένως, αυτή η μεταβλητή προσθέτει ελάχιστη αξία στο προβλεπτικό μοντέλο, καθώς οι τιμές για όλες τις εγγραφές είναι ίδιες, εκτός από το 0,5%. Έτσι, και αυτή η μεταβλητή αφαιρείται.

Τελικά, οι μεταβλητές που έμειναν για την ανάλυση είναι οι εξεις:

'MONTH' : ο μήνας εκτέλεσης της πτήσης

'DAY' : η ημέρα εκτέλεσης της πτήσης

'DAY_OF_WEEK' :

'FLIGHT_NUMBER'

'TAIL_NUMBER'

'ORIGIN_AIRPORT'

'DESTINATION_AIRPORT'

'SCHEDULED_DEPARTURE'

'DEPARTURE_DELAY'

'AIR_TIME'

'DISTANCE'

'SCHEDULED_ARRIVAL'

'CANCELLED'

'AIRLINE'

3.3.4 Data Sampling

Για τη βελτίωση της ταχύτητας του μοντέλου, επιλέχθηκε να χρησιμοποιηθεί μια μέθοδος δειγματοληψίας δεδομένων, διατηρώντας περίπου το 10% του αρχικού συνόλου. Η εν λόγω επιλογή βασίστηκε σε πολλούς παράγοντες που επηρεάζουν την αποδοτικότητα της ανάλυσης δεδομένων.

Ένας από αυτούς τους παράγοντες αφορά την πολυπλοκότητα των δεδομένων. Σε περιπτώσεις μεγάλων συνόλων δεδομένων, ο χρόνος εκτέλεσης των μοντέλων μπορεί να αυξηθεί σημαντικά, διακυβεύοντας την αποδοτικότητα της ανάλυσης. Από τη στιγμή που το μέγεθος του συνόλου δεδομένων μειώνεται, τα μοντέλα μπορούν να εκπαιδευτούν και να αξιολογηθούν πιο γρήγορα, χωρίς να υποβαθμίζεται δραστικά η ακρίβεια των αποτελεσμάτων.

Επιπλέον, η μείωση του μεγέθους των δεδομένων μπορεί να βοηθήσει στον περιορισμό της πολυπλοκότητας των μοντέλων και να μειώσει τον κίνδυνο υπερεκπαίδευσης. Το αποτέλεσμα αυτό συμβάλλει στη βελτίωση της ικανότητας γενίκευσης των μοντέλων και αυξάνει την αποδοτικότητά τους σε νέα δεδομένα.

Επίσης, η επιλογή της δειγματοληψίας οδηγεί σε μείωση των απαιτήσεων αποθήκευσης δεδομένων, επιτρέποντας την επιτάχυνση της εκτέλεσης των μοντέλων σε συστήματα με περιορισμένους πόρους.

Συνολικά, η απόφαση να χρησιμοποιηθεί δειγματοληψία για τη διατήρηση περίπου 10% των δεδομένων αποδείχθηκε αποτελεσματική, βελτιώνοντας την ταχύτητα της ανάλυσης, ενώ παράλληλα διατηρήθηκε ικανοποιητική ακρίβεια στα αποτελέσματα των μοντέλων.

3.3.5 Κωδικοποίηση Ετικετών

Ορισμένα από τα χαρακτηριστικά είναι σε μορφή συμβολοσειράς. Αυτά μετατρέπονται σε αριθμητικές τιμές χρησιμοποιώντας τον κωδικοποιητή ετικετών (Label Encoder), με τιμές που ξεκινούν από το μηδέν. Αυτό γίνεται

για να γίνει το σύνολο δεδομένων πιο φιλικό προς τη μηχανική μάθηση, καθώς τα μοντέλα συνήθως δεν αποδίδουν καλά με συμβολοσειρές ως χαρακτηριστικά.

3.3.6 Κανονικοποίηση τιμών και κλιμάκωση

Χρησιμοποιώντας τη βιβλιοθήκη της Python που ονομάζεται "StandardScaler", το σύνολο δεδομένων κλιμακώνεται έτσι ώστε να έχει μέση τιμή μηδέν και τυπική απόκλιση 1. Αυτό γίνεται για να φέρουμε όλα τα χαρακτηριστικά στην ίδια κλίμακα, καθώς οι τιμές των χαρακτηριστικών διαφέρουν σε μεγέθη, και οι αλγόριθμοι που χρησιμοποιούνται εσωτερικά χρησιμοποιούν μετρικές απόστασης που είναι ευαίσθητες σε μεγάλες διακυμάνσεις στις τιμές.

3.3.7 Δημιουργία νέου χαρακτηριστικού για την ταξινόμηση

Δημιουργήθηκε ένα νέο χαρακτηριστικό με δυαδική τιμή 0 και 1 για την εκτέλεση του μοντέλου ταξινόμησης. Αυτό το χαρακτηριστικό δημιουργήθηκε με βάση το χρόνο καθυστέρησης. Εάν ήταν μεγαλύτερος από 0, η τιμή ορίστηκε σε 1, αλλιώς ορίστηκε σε 0.

3.4 Μοντέλα που χρησιμοποιήθηκαν

Χρησιμοποιήθηκαν πολλά μοντέλα τόσο για παλινδρόμηση όσο και για ταξινόμηση. Τα μοντέλα που χρησιμοποιήθηκαν για παλινδρόμηση και ταξινόμηση είναι τα ακόλουθα:

3.4.1 Παλινδρόμηση:

3.4.1.1 Απλή γραμμική παλινδρόμηση:

Αυτή η προσέγγιση προσπαθεί να βρει μια γραμμική σχέση μεταξύ της τιμής που πρόκειται να προβλεφθεί και των χαρακτηριστικών που χρησιμοποιούνται για την πρόβλεψη. Αυτό είναι ένα από τους πιο απλούς αλγόριθμους μηχανικής μάθησης που λειτουργεί προσπαθώντας να βρει μια συνάρτηση πρόβλεψης μιας μεταβλητής χρησιμοποιώντας άλλες μεταβλητές, προσφέροντας προϋποθέσεις για την ύπαρξη αιτιακής σχέσης μεταξύ τους. Η βασική έννοια της γραμμικής παλινδρόμησης μπορεί να εκφραστεί από τον παρακάτω τύπο όπου τα f αντιπροσωπεύουν τα χαρακτηριστικά που χρησιμοποιούμε για την πρόβλεψη, Δ αντιστοιχεί στα βάρη κάθε ενός από αυτά και $\in$ είναι μια αυθαίρετη σταθερά (Montgomery, Peck, & Vining, 2021).

$$Y_{pred} = \Delta_1 f_1 + \Delta_2 f_2 + \Delta_3 f_3 + \cdots + \Delta_n f_n + \in$$

3.4.1.2 Παλινδρόμηση με τυχαίο δάσος (Random Forest Regression):

Αυτή είναι μια τεχνική συνόλου (ensemble) που μπορεί να χρησιμοποιηθεί τόσο για προβλήματα παλινδρόμησης όσο και για προβλήματα ταξινόμησης. Δημιουργεί πολλά δέντρα αποφάσεων χρησιμοποιώντας μια τεχνική που ονομάζεται "bagging". Το bagging περιλαμβάνει την εκπαίδευση όλων των δέντρων αποφάσεων σε διαφορετικά δείγματα δεδομένων. Η τελική πρόβλεψη γίνεται συνδυάζοντας τα αποτελέσματα όλων των δέντρων αποφάσεων αντί να εξαρτόμαστε αποκλειστικά από ένα από αυτά (Segal, 2004).

3.4.1.3 Ενισχυμένη γραμμική παλινδρόμηση (Boosted Linear Regression)

Αυτή είναι μια μέθοδος συνόλου μηχανικής μάθησης που συνδυάζει πολλά αδύναμα μοντέλα σε ένα. Κατασκευάζει το μοντέλο σταδιακά, με την έννοια ότι το επόμενο καλύτερο μοντέλο όταν συνδυαστεί με όλα τα προηγούμενα μοντέλα θα ελαχιστοποιήσει τα συνολικά σφάλματα πρόβλεψης (Ilic, Gorgulu, Cevik, & Baydogan, 2021).

3.4.1.4 XGBoost

Το XGBoost (Extreme Gradient Boosting) είναι ένα ισχυρό εργαλείο της μηχανικής μάθησης που χρησιμοποιείται τόσο για προβλήματα παλινδρόμησης όσο και ταξινόμησης. Αποτελεί μια προηγμένη μέθοδο ενισχυτικής μάθησης, βασισμένη στην αρχή της ενίσχυσης των δέντρων απόφασης.

Η βασική του ιδέα είναι να δημιουργήσει μια σειρά από αδύναμα μοντέλα (συνήθως δέντρα απόφασης) και να τα συνδυάσει σε ένα ισχυρό μοντέλο, αποφεύγοντας την υπερεκπαίδευση και βελτιστοποιώντας την απόδοση.

Το XGBoost χρησιμοποιεί την ιδέα της ενίσχυσης των κλάσεων, δημιουργώντας μοντέλα που επικεντρώνονται στα σφάλματα που προκλήθηκαν από τα προηγούμενα μοντέλα. Αυτό οδηγεί σε βελτιωμένες προβλέψεις καθώς το μοντέλο συνεχίζει να μαθαίνει από τα λάθη του.

Οι προηγμένες δυνατότητες του XGBoost περιλαμβάνουν αυτόματη ρύθμιση παραμέτρων, υποστήριξη για πολλαπλά μοντέλα ταυτόχρονα και εξαιρετική απόδοση σε μεγάλα και πολύπλοκα σύνολα δεδομένων (Chen, et al., 2015).

3.4.2 Ταξινόμηση

3.4.2.1 Ταξινομητής k-Πλησιέστερου Γείτονα (K Neighbors Classifier)

Αυτός ο αλγόριθμος λειτουργεί αρχικά υπολογίζοντας τους k πλησιέστερους γείτονες της τιμής που χρειάζεται να προβλεφθεί και με βάση αυτούς τους γείτονες αναθέτει μια τιμή κλάσης. Ο αλγόριθμος K-πλησιέστερων γειτόνων λειτουργεί ως εξής (Laaksonen & Oja, 1996):

- Βρίσκει την τιμή k, που είναι ο αριθμός των πλησιέστερων γειτόνων
- Υπολογίζει τις αποστάσεις μεταξύ των δεδομένων προς πρόβλεψη και των δεδομένων εκπαίδευσης και τις ταξινομεί.
- Ελέγχει τις ετικέτες των k πλησιέστερων γειτόνων και αναθέτει την τιμή της μεγαλύτερης κλάσης ως πρόβλεψη.

3.4.2.2 Λογιστική παλινδρόμηση (Logistic Regression)

Αυτός ο αλγόριθμος χρησιμοποιεί ένα λογαριθμικό μοντέλο που έχει μια καμπύλη 'S' για να προβλέψει τιμές. Αυτό το μοντέλο χρησιμοποιείται όταν υπάρχουν 2 κλάσεις εξόδου όπως αληθές ή ψευδές. Λειτουργεί υπολογίζοντας πιθανότητες και μετατρέποντας τις σε μια συνάρτηση (Wright, 1995). Χρησιμοποιεί την εξίσωση υπόθεσης:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \quad \text{όπου } \theta^T x = \theta_0 + \sum_{j=1}^{n} \theta_j - x_j$$

Μπορούμε να βρούμε την παράμετρο θ χρησιμοποιώντας τη μέθοδο ανόδου της κλίσης (gradient ascent) και τον εκτιμητή της μέγιστης πιθανότητας, όπως παρακάτω:

$$\theta := \theta + a\nabla_\theta 1(\theta)$$

$$1(\theta) = \sum_{i=1}^{m} \log\left(h\left(x^i\right)\right) + \left(1 - y^i\right)\log(1 - hx^i)$$

3.4.2.3 Δέντρα αποφάσεων:

Η βασική ιδέα αυτού του αλγορίθμου είναι να δημιουργήσει μια δενδρικό δομή με κάθε κόμβο να περιέχει μια επιλογή για να πάει προς το κλαδί αριστερά ή δεξιά. Σε κάθε επίπεδο, ο κόμβος θέτει μια απλή ερώτηση στην οποία η απάντηση είναι είτε αληθής είτε ψευδής, και βάσει αυτού το δεδομένο διαχωρίζεται σε 2 υποσύνολα. Ο στόχος αυτού του αλγορίθμου είναι να συνεχίσει να κάνει ερωτήσεις και να χτίζει το δέντρο μέχρι να μπορέσει να φτάσει στα πιο καθαρά δυνατά διαχωρισμένα υποσύνολα. Για να αξιολογήσει την ακαθαρσία σε κάθε επίπεδο, τα δέντρα αποφάσεων χρησιμοποιούν μετρικές όπως η Εντροπία ή η τιμή Gini για να ποσοτικοποιήσουν τις ακαθαρσίες. Συνήθως η διαδικασία εισαγωγής είναι πολύ αργή, αλλά η εξαγωγή πληροφορίας είναι πολύ γρήγορη, καθώς απλά χρειάζεται να διασχίσει το δημιουργημένο δέντρο και να φτάσει στο φύλλο (De Ville, 2013).

3.5 Μετρικές Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν διάφορες μετρικές ανάλογα την κάθε περίπτωση.

3.5.1 Παλινδρόμηση

Οι ακόλουθες μετρικές χρησιμοποιούνται για να αξιολογήσουν την απόδοση των μοντέλων που χρησιμοποιούνται για παλινδρόμηση:

- Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE) - Αυτό μας δίνει τις διακυμάνσεις μεταξύ των αναμενόμενων και των πραγματικών τιμών των προβλέψεων.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_i - x_i|$$

όπου y_i είναι η προβλεπόμενη τιμή και x_i είναι η πραγματική τιμή για την i-οστή παράμετρο.

- Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE) - Αυτό μετρά το μέσο του αθροίσματος των τετραγώνων των σφαλμάτων.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_i - x_i)^2$$

όπου y_i είναι η προβλεπόμενη τιμή και x_i είναι η πραγματική τιμή για την i-οστή παράμετρο.

- Τετραγωνική Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error - RMSE) - Αυτή είναι απλώς η τετραγωνική ρίζα του MSE, και χρησιμοποιείται συχνά ως μέτρο διαφοράς αντί του MSE, καθώς οι μονάδες τελικά καταλήγουν να είναι ίδιες με τις αρχικές.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_i - x_i)^2}$$

όπου y_i είναι η προβλεπόμενη τιμή και x_i είναι η πραγματική τιμή για την i-οστή παράμετρο.

- R^2 (Συντελεστής Προσδιορισμού) - Είναι μια στατιστική μέτρηση του πόσο κοντά είναι οι πραγματικές τιμές στην εφαρμοσμένη γραμμή παλινδρόμησης.

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

όπου y_i είναι η πραγματική τιμή, $\hat{y}_i$ είναι η προβλεπόμενη τιμή και $\bar{y}_i$ είναι ο μέσος όρος για την i-οστή παράμετρος

3.5.2 Classification

Οι παρακάτω μετρικές χρησιμοποιούνται για να αξιολογήσουν την απόδοση των μοντέλων που χρησιμοποιούνται για ταξινόμηση:

- Σκορ (Score): Αυτή είναι η προεπιλεγμένη μέθοδος αξιολόγησης που έχει ενσωματωθεί σε κάθε έναν από τους ταξινομητές.
- Ακρίβεια (Precision): Αυτή μας λέει πόσες από τις τιμές που ο ταξινομητής προέβλεψε ως αληθείς, πραγματικά ήταν αληθείς.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- Ανάκληση (Recall): Η ανάκληση μας λέει πόσες από τις πραγματικά αληθείς τιμές κατάφερε ο ταξινομητής να ανακαλέσει σωστά από αυτές που έχει μάθει.

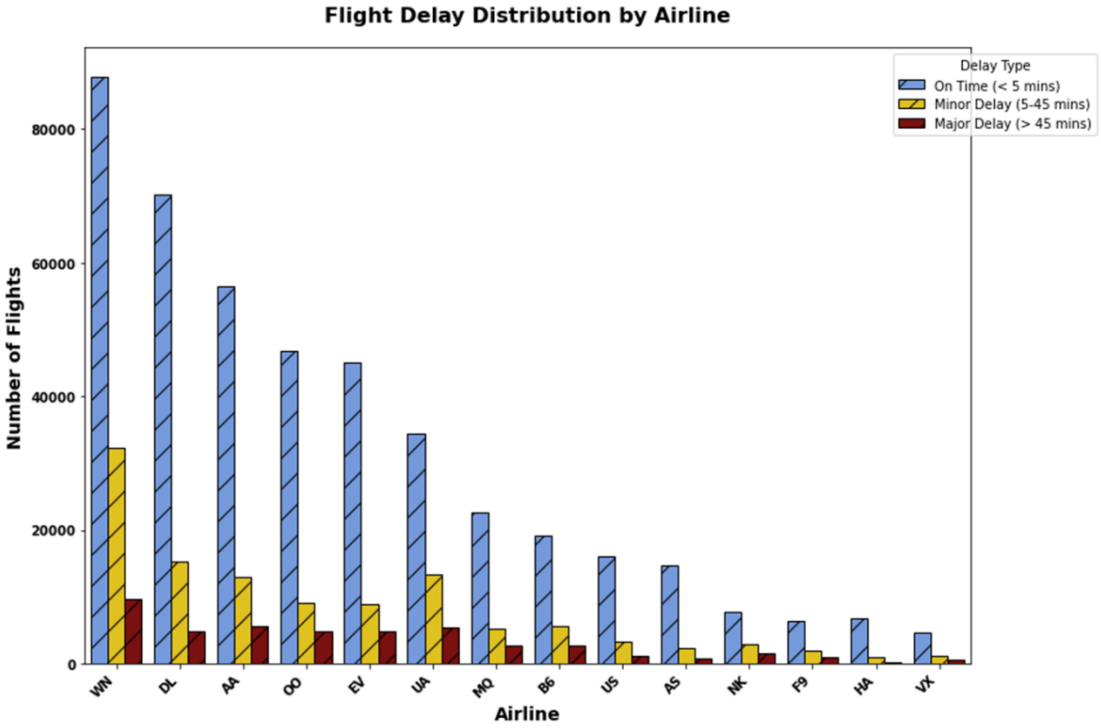
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- Σκορ F1 (F1 Score): Αυτό είναι το αρμονικό μέσο της ακρίβειας και της ανάκλησης.

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

3.6 Αποτελέσματα και Παρουσίαση της Ανάλυσης

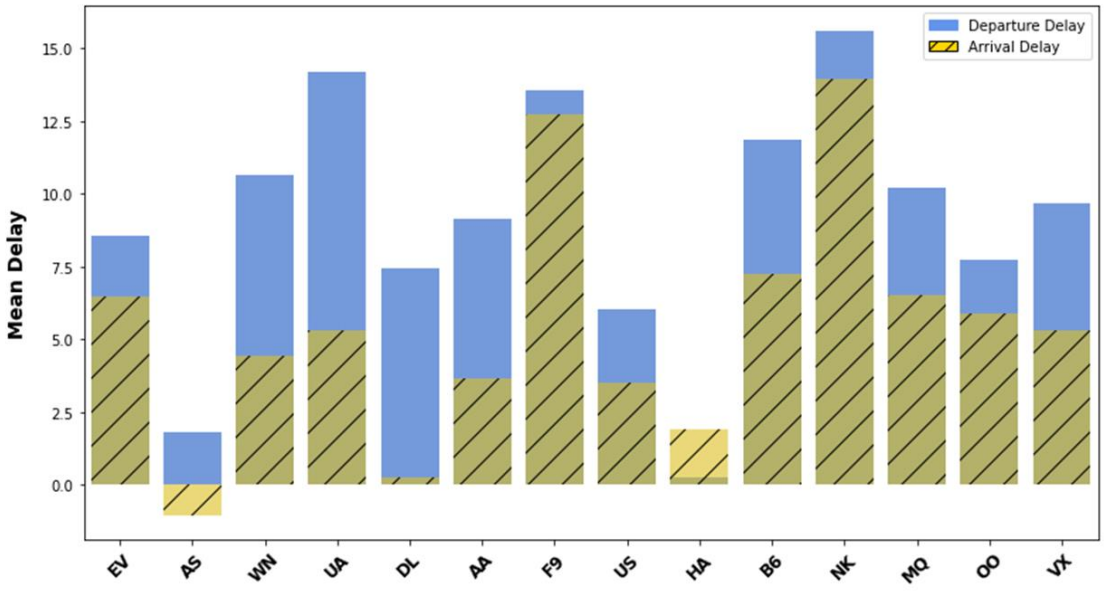
Παρακάτω παρουσιάζεται ένα διάγραμμα που αποτυπώνει τους αερομεταφορείς και τον αριθμό των πτήσεων χωρίζοντάς τες σε αυτές που δεν είχαν καμία σημαντική καθυστέρηση (καθυστέρηση έως 5 λεπτά), αυτές που είχαν μικρή καθυστέρηση (καθυστέρηση από 5 εως 45 λεπτά) και αυτές που παρουσίασαν σημαντική καθυστέρηση (μεγαλύτερη των 45 λεπτών).



Γράφημα 5 Καθυστέρηση πτήσεων

Όπως φαίνεται παραπάνω, ορισμένες αεροπορικές έχουν μεγαλύτερο ποσοστό πτήσεων που παρουσίασαν σημαντική καθυστέρηση συγκριτικά με τον συνολικό αριθμό πτήσεων.

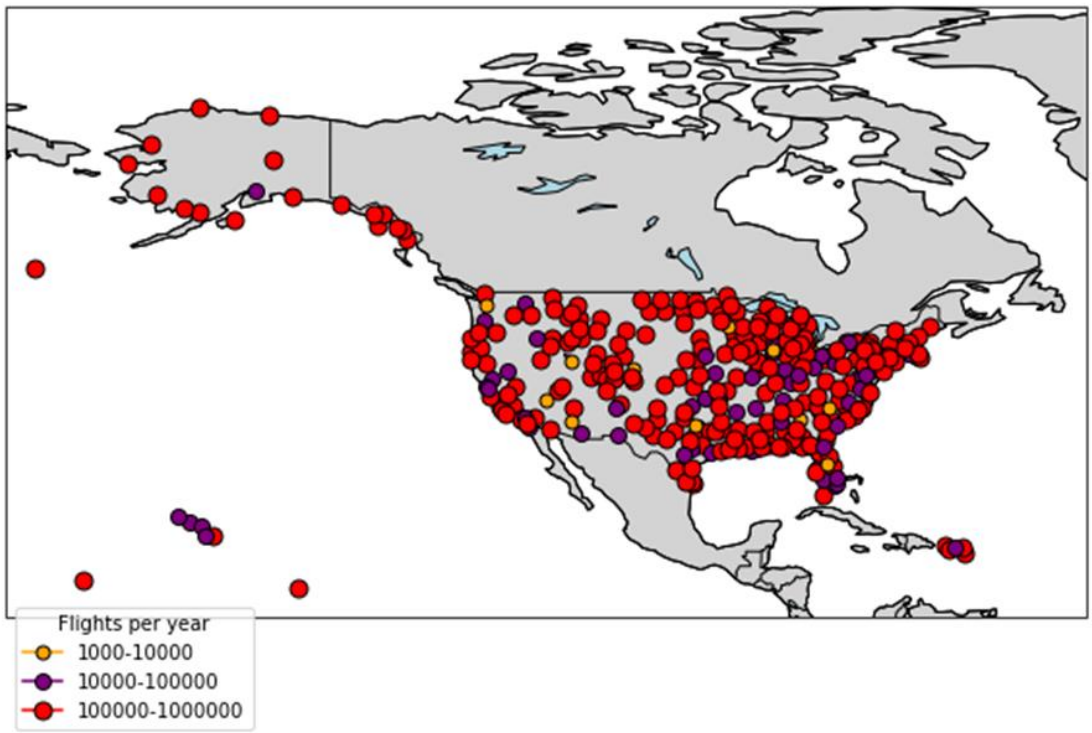
Στο επόμενο γράφημα παρουσιάζονται οι μέσες τιμές των καθυστερήσεων ανά αεροπορική για τις αφίξεις και τις αναχωρήσεις.



Γράφημα 6 Μέση τιμή καθυστερήσεων (Αναχωρήσεις και Αφίξεις)

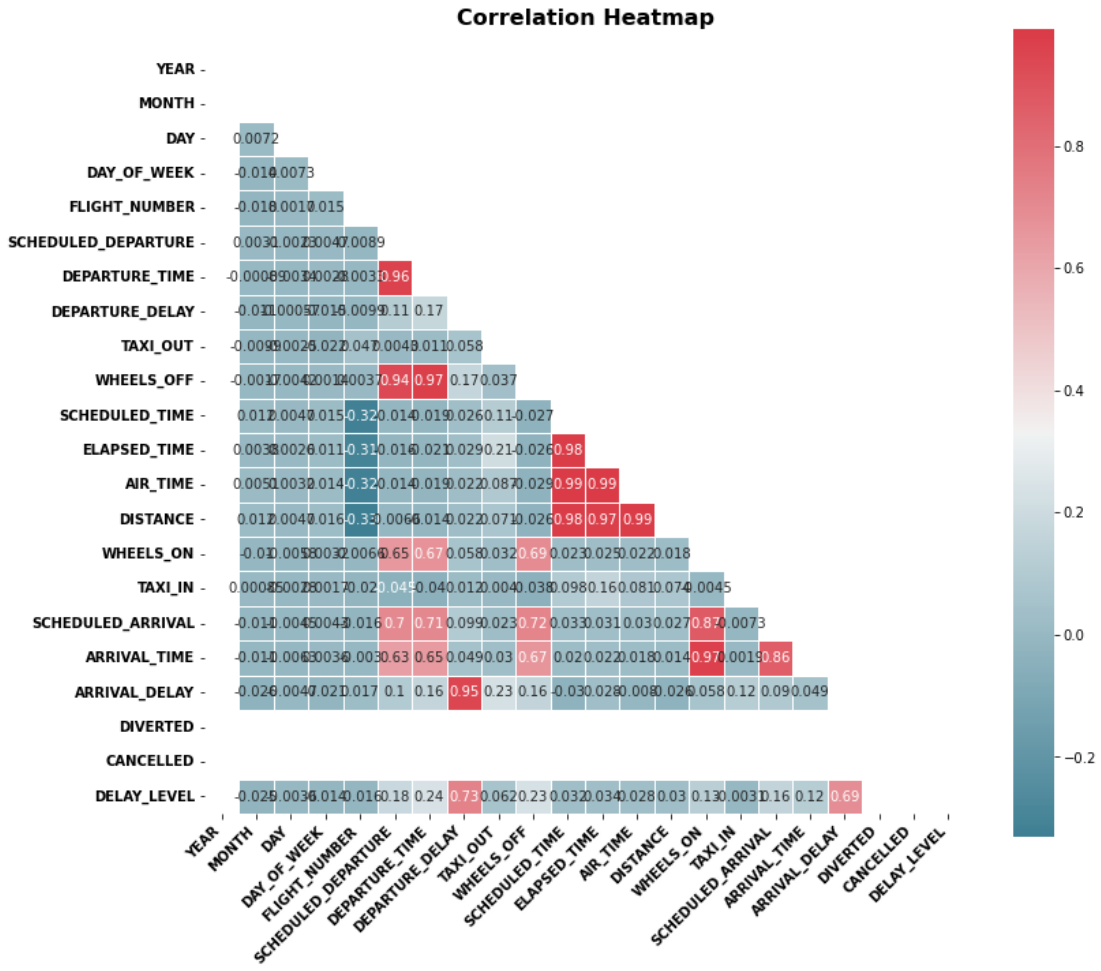
Ομοίως, ορισμένες αεροπορικές έχουν σημαντικά μεγάλο χρόνο καθυστέρησης τόσο για τις αφίξεις όσο και για τις αναχωρήσεις σε σύγκριση με τις υπόλοιπες αεροπορικές εταιρείες.

Παρακάτω παρατίθεται ένας χάρτης με την πυκνότητα των πτήσεων ανά περιοχή.



Γράφημα 7 Πυκνότητα Πτήσεων

Τέλος, παρουσιάζεται ο πίνακας συσχέτισης.



Γράφημα 8 Πίνακας Συσχέτισης

Όπως φαίνεται υπάρχει δυνατή συσχέτιση μεταξύ του ARRIVAL_DELAY και του DEPARTURE_DELAY.

3.6.1 Τεχνικές παλινδρόμησης

Έχουν διεξαχθεί τρία σετ εκτελέσεων με διαφορετικά μεγέθη δεδομένων και τεχνικές μεντολοποίησης. Θα αναλύσουμε τα αποτελέσματα κάθε βήματος, θα συζητήσουμε τις τάσεις και θα παρέχουμε μια συνολική επισκόπηση της μεθόδου που φαίνεται να είναι η καλύτερη.

3.6.1.1 Βήμα 1: Αρχική ανάλυση πλήρους συνόλου δεδομένων

Σε αυτό το βήμα, χρησιμοποιήσαμε ολόκληρο το σύνολο δεδομένων και λάβαμε τα ακόλουθα αποτελέσματα παλινδρόμησης:

Metric	Linear Regression	Random Forest Regressor	Decision Tree Regression	ADA Boost Regression	XG Boost
MAE	8.50	-	9.23	12.67	7.34
MSE	145.92	-	195.34	280.97	114.76
RMSE	12.08	-	13.98	16.76	10.71
R2	0.91	-	0.87	0.82	0.93

Πίνακας 3 Αποτελέσματα Παλινδρόμησης (Ολόκληρο Dataset)

3.6.1.2 Βήμα 2: Ανάλυση σε δειγματοληπτικό σύνολο δεδομένων

Το μέγεθος δεδομένων μειώθηκε σε περίπου 10% για να επιταχυνθούν οι υπολογισμοί. Παρακάτω τα αποτελέσματα αφού έτρεξαν ξανά όλα τα μοντέλα στο νέο περιορισμένο σύνολο δεδομένων:

Metric	Linear Regression	Random Forest Regressor	Decision Tree Regression	ADA Boost Regression	XG Boost
MAE	8.52	-	11.57	9.42	7.34
MSE	147.50	-	273.51	166.68	114.76
RMSE	12.15	-	16.54	12.91	10.71
R2	0.91	-	0.83	0.90	0.93

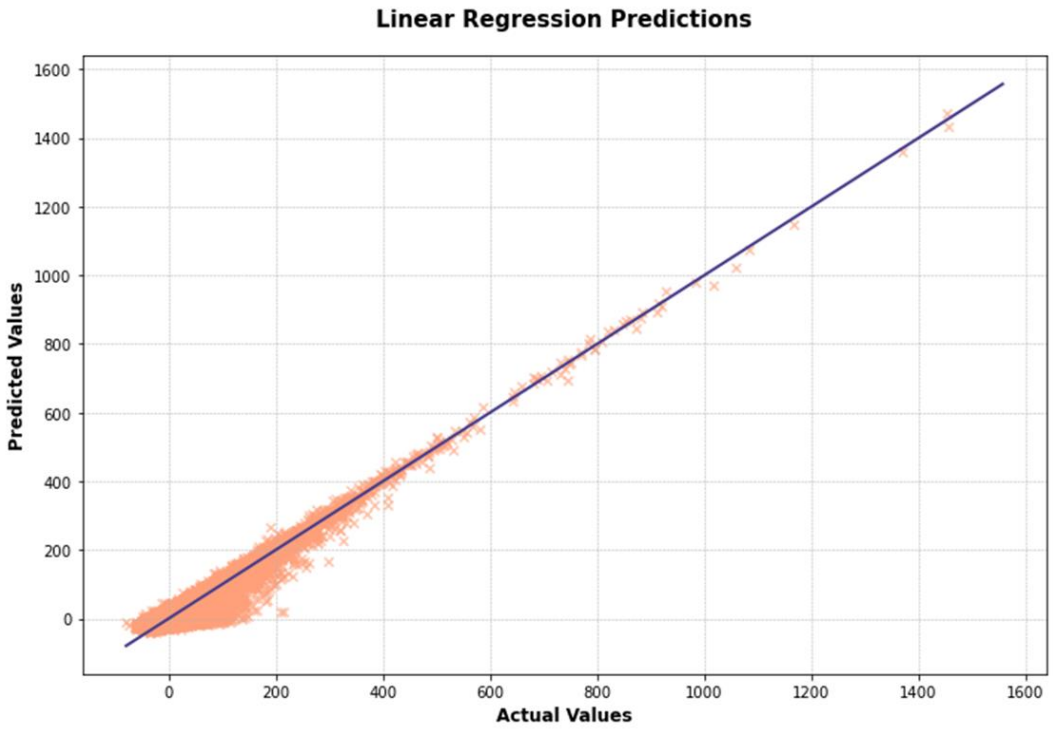
Πίνακας 4 Αποτελέσματα Παλινδρόμησης (Μειωμένο Dataset)

3.6.1.3 Βήμα 3: Βελτιστοποιημένα μοντέλα σε δειγματοληπτικό σύνολο δεδομένων

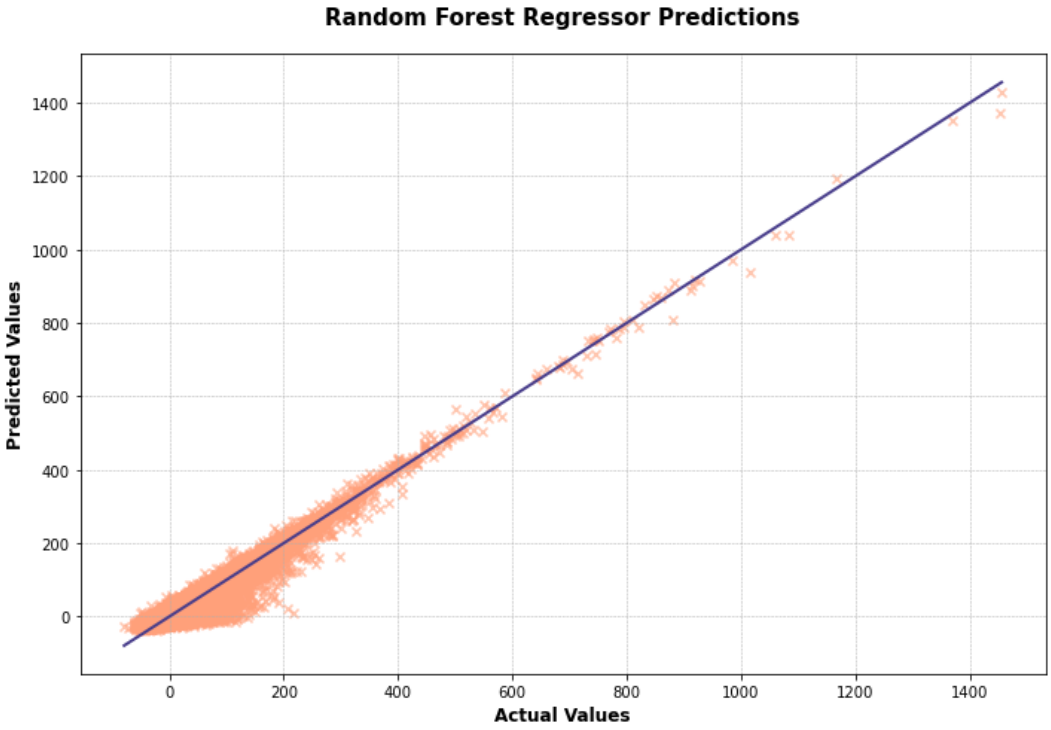
Τα μοντέλα βελτιστοποιήθηκαν στο σύνολο δεδομένων του δείγματος αλλάζοντας ορισμένες από τις παραμέτρους και τα αποτελέσματα που έφεραν είναι τα παρακάτω:

Metric	Linear Regression	Random Forest Regressor	Decision Tree Regression	ADA Boost Regression	XG Boost
MAE	8.16	7.83	8.95	8.52	6.80
MSE	135.59	129.58	162.97	147.50	101.68
RMSE	11.64	11.38	12.77	12.15	10.08
R2	0.92	0.92	0.90	0.91	0.94

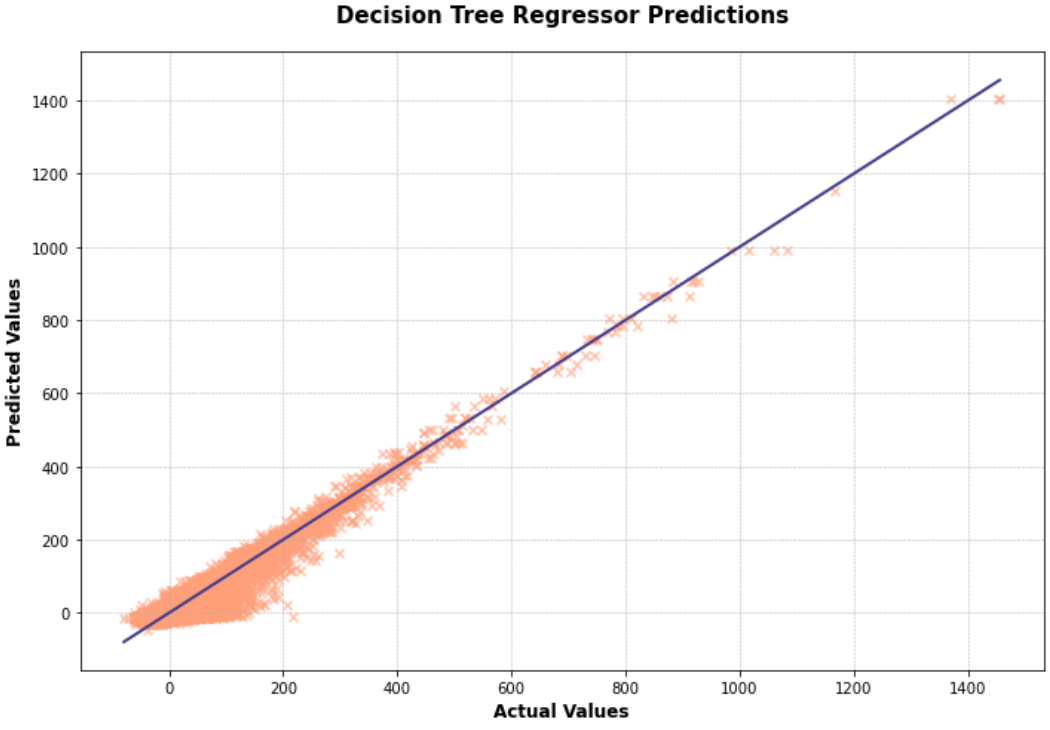
Πίνακας 5 Αποτελέσματα Παλινδρόμησης (Βελτιστοποιημένα σε μειωμένο Dataset)



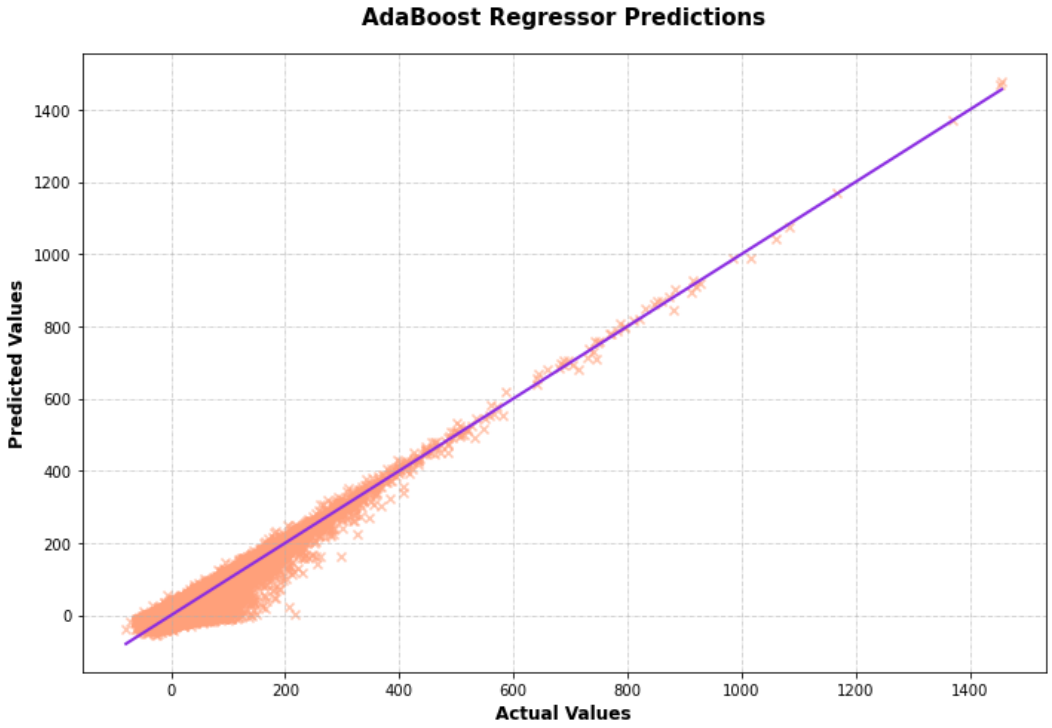
Γράφημα 9 Lineear Regression



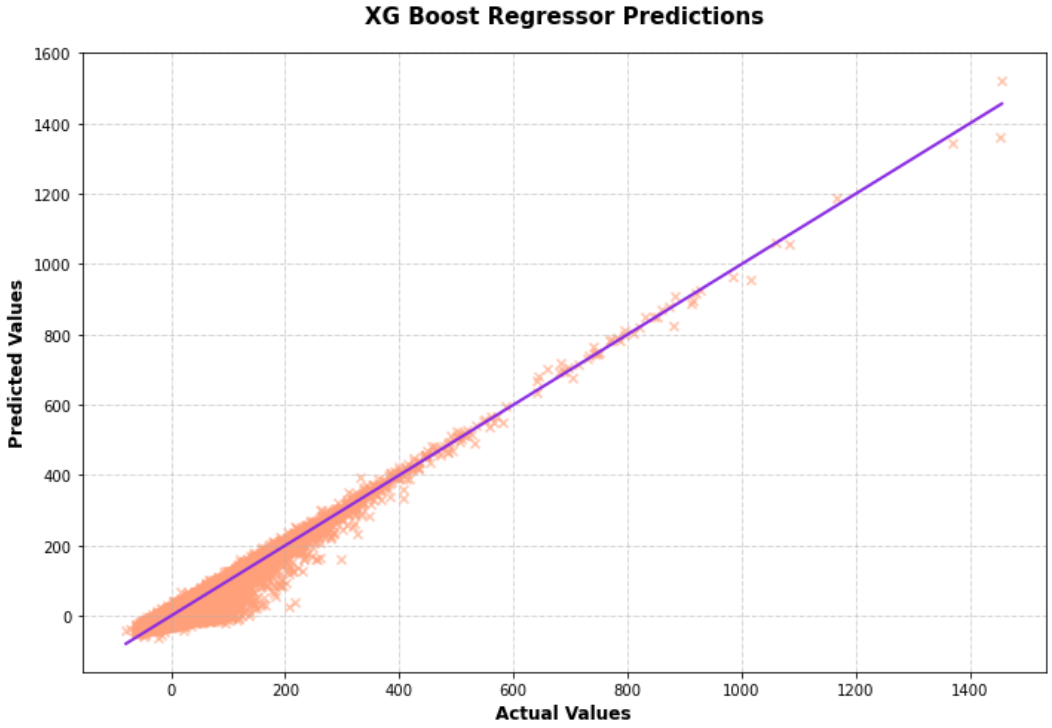
Γράφημα 10 Random Forest



Γράφημα 11 Decision Tree



Γράφημα 12 AdaBoost



Γράφημα 13 XGBoost

3.6.1.4 Συνολική σύγκριση και ανάλυση:

Η ανάλυση απεικονίζει μια σαφή εξέλιξη στην απόδοση του μοντέλου σε μικρότερα μεγέθη δεδομένων και επαναλήψεις βελτιστοποίησης. Είναι αξιοσημείωτο ότι το μοντέλο παλινδρόμησης XG Boost αναδεικνύεται σταθερά ως η κορυφαία απόδοση, παρουσιάζοντας ανώτερες τιμές R2, χαμηλότερο RMSE και μειωμένο MAE σε όλα τα στάδια. Αυτό υπογραμμίζει την ικανότητα του XG Boost να καταγράφει περίπλοκες σχέσεις που είναι εγγενείς στα δεδομένα.

Τα ευρήματα υπογραμμίζουν την επαναληπτική φύση της ανάπτυξης του μοντέλου, τονίζοντας την περίπλοκη αλληλεπίδραση μεταξύ του μεγέθους των δεδομένων, της πολυπλοκότητας του μοντέλου και της απόδοσης.

Βελτίωση μοντέλου: Έχει δείξει σαφή βελτίωση σε όλα τα μοντέλα καθώς προχωρούσε από την αρχική ανάλυση πλήρους συνόλου δεδομένων στα βελτιστοποιημένα μοντέλα στο σύνολο δεδομένων του δείγματος. Αυτή η βελτίωση είναι εμφανής στις μετρήσεις R2, RMSE και MAE.

Μοντέλα με τις καλύτερες επιδόσεις: Μεταξύ των μοντέλων που εξετάστηκαν, το μοντέλο XG Boost ξεχωρίζει σταθερά ως το κορυφαίο σε όλα τα βήματα, με τις υψηλότερες τιμές R2 και γενικά τις χαμηλότερες τιμές RMSE και MAE. Αυτό υποδηλώνει ότι το XG Boost καταγράφει αποτελεσματικά πολύπλοκες σχέσεις στα δεδομένα.

Αντίκτυπος συνόλου δεδομένων δειγμάτων: Ενώ η εργασία με μικρότερο σύνολο δεδομένων επιταχύνει τους υπολογισμούς, είναι σημαντικό να σημειωθεί ότι η απόδοση του μοντέλου μπορεί να ποικίλλει ανάλογα με το μέγεθος του συνόλου δεδομένων. Ορισμένα μοντέλα ενδέχεται να μην έχουν εξίσου καλή απόδοση σε μικρότερα σύνολα δεδομένων λόγω μειωμένης ποικιλομορφίας και πολυπλοκότητας δεδομένων.

Βελτιστοποίηση μοντέλων: Η βελτιστοποίηση μοντέλων φαίνεται να οδηγεί σε βελτιωμένη απόδοση. Ωστόσο, απαιτείται προσεκτικός συντονισμός, όπως φαίνεται στα βελτιωμένα αποτελέσματα της Γραμμικής παλινδρόμησης και της ADA Boost Regression στο Βήμα 3.

Τελική σύσταση: Με βάση τα παρεχόμενα αποτελέσματα, το μοντέλο XG Boost παρουσιάζει σταθερά την καλύτερη συνολική απόδοση όσον αφορά τόσο την ακρίβεια παλινδρόμησης όσο και την ικανότητα πρόβλεψης. Επομένως, για το μοντέλο πρόβλεψης αναλυτικών στοιχείων, η χρήση του XG Boost με τις βελτιστοποιημένες παραμέτρους θα ήταν μια ισχυρή σύσταση.

3.6.2 Τεχνικές Ταξινόμησης

Εξετάσθηκαν πέντε διαφορετικά μοντέλα ταξινόμησης και αναλύθηκαν οι μετρικές απόδοσης προκειμένου να αξιολογηθεί η καταλληλότητά τους για το εν λόγω πρόβλημα.

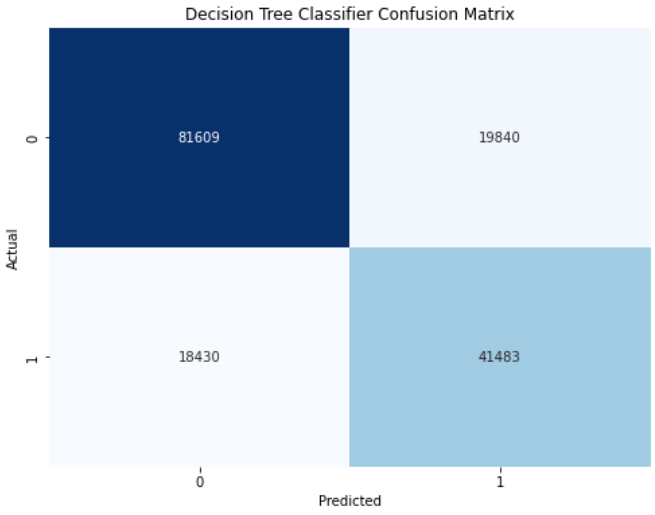
Αποτελέσματα Μοντέλων Ταξινόμησης:

Metric	Decision Tree	K-nearest Neighbour	Logistic Regression Classifier	Naïve Bayes Classifier	XG Boost
Accuracy	0.76	0.75	0.83	0.82	0.84
Precision	0.75	0.74	0.82	0.84	0.84
Recall	0.75	0.71	0.80	0.77	0.80
F1 Score	0.75	0.71	0.81	0.78	0.81

Πίνακας 6 Αποτελέσματα Ταξινόμησης

3.6.2.1 Δέντρο Αποφάσεων:

- Ορθότητα: 0.763
- Ακρίβεια: 0.746
- Ανάκληση: 0.748
- Σκορ F1: 0.747

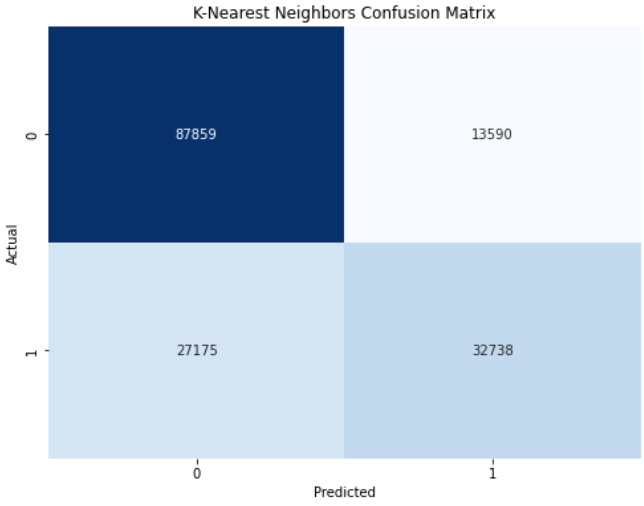


Γράφημα 14 Decision Tree

Το μοντέλο Decision Tree παρουσιάζει ικανοποιητική ακρίβεια, με εξισορροπημένες τιμές ορθότητας και ανάκλησης. Οι τιμές αυτές υποδεικνύουν την ικανότητα του μοντέλου να ταξινομεί σωστά τις περιπτώσεις καθυστερήσεων.

3.6.2.2 Κ-Κοντινότερος Γείτονας:

- Ορθότητα: 0.747
- Ακρίβεια: 0.735
- Ανάκληση: 0.706
- Σκορ F1: 0.714

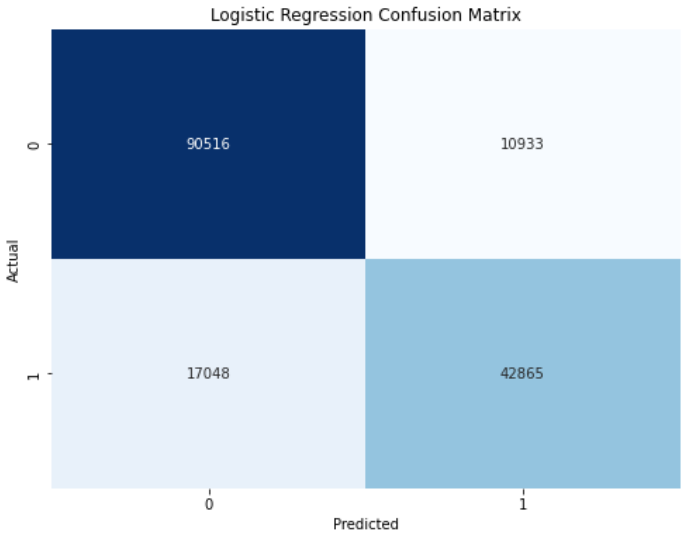


Γράφημα 15 K-Nearest Neighbors

Το μοντέλο Κ-Κοντινότερου Γείτονα εμφανίζει λογική ακρίβεια, αλλά ενδέχεται να έχει δυσκολία στην ανάκληση και το σκορ F1. Οι τιμές αυτές μπορούν να υποδείξουν ένα ενδεχόμενο πρόβλημα στην ταξινόμηση των περιπτώσεων.

3.6.2.3 Λογιστική Παλινδρόμηση:

- Ορθότητα: 0.827
- Ακρίβεια: 0.819
- Ανάκληση: 0.804
- Σκορ F1: 0.810

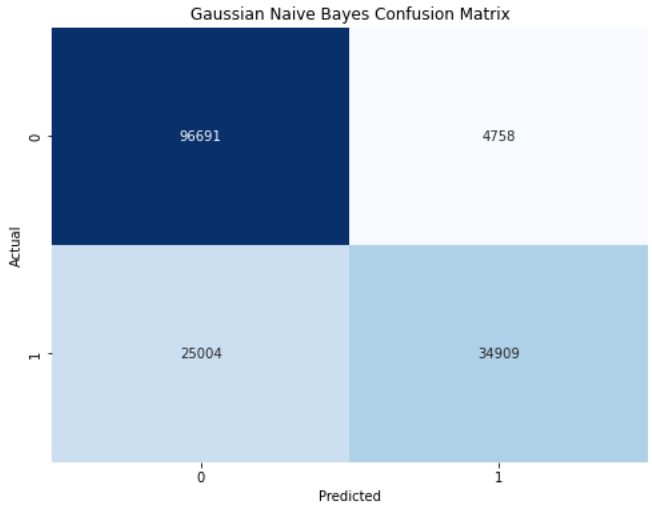


Γράφημα 16 Logistic Regression

Ο Ταξινομητής Λογιστικής Παλινδρόμησης παρουσιάζει υψηλή ακρίβεια και εξισορροπημένες τιμές ορθότητας και ανάκλησης. Οι τιμές αυτές υποδεικνύουν την ικανότητα του μοντέλου να αναγνωρίζει και τις δύο κλάσεις με αξιοπιστία.

3.6.2.4 Naïve Bayes Classifier:

- Ορθότητα: 0.816
- Ακρίβεια: 0.837
- Ανάκληση: 0.768
- Σκορ F1: 0.784

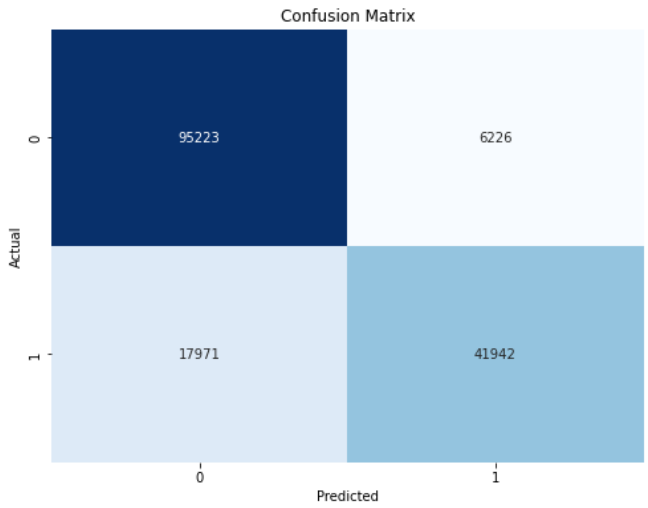


Γράφημα 17 Gaussian Naive Bayes

Ο Ταξινομητής Αφελούς Βάιες παρουσιάζει αξιοπιστία στην ακρίβεια, ενώ η ανάκληση μπορεί να υποδεικνύει ένα μικρό πρόβλημα στην αναγνώριση των αρνητικών περιπτώσεων.

3.6.2.5 XGBoost

- Ορθότητα: 0.836
- Ακρίβεια: 0.841
- Ανάκληση: 0.803
- Σκορ F1: 0.815



Γράφημα 18 XGBoost

Το μοντέλο XGBoost παρουσιάζει υψηλή ακρίβεια, εξαιρετική προσέγγιση και αξιοπιστία στην ανάκληση. Οι τιμές αυτές επιβεβαιώνουν την απόδοση του μοντέλου στην κατηγοριοποίηση των περιπτώσεων.

Συμπεράσματα: Βάσει των αποτελεσμάτων της παρούσας ανάλυσης, το μοντέλο **XGBoost** καταδεικνύει την καλύτερη συνολική απόδοση με βάση τις μετρικές αξιολόγησης. Το μοντέλο αυτό προσφέρει υψηλή ακρίβεια, προσέγγιση και ανάκληση, επιβεβαιώνοντας την αποτελεσματικότητά του στην πρόβλεψη καθυστερήσεων στις πτήσεις. Οι λοιποί ταξινομητές επίσης παρουσιάζουν αξιοπιστία και θα μπορούσαν να χρησιμοποιηθούν για το εν λόγω πρόβλημα, αν και το XGBoost ξεχωρίζει ως το πλέον αποδοτικό μοντέλο.

4 Συμπεράσματα

Στη σύγχρονη εποχή της τεχνολογίας και της παγκοσμιοποίησης, η αεροπορική βιομηχανία είναι ζωτικής σημασίας για την παγκόσμια οικονομία. Η αεροπορία προσφέρει ταχύτητα και εμβέλεια, αλλάζοντας τον τρόπο που αντιλαμβανόμαστε το χρόνο και τον χώρο. Πολλές φορές όμως, για ποικίλους λόγους οι πτήσεις παρουσιάζουν καθυστερήσεις, με αποτέλεσμα αυτό να διαταράσσει τόσο την εφοδιαστική αλυσίδα όσο και τους επιβάτες.

Για το λόγο αυτό, η χρήση μεθόδων για την καταπολέμηση και πρόγνωση των καθυστερήσεων αυτών είναι σημαντική. Κατά καιρούς αναπτύσσονται μοντέλα Μηχανικής Μάθησης για το σκοπό αυτό αλλά λόγω της πολυπλοκότητας, δεν παρουσιάζουν ιδιαίτερα ενθαρρυντικά αποτελέσματα.

Στην παρούσα εργασία επιχειρήθηκε να αναπτυχθεί ένα μοντέλο για την πρόβλεψη των καθυστερήσεων τόσο με χρήση μοντέλων παλινδρόμησης όσο και με μοντέλα ταξινόμησης. Μπόρεσε να προβλεφθεί με επιτυχία σε ένα πολύ μεγάλο βαθμό η καθυστέρηση για τις αφίξεις, αλλά αυτό γινόταν με τη χρήση μιας μεταβλητής που παρουσίαζε ιδιαίτερα υψηλή συσχέτιση, αυτή των καθυστερήσεων στην αναχώρηση. Χωρίς τη χρήση αυτής της μεταβλητής φαινόταν πρακτικά αδύνατο να υπάρξει η οποιαδήποτε σωστή πρόβλεψη και για αυτό το λόγο δεν συμπεριλήφθηκε στην παρούσα εργασία.

Στο μέλλον, για να μπορεί να δημιουργηθεί ένα πιο συμπαγές μοντέλο, ίσως θα πρέπει να καταγραφούν περισσότερες μεταβλητές, όπως ο καιρός, ο αριθμός επιβατών κ.α. για να μπορέσουν οι μελλοντικοί ερευνητές να έχουν περισσότερα εργαλεία στην πρόβλεψη των καθυστερήσεων.

5 Βιβλιογραφία

Asfe, M., Zehi, M., Tash, M., & Yaghoubi, N. (2014). Ranking different factors influencing flight delay. *Management Science Letters*, 1397-1400.

Badea, V., Zamfiroiu, A., & Boncea, R. (2018). Big data in the aerospace industry. *Informatica Economica*, 17-24.

Bisong, E. (2019). Introduction to Scikit-learn. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 215-229.

Brosky, S., & Unterberger, C. (2019). Bad weather and flight delays: The impact of sudden and slow onset weather events. *Economics of Transportation*, 10-26.

Burkov, A. (2019). *The Hundred-page Machine Learning Book.* Andriy Burkov.

Cavusoglou, S., & Macario, R. (2021). Minimum delay or maximum efficiency? Rising productivity of available capacity at airports: Review of current practice and future needs. *Journal of Air Transport Management*, 90.

Chen, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Zhou, T. (2015). Xgboost: extreme gradient boosting. *package version*, 1-4.

Chen, Z., & Wange, Y. (2019). Impacts of severe weather events on high-speed rail and aviation delays. *Transportation research part D: transport and environment*, 168-183.

De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 448-455.

Demir, E., & Demir, V. (2017). Predicting flight delays with artificial neural networks: case study of an airport. *25th Signal Processing and Communications Applications Conference (SIU)* (σσ. 1-4). IEEE.

Dhanawade, R., Deo, M., Khanna, N., & Deolekar, R. (2019). Analyzing factors influencing flight delay prediction. *2019 6th International Conference on*

Computing for Sustainable Global Development (INDIACom) (σσ. 1003-1007). IEEE.

Gao, Y., Huyan, Z., & Ju, F. (2015). 2015. *8th International Symposium on Computational Intelligence and Design (ISCID)* (σσ. 219-222). IEEE.

Gollapudi, S. (2016). *Practical machine learning.* Packt Publishing Ltd.

Goodfellow, I., Yoshua, B., & Courville, A. (2016). *Deep Learning.* Boston: MIT Press.

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *IEEE Transactions on Vehicular Technology* , 140-150.

Hassan, L., Santas, B., & Vink, J. (2021). Airline disruption management: A literature review and practical challenges. *Computers & Operations Research*, 105-137.

Ilic, I., Gorgulu, B., Cevik, M., & Baydogan, M. (2021). Explainable boosted linear regression for time series forecasting. *Pattern Recognition*, 108-144.

Jiang, Y., Miao, J., Zhang, X., & Le, N. (2020). A multi-index prediction method for flight delay based on long short-term memory network model. *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology* (σσ. 159-163). ICCASIT.

Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016). A deep learning approach to flight delay prediction. *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 1-6.

Laaksonen, J., & Oja, E. (1996). Classification with learning k-nearest neighbors. *Proceedings of international conference on neural networks* (σσ. 1480-1483). IEEE.

L'heureux, A., Grolinger, K., Elyamany, H., & Capretz, M. (2017). Machine learning with big data: Challenges and approaches. *Ieee Access*, 7776-7797.

Li, Q. J. (2021). Generation and prediction of flight delays in air transport. *Intelligent Transport Systems*, 740-753.

Liu, Y. (2019). Evaluating Impact of Flight Delay on Cargo and Overnight Package Delivery Firms. *Conference: Transportation Research Board 94th Annual MeetingTransportation Research Board*.

Liu, Y., Yin, M., & Hansen, M. (2019). Economic costs of air cargo flight delays related to late package deliveries. *Transportation Research Part E: Logistics and Transportation Review*, 388-401.

Montgomery, D. C., Peck, E., & Vining, G. (2021). *Introduction to linear regression analysis.* John Wiley & Sons.

Motchell, T. M. (1997). *Machine Learning.* McGraw-Hill .

Najafabadi, M., Villanustre, F., Khoshgoftaar, T., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big data*, 1-21.

Pamplona, D., Weigang, L., de Barros, A., Shiguemori, E., & Alves, C. (2018). Supervised neural network with multilevel input layers for predicting of air traffic delays. *2018 International Joint Conference on Neural Networks (IJCNN)* , 16.

Schultz, M., Lorenz, S., Schmitz, R., & Delgado, L. (2019). Weather impact on airport performance. *Aerospace*, 109.

Seelhorst, M. T. (2014). *Flight Cancellation Behavior and Aviation System Performance.* Berkeley: UniversityofCalifornia,Berkeley.

Segal, M. (2004). *Machine learning benchmarks and random forest regression.*

Wright, R. E. (1995). *Logistic regression.* Wright.

Xing, Z., & Tang, Y. (2016). 'The model for optimizing airport flight delays allocation. *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics* (σσ. 188-191). IHMSC.

Yuan, X., Low, J., & Tang, C. (2010). Roles of the airport and logistics services on the economic outcomes of an air cargo supply chain. *International Journal of Production Economics*, 215-225.

Zamkova, M., Rojik, S., Prokop, M., & Stolin, R. (2022). Factors Affecting the International Flight Delays and Their Impact on Airline Operation and Management and Passenger Compensations Fees in Air Transport Industry: Case Study of a Selected Airlines in Europe. *Sustainability* , 14-22.