Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική εργασία

ΑΝΑΠΤΥΞΗ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ ΙΚΑΝΟΠΟΙΗΣΗΣ ΕΠΙΒΑΤΩΝ ΜΙΑΣ ΑΕΡΟΠΟΡΙΚΗΣ ΕΤΑΙΡΕΙΑΣ

του

Λάζαρου Αμανατίδη του Χαράλαμπου

Επιβλέπων καθηγητής: Κωνσταντάρας Ιωάννης

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Φεβρουάριος 2023

ΑΦΙΕΡΩΣΕΙΣ

Θα ήθελα να αφιερώσω την παρούσα διπλωματική εργασία στην οικογένεια μου και συγκεκριμένα στους γονείς μου και την αδερφή μου για την συμπαράσταση που μου πρόσφεραν καθ' όλη τη διάρκεια των σπουδών μου.

ΕΥΧΑΡΙΣΤΙΕΣ

Με την παρούσα διπλωματική εργασία ολοκληρώνεται η φοίτησή μου στο πρόγραμμα μεταπτυχιακών σπουδών στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων.

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κύριο Ιωάννη Κωνσταντάρα που μου έδωσε την ευκαιρία να συνεργαστώ μαζί του για την υπομονή και κατανόηση που μου έδειξε αλλά και για την καθοδήγηση και την υποστήριξη που μου παρείχε.

ΠΕΡΙΛΗΨΗ

Μια αεροπορική εταιρεία δίνει πρωταρχική σημασία στην ικανοποίηση των επιβατών της. Ως εκ τούτου, στόχος της παρούσας εργασίας είναι η εύρεση παραγόντων που επηρεάζουν την ικανοποίηση του επιβάτη, η δημιουργία ενός μοντέλου που να προβλέπει την ικανοποίηση του επιβάτη βάση της αξιολόγησης ορισμένων υπηρεσιών και η εξαγωγή συμπερασμάτων σχετικά με την ικανοποίηση των επιβατών. Ακολουθώντας τη μεθοδολογία CRISP-DM, ξεκινώντας από τους στόχους επιχειρηματικού ενδιαφέροντος, γίνεται συλλογή και επεξεργασία δεδομένων που επιτρέπουν την εύρεση των σημαντικότερων παραγόντων που επηρεάζουν την ικανοποίηση των επιβατών καθώς και τη δημιουργία μοντέλων Μηχανικής Μάθησης που να προβλέπουν το βαθμό ικανοποίησης του επιβάτη λαμβάνοντας υπόψη τις αξιολογήσεις των σημαντικότερων παραγόντων-υπηρεσιών της πτήσης. Συγκεκριμένα, αναπτύσσονται τα μοντέλα Δέντρου Απόφασης και Τυχαίου Δάσους που αποτελούν 2 αλγορίθμους κατηγοριοποίησης και επιλέγεται ο πιο βέλτιστος βάση των μετρικών αξιολόγησης που έχουν τεθεί. Τέλος, με χρήση όλων αυτών των δεδομένων και εργαλείων γλωσσών προγραμματισμού, εξάγονται συμπεράσματα σχετικά με το ποσοστό ικανοποιημένων επιβατών, τις υπηρεσίες με καλύτερες επιδόσεις καθώς και τις υπηρεσίες με τις χαμηλότερες επιδόσεις προκειμένου η Διοίκηση να λάβει αποφάσεις και να προτείνει μέτρα βελτίωσης.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ 1
ΚΕΦΑΛΑΙΟ 2: ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ 3
ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ 8
ΚΕΦΑΛΑΙΟ 4: ΑΝΑΠΤΥΞΗ ΘΕΜΑΤΟΣ 14
ΚΕΦΑΛΑΙΟ 5: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΡΜΗΝΕΙΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ 35
ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ-ΠΡΟΤΑΣΕΙΣ 44
ΒΙΒΛΙΟΓΡΑΦΙΑ/ΑΡΘΟΓΡΑΦΙΑ 47
ΠΑΡΑΡΤΗΜΑΤΑ …..... 55

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ ΚΑΙ ΕΙΚΟΝΩΝ

ΠΙΝΑΚΕΣ

Πίνακας 1: Χωρισμός δεδομένων βάση το είδος ερώτησης και των τιμών των μεταβλητών 18
Πίνακας 2: Ψευδοκώδικας του αλγορίθμου Δέντρου Απόφασης..... 23
Πίνακας 3: Ψευδοκώδικας του αλγορίθμου Τυχαίο Δάσος 28
Πίνακας 4: Σημαντικότεροι Παράγοντες που επηρεάζουν την ικανοποίηση ενός επιβάτη μιας πτήσης..... 36
Πίνακας 5: Σύγκριση αποτελεσμάτων των 2 μοντέλων..... 42

ΕΙΚΟΝΕΣ

Εικόνα 1: Σχηματική απεικόνιση μεθοδολογίας CRISP-DM 8
Εικόνα 2: Απεικόνιση δεδομένων στην ιστοσελίδα 16
Εικόνα 3:Κώδικας στην Python για ανάγνωση δεδομένων και δημιουργία ενός πλαισίου δεδομένων 17
Εικόνα 4: Διάγραμμα ροής του αλγορίθμου Δέντρου Απόφασης..... 26
Εικόνα 5: Κώδικας και επεξηγήσεις αλγορίθμου Δέντρου Απόφασης..... 32
Εικόνα 6: Διάγραμμα ροής του αλγορίθμου Τυχαίο Δάσος 33
Εικόνα 7: Κώδικας και επεξηγήσεις αλγορίθμου Τυχαίο Δάσος 34
Εικόνα 8: Γραφική απεικόνιση Δέντρου Απόφασης 29
Εικόνα 9: Μορφή σύντομου ερωτηματολογίου όπως θα αποστέλλεται στους επιβάτες 32
Εικόνα 10: Προβολή των απαντήσεων επιβατών και της πρόβλεψης που δίνει το μοντέλο σε Φύλλο εργασίας Excel..... 33
Εικόνα 11: Σχηματική περιγραφή της διαδικασίας εφαρμογής του μοντέλου..... 34
Εικόνα 12: Προβολή αποτελεσμάτων αξιολόγησης Δέντρου Απόφασης..... 58
Εικόνα 13: Προβολή αποτελεσμάτων αξιολόγησης Τυχαίου Δάσους 60
Εικόνα 14: Κώδικας και output της εντολής info 58
Εικόνα 15: Κώδικας και output της εντολής describe για αριθμητικές μεταβλητές 59

Εικόνα 16: Κώδικας και output της εντολής describe για κατηγορικές μεταβλητές..... 59
Εικόνα 17: Κώδικας και output για δημιουργία προβολή ιστογραμμάτων..... 60
Εικόνα 18: Κώδικας και output για τη δημιουργία προβολή διαγραμμάτων μπάρας για τις κατηγορικές μεταβλητές 60
Εικόνα 19: Κώδικας και output για τη δημιουργία προβολή συσχετίσεων μεταξύ αριθμητικών μεταβλητών..... 61
Εικόνα 20: Κώδικας για χειρισμό τιμών που λείπουν 62
Εικόνα 21: Κώδικας για την επιλογή 3 σημαντικότερων αριθμητικών μεταβλητών και εμφάνιση αυτών..... 63
Εικόνα 22: Κώδικας για έλεγχο συσχετίσεων κατηγορικών μεταβλητών..... 64
Εικόνα 23: Κώδικας για κωδικοποίηση κατηγορικών μεταβλητών..... 64
Εικόνα 24: Κώδικας για χωρισμό του συνόλου 65
Εικόνα 25: Κώδικας υπολογισμού ορθότητας 66
Εικόνα 26: Κώδικας εύρεσης πίνακα σύγχυσης και αναφοράς ταξινόμησης..... 67

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΡΤΗΜΑ 1: ΜΟΡΦΗ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟΥ 55
ΠΑΡΑΡΤΗΜΑ 2: ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ PYTHON..... 56
ΠΑΡΑΡΤΗΜΑ 3: ΕΞΕΡΕΥΝΗΣΗ ΔΕΔΟΜΕΝΩΝ 58
ΠΑΡΑΡΤΗΜΑ 4: ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ 62
ΠΑΡΑΡΤΗΜΑ 5: ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ..... 65

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

Σύμφωνα με τον BurgueñoSalas (2022), η βιομηχανία των αερομεταφορών αποτελεί την ταχύτερη βιομηχανία του κλάδου των μεταφορών στον κόσμο. Πράγματι, η ύπαρξη μεταφορικού μέσου καθίσταται απαραίτητη για την κάλυψη ποικίλλων αναγκών όπως προσωπικές ή επαγγελματικές υποχρεώσεις καθώς και για τον τουρισμό.(Πρίντεζης, 1995) Η μεταφορά με αεροπλάνο προτιμάται λόγω της μεγάλης απόστασης ενός ταξιδιού σε συνδυασμό με την ελαχιστοποίηση του χρόνου που προσφέρει λόγω της άνεσης και της ταχύτητας του. Σημειώνεται ότι το αεροπλάνο θεωρείται το ασφαλέστερο μεταφορικό μέσο για μεγάλες διαδρομές (IATA, 2018).

Οι αεροπορικές εταιρείες δίνουν πλέον ιδιαίτερη σημασία στην ικανοποίηση των επιβατών και διαθέτουν συγκεκριμένα τμήματα (π.χ. «CustomerRelationsDepartment») που διεξάγουν σχετικές έρευνες. Μια αυξημένη ικανοποίηση των επιβατών οδηγεί σε περισσότερα κέρδη και αυξάνει τα επίπεδα της πιστότητας (Namukasa, 2013) καθώς όλο και περισσότεροι επιβάτες θα επιλέξουν ξανά την ίδια εταιρεία για το ταξίδι τους. Παράλληλα, όπως αναφέρει η Scheffler (2018) η ύπαρξη όλων και περισσότερων ανταγωνιστών στην αγορά επιτείνει την ανάγκη για ικανοποιημένους επιβάτες, με τη δημιουργία ενός υψηλής ποιότητας προϊόντος που θα διαφοροποιήσει την εταιρεία από τους ανταγωνιστές της.

Έτσι, ο σκοπός του έργου αυτού είναι να αξιολογηθεί η επίδραση διαφορετικών παραμέτρων (όπως το προφίλ του επιβάτη και η άποψη του για το επίπεδο των διαφόρων υπηρεσιών που παρέχονται) στην ικανοποίησή του επιβάτη και εν συνεχεία να μπορεί να προβλεφθεί βάση των σημαντικότερων παραγόντων αν ο επιβάτης είναι ικανοποιημένος ή όχι.

Συγκεκριμένα, οι στόχοι του έργου συνοψίζονται ως εξής:

- Εύρεση των σημαντικότερων παραγόντων που επιδρούν στην ικανοποίηση του επιβάτη
- Πρόβλεψη της ικανοποίησης (ή της μη ικανοποίησης) βάση της άποψης που έχει ο επιβάτης για τους σημαντικότερους παράγοντες επίδρασης
- Αξιολόγηση συμπερασμάτων ή αποτελεσμάτων από τους παράγοντες και το μοντέλο πρόβλεψης για τη δημιουργία νέων ανταγωνιστικών προϊόντων που αυξάνουν την ικανοποίηση του επιβάτη

ΚΕΦΑΛΑΙΟ 2: ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Στη συνέχεια, θα γίνει σύντομη αναφορά σε ερευνητικά άρθρα τα οποία σχετίζονται με μεθοδολογίες και τεχνικές τόσο σχετικά με προβλήματα Μηχανικής Μάθησης όσο και γενικότερα του κλάδου της επιστήμης δεδομένων.

Η βιβλιογραφία που αναζητήθηκε χωρίζεται σε μεθοδολογία προσέγγισης έργου επιστήμης δεδομένων, τρόποι εύρεσης σημαντικότερων μεταβλητών προκειμένου να χρησιμοποιηθούν σε μοντέλα, αλγόριθμοι μηχανικής μάθησης κατηγοριοποίησης και τεχνικές προετοιμασίας δεδομένων που χρειάζονται σε τέτοιου είδους προβλήματα καθώς και τι είδους δεδομένα θα χρησιμοποιηθούν:

i. Μεθοδολογία προσέγγισης έργου επιστήμης Δεδομένων

- Σε τι διαφέρει το crisp DM από το Semma;

  https://www.datascience-pm.com/crisp-dm-2/

Το άρθρο αυτό αντιπαραβάλλοντας τις τεχνικές crispDM και Semma προσδιορίζει τη χρησιμότητα της μεθόδου crispDM και παρουσιάζει αναλυτικά κάθε φάση της. Πράγματι, το επιχειρηματικό ενδιαφέρον που έχει η πρόβλεψη της ικανοποίησης του επιβάτη μιας αεροπορικής εταιρείας καθιστά αναγκαία τη χρήση της τεχνικής CRISPDM που έχει αρχικό βήμα την κατανόηση επιχειρηματικών αναγκών. Παράλληλα, ο χωρισμός των ενεργειών ενός έργου ανάλυσης δεδομένων σε φάσεις καθιστά ευκολότερη την υλοποίησή του.

ii. Τρόποι εύρεσης σημαντικότερων μεταβλητών

- Πώς γίνεται επιλογή ιδιοτήτων που είναι αριθμητικές μεταβλητές;

https://machinelearningmastery.com/feature-selection-with-numerical-input-data/

Το άρθρο περιγράφει δύο μεθόδους για την επιλογή των πιο σημαντικών αριθμητικών μεταβλητών σε ένα πρόβλημα κατηγοριοποίησης (Kuhn, 2019). Εξηγούνται σύντομα και δίνεται αλγόριθμος για κάθε μία από αυτές. Η μία βασίζεται στο F-ANOVA test που υπολογίζει συσχετίσεις μεταξύ της εξαρτημένης κατηγορικής μεταβλητής με κάθε μια από τις αριθμητικές μεταβλητές. Η άλλη λέγεται «Αμοιβαία Πληροφόρηση», και υπολογίζει την αβεβαιότητα πρόβλεψης της εξαρτημένης μεταβλητής με κάθε μια από τις ανεξάρτητες αριθμητικές μεταβλητές. Στην περίπτωση της ανάλυσης της ικανοποίησης των επιβατών, αυτές οι τεχνικές και κατ' επέκταση οι αλγόριθμοι τους, θα επιτρέψουν την εύρεση των πιο σημαντικών αριθμητικών μεταβλητών που μπορούν να χρησιμοποιηθούν για την πρόβλεψη ικανοποίησης του επιβάτη.

- Επιλογή ιδιοτήτων με χρήση X^2τεστ

https://medium.com/analytics-vidhya/categorical-feature-selection-using-chi-squared-test-e4c0d0af6b7e

Το άρθρο εστιάζει στην περίπτωση προβλήματος κατηγοριοποίησης που η εξαρτημένη μεταβλητή είναι κατηγορική αλλά και οι ανεξάρτητες μεταβλητές είναι κατηγορικές (Anand, 2020). Οπότε, προτείνεται μέθοδος με X^2 τεστ για την εύρεση σημαντικότερων κατηγορικών ανεξάρτητων μεταβλητών. Αυτό γίνεται με τη μελέτη της κάθε συσχέτισης. Πρώτα όμως χρειάζεται η κωδικοποίηση των κατηγορικών μεταβλητών και η μέτρηση των συχνοτήτων κάθε τιμής για κάθε κατηγορία. Με ελέγχους υποθέσεων μπορούν να βρεθούν οι σημαντικότερες μεταβλητές με χρήση του p-valueπου καθορίζει πόσο συσχετισμένες είναι 2 μεταβλητές. Στο έργο αυτό μπορεί να εφαρμοστεί η τεχνική αυτή για να

αποφασιστούν οι σημαντικότεροι παράγοντες που είναι κατηγορικές μεταβλητές μελετώντας συσχετίσεις τους με την ικανοποίηση του επιβάτη.

- Βήμα-Βήμα Εξήγηση της Ανάλυσης Βασικών Παραγόντων (PCA)

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Το άρθρο εστιάζει στην εύρεση των σημαντικότερων παραγόντων που είναι αριθμητικές μεταβλητές, εξηγώντας τη σημασία της διαδικασίας αυτής για την υλοποίηση αλγορίθμων μηχανικής μάθησης και τη μεγιστοποίηση της ορθότητας των αλγορίθμων. (Jaadi, 2022) Επίσης περιγράφονται τα βήματα της διαδικασίας αυτής αναλυτικά. Βασίζεται στη δημιουργία πίνακα συνδιασποράς μεταξύ των μεταβλητών, στον υπολογισμό ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα προκειμένου να υπολογιστεί η μεταβλητότητα που δημιουργεί κάθε μεταβλητή στην απαντητική μεταβλητή. Όσο μεγαλύτερη είναι η μεταβλητότητα τόσο σημαντικότερη είναι για να επιλεχθεί ως παράγοντας. Στην περίπτωση αυτή μπορεί να χρησιμοποιηθεί για τις αριθμητικές μεταβλητές που επηρεάζουν την ικανοποίηση του επιβάτη.

iii. Αλγόριθμοι μηχανικής μάθησης κατηγοριοποίησης

- Αλγόριθμος Μηχανικής Μάθησης Δέντρου Απόφασης

https://www.analyticsvidhya.com/blog/2022/03/decision-tree-machine-learning-using-python/

Το άρθρο περιγράφει συνοπτικά πώς λειτουργεί ο αλγόριθμος Δέντρου Απόφασης για την κατηγοριοποίηση μιας μεταβλητής και εξηγείται πώς θα υλοποιηθεί στη γλώσσα προγραμματισμού Python.(Amrutha, 2022) Επιλέγεται μια ιδιότητα και δημιουργείται μια συνθήκη με στόχο να χωριστεί το δείγμα στην κάθε κατηγορία. Αυτό μπορεί να εφαρμοστεί με χρήση των σημαντικότερων παραγόντων που επηρεάζουν την ικανοποίηση του επιβάτη προκειμένου να χωριστούν οι επιβάτες σε ικανοποιημένους και μη, βάση των τιμών που λαμβάνει κάθε παράγοντας.

- Κατανοώντας το Τυχαίο Δάσος

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Το άρθρο αυτό αναφέρει συνοπτικά πώς λειτουργεί το Τυχαίο Δάσος και ποιες είναι οι διαφορές του με τον αλγόριθμο Δέντρου Απόφασης. (Sruthi, 2022) Επίσης δίνεται πάλι ο κώδικας στην Python για να υλοποιηθεί το Τυχαίο Δάσος. Το Τυχαίο Δάσος αποτελείται από πολλά Δέντρα Απόφασης με σκοπό τη μεγιστοποίηση της ακρίβειας του μοντέλου. Όμοια λοιπόν εφαρμόζεται ο αλγόριθμος Τυχαίο Δάσους στο σύνολο δεδομένων με τα στοιχεία των επιβατών προκειμένου να κατηγοριοποιηθούν σε ικανοποιημένους και μη.

- Λογιστική Παλινδρόμηση – Αλγόριθμος Εποπτευόμενης μάθησης για Κατηγοριοποίηση

https://www.analyticsvidhya.com/blog/2021/05/logistic-regression-supervised-learning-algorithm-for-classification/

Το άρθρο παρουσιάζει έναν άλλο τρόπο κατηγοριοποίησης με χρήση παλινδρόμησης (Agrawal, 2021). Αντίστοιχα θα εφαρμοστεί ο αλγόριθμος Λογιστικής Παλινδρόμησης στο σύνολο δεδομένων με τα στοιχεία των επιβατών προκειμένου να κατηγοριοποιηθούν σε ικανοποιημένους και μη.

- Μετρικές Αξιολόγησης για τη Μηχανική Μάθηση- Ορθότητα,Ακρίβεια, Ανάκληση και F1-score

http://wiki.pathmind.com/accuracy-precision-recall-f1

Ο Nicholson (2022) παρουσιάζει μετρικές που θα χρησιμοποιηθούν για να αξιολογηθούν τα μοντέλα και πόσο καλές είναι οι προβλέψεις. Στην περίπτωση αυτή, βάση μετρικών θα αξιολογηθεί κάθε μοντέλο πόσο καλά προβλέπει αν ένας επιβάτης είναι ικανοποιημένος ή όχι.

iv. Τεχνικές προετοιμασίας δεδομένων

- Πώς ο αλγόριθμος CatBoost λειτουργεί στη Μηχανική Μάθηση;

https://dataaspirant.com/catboost-algorithm/

Το άρθρο αυτό δίνει πληροφορίες για την επεξεργασία δεδομένων πριν χρησιμοποιηθούν στους αλγορίθμους Μηχανικής Μάθησης. Γίνεται αναφορά στην κωδικοποίηση των κατηγορικών μεταβλητών. (Adebayo, 2021) Στην περίπτωση αυτή θα χρησιμοποιηθεί για την κωδικοποίηση κατηγορικών μεταβλητών που μπορεί να θεωρηθούν ως παράγοντες που επηρεάζουν την ικανοποίηση του επιβάτη.

v. Δεδομένα που θα χρησιμοποιηθούν

- Ποιοι παράγοντες επηρεάζουν την ικανοποίηση των επιβατών μιας αεροπορικής εταιρείας;

https://www.wise-geek.com/what-factors-affect-airline-customer-satisfaction.htm

Ο Barnett (2023), δίνει πληροφορίες για στοιχεία που πρέπει να αναζητηθούν από τους επιβάτες τα οποία θα μπορούσαν να επηρεάσουν την ικανοποίησή του. Οι παράγοντες αυτοί θα χρησιμοποιηθούν για να γίνουν ερωτηματολόγια και τελικά να σχηματιστεί το σύνολο δεδομένων που θα χρησιμοποιηθεί για να αναπτυχθούν τα μοντέλα.

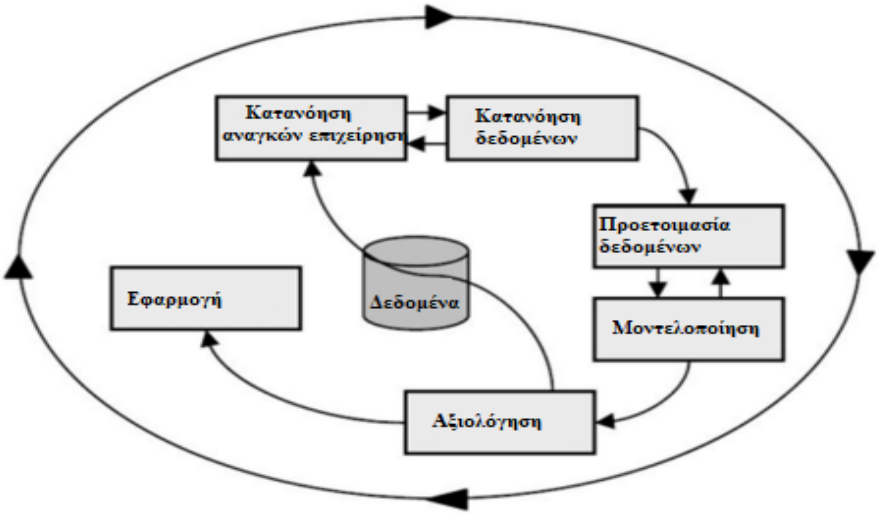
ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ

3.1: ΜΕΘΟΔΟΛΟΓΙΑ CRISP-DM: ΠΕΔΙΟ ΕΦΑΡΜΟΓΗΣ

Ακολουθείται η μεθοδολογία CRISP-DM, η οποία είναι κατάλληλη για την υλοποίηση έργων ανάλυσης δεδομένων σε επιχειρηματικό περιβάλλον (Schröer etal., 2021) όπως είναι μια αεροπορική εταιρεία. Αυτή η μεθοδολογία επιτρέπει ο τελικός χρήστης (πελάτης) να μη συμμετέχει στην ανάπτυξη του έργου. Επιπλέον, αυτή η μεθοδολογία επιτρέπει την ανάπτυξη και τη συντήρηση του έργου στην περίπτωση που το επιλεχθέν μοντέλο δεν είναι ιδανικό. (Santos & Azevedo, 2005) Επίσης, χάρη στη μεθοδολογία αυτή το έργο που αναπτύσσεται μπορεί εύκολα να συσχετιστεί με άλλα έργα. Δηλαδή, άλλες ομάδες χρησιμοποιούν τη γνώση που παράγεται από το εν λόγω έργο και δουλεύουν σε άλλο έργο χρησιμοποιώντας αρχικά την παραχθείσα γνώση.

3.2: ΜΕΘΟΔΟΛΟΓΙΑ CRISP-DM: ΒΗΜΑΤΑ

Ο κύκλος του έργου, σύμφωνα με τη μεθοδολογία CRISP-DM, αναπαρίσταται στο παρακάτω σχήμα:



Εικόνα 1: Σχηματική απεικόνιση μεθοδολογίας CRISP-DM

Ανάλογα με τη διεύθυνση των βελών, επιτρέπεται η κίνηση εμπρός ή και πίσω σε διαφορετικές φάσεις. Ο εξωτερικός κύκλος συμβολίζει την κυκλική φύση του έργου. Αυτό σημαίνει ότι το έργο δεν τελειώνει με το που θα προταθεί η λύση του. Αντίθετα, με βάση τα αποτελέσματα που θα λαμβάνονται κατά τη διάρκεια χρήσης της προτεινόμενης λύσης, θα βρίσκονται σημεία βελτίωσης.

Στη συνέχεια θα εξηγηθεί περιληπτικά κάθε μία από τις φάσεις:

i. Κατανόηση αναγκών επιχείρησης: Επικεντρώνεται στην κατανόηση των στόχων του έργου, την ανίχνευση του προβλήματος και το σχεδιασμού πλάνου επίλυσης.

ii. Κατανόηση των δεδομένων: Περιλαμβάνει τη συλλογή δεδομένων, ενέργειες που επιτρέπουν την εξοικείωση με τα δεδομένα καθώς και την ανακάλυψη υποσυνόλων δεδομένων που παρουσιάζουν ενδιαφέρον για να σχηματιστούν υποθέσεις σχετικά με την «κρυμμένη» πληροφορία.

iii. Προετοιμασία των δεδομένων: Καλύπτει όλες τις ενέργειες για τη δημιουργία του τελικού συνόλου δεδομένων από τα αρχικά ακατέργαστα δεδομένα, που πρόκειται να χρησιμοποιηθεί για την εκτέλεση του μοντέλου. Αυτές οι λειτουργίες σχετίζονται με τη μετατροπή και τον καθαρισμό των δεδομένων.

iv. Μοντελοποίηση: Σε αυτήν τη φάση επιλέγονται και εφαρμόζονται οι τεχνικές που μπορούν να επιλύσουν το πρόβλημα και προσδιορίζονται οι παράμετροι λαμβάνοντας τις βέλτιστες τιμές τους. Κάθε φορά που κάποιο δεδομένο χρειάζεται μετατροπή, μπορεί να γίνει μετάβαση στην προηγούμενη φάση της προετοιμασίας των δεδομένων.

v. Αξιολόγηση: Συγκρίνεται το αποκτημένο μοντέλο με τους στόχους της επιχείρησης. Επίσης, καθορίζεται αν υπάρχει κάποια σημαντική επιχειρηματική ερώτηση που δεν έχει ληφθεί υπόψη.

vi. Εφαρμογή: Παρουσιάζεται το μοντέλο σε τελική μορφή προκειμένου να το χρησιμοποιήσει ο τελικός χρήστης. Μπορεί να έχει μορφή αναφοράς ή να είναι μια περιοδική έκδοση ή αυτοματοποιημένη έκδοση διαμέσου μιας διαδικασίας ανάλυσης δεδομένων στον οργανισμό. (Chapman et al, 2000)

3.3: ΜΕΘΟΔΟΛΟΓΙΑ CRISP-DM: ΠΕΡΙΟΡΙΣΜΟΙ

Όπως φαίνεται η μεθοδολογία CRISP-DM είναι προσανατολισμένη στις επιχειρηματικές ανάγκες. Ωστόσο, σε πολλά έργα η διαδικασία χρειάζεται να συντομευθεί και αυτό οδηγεί σε μια κατεστραμμένη έκδοση της προσέγγισης που περιγράφηκε παραπάνω. Έτσι μπορεί να προκύψουν οι παρακάτω περιορισμοί:

- Έλλειψη σαφήνειας

Είναι πιθανόν να μην υπάρχει εμβάθυνση των επιχειρηματικών αναγκών του προβλήματος. Ως εκ τούτου, γίνεται ανάλυση των δεδομένων χωρίς την πλήρη κατανόηση των επιχειρηματικών στόχων. Αποτελεσματικά, δημιουργούνται ενδιαφέροντα μοντέλα που ωστόσο δεν ανταποκρίνονται σε μια πραγματική επιχειρηματική ανάγκη.

- Άστοχη αξιολόγηση μοντέλου

Η ομάδα της ανάλυσης συχνά αξιολογεί τα αποτελέσματα με κλασσικές μεθόδους και μετρικές χωρίς να εστιάζει στους επιχειρηματικούς στόχους. Οπότε σε περίπτωση μη ικανοποιητικού μοντέλου συχνά αναζητούνται επιπρόσθετα δεδομένα ή υλοποίηση διαφορετικών τεχνικών μοντελοποίησης. Αντίθετα, συχνά χρειάζεται η επαναξιολόγηση του επιχειρηματικού προβλήματος με συναντήσεις με την επιχειρηματική ομάδα.

- Ελλιπής συνεργασία με το τμήμα IT

Αφότου αναπτυχθεί το μοντέλο από την ομάδα ανάλυσης, ακολουθεί η εφαρμογή του σε πραγματικά δεδομένα σε λειτουργικές αποθήκες δεδομένων ή η ενσωμάτωσή του σε λειτουργικά συστήματα. Η αναλυτική ομάδα συχνά δεν εμπλέκεται σε καθήκοντα εφαρμογής, με αποτέλεσμα αν και πολλά μοντέλα είναι πολύ χρήσιμα, είναι δύσκολη η εφαρμογή τους. Λόγω της έλλειψης συνεργασίας με την αντίστοιχη ομάδα, αυξάνεται ο χρόνος και το κόστος και πολλές φορές καταλήγει ως ένα μοντέλο που ποτέ δεν προσφέρει στην επιχείρηση.

- Αποτυχία επανάληψης της μεθοδολογίας

Τα μοντέλα δε μπορούν να εξακολουθούν να είναι πολύτιμα για πάντα δεδομένου ότι οι επιχειρηματικές συνθήκες αλλάζουν συνεχώς. Ωστόσο συχνά δεν καθορίζεται πώς θα παρακολουθείται η επιχειρηματική απόδοση του μοντέλου και δε γίνεται επένδυση για αναθεώρηση του μοντέλου. Έτσι χωρίς παρακολούθηση και συντήρηση υπονομεύεται η μακροπρόθεσμη αξία των αναλυτικών στοιχείων. (Taylor,2017)

3.4: ΜΕΘΟΔΟΛΟΓΙΑ ΓΙΑ ΑΝΑΠΤΥΞΗ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟΥ

Σχετικά με τα ερωτήματα που πρέπει να ερωτηθούν οι επιβάτες για να μετρηθεί η ικανοποίησή τους λαμβάνονται υπόψη παράγοντες που αφορούν σχεδόν κάθε επιχείρηση παροχής υπηρεσιών. Η μεγάλη διαφορά της αεροπορικής εταιρείας από τις άλλες επιχειρήσεις είναι η επίδραση των δραστηριοτήτων της από κανονισμούς Αρχών και Οργανισμών όπως η Αρχή Πολιτικής Αεροπορίας ή ο ICAO (Taha, 2016). Τέτοιοι παράγοντες είναι κανονισμοί ασφάλειας, μετεωρολογικές συνθήκες που οδηγούν σε καθυστερήσεις ή ακυρώσεις πτήσεων είτε λόγω οδηγίας από τις σχετικές Αρχές είτε λόγω απόφασης της εταιρείας που θέτει ως προτεραιότητα την ασφάλεια σε σχέση με την ικανοποίηση του επιβάτη. Ως εκ τούτου πολλές φορές, η ικανοποίηση του επιβάτη δε μπορεί να ελεγχθεί από την εταιρεία οπότε σε αυτές τις περιπτώσεις λαμβάνεται υπόψη ο βαθμός στον οποίο μπορεί να διαχειρίζεται επιτυχώς τα προβλήματα στα οποία δεν ευθύνεται η ίδια. Συγκεκριμένα λοιπόν αξιολογείται η αποδοτικότητα των τμημάτων εξυπηρέτησης επιβατών και διαχείρισης παραπόνων (Tegaretal, 2018).

Παράλληλα, σημαντικό ρόλο αποτελούν και οι υπηρεσίες εν πτήσει. Πράγματι περικοπές στις υπηρεσίες εν πτήσει μπορεί να επηρεάσουν αρνητικά τον επιβάτη ακόμη και αν η τιμή του εισιτηρίου είναι χαμηλότερη (Sanyal et al, 2016). Οι υπηρεσίες εν πτήσει αφορούν την άνεση του καθίσματος, το χώρο για τα πόδια, διασκέδαση εν πτήσει (ταινίες, μουσική κτλ.) καθώς και την προσφορά φαγητού.

Επίσης, χρειάζεται να αξιολογηθεί η εξυπηρέτηση και η άνεση στο αεροδρόμιο πριν και μετά την παραμονή στο αεροσκάφος.

Σύμφωνα με τους Freitasetal (2021), το προφίλ του επιβάτη επηρεάζει την ικανοποίηση του ως προς τις προσφερόμενες υπηρεσίες. Γι’ αυτό, πέρα από την αξιολόγηση όλων αυτών των φάσεων θα ζητηθούν και μερικές λεπτομέρειες του επιβάτη όπως η ηλικία, το φύλο, λόγος ταξιδιού , είδος θέσης (οικονομική/διακεκριμένη), πτήση(διάρκεια/απόσταση) και η πιστότητα του επιβάτη. Αυτά τα στοιχεία θα επιτρέψουν επεξεργασία δεδομένων για την καλύτερη μοντελοποίηση και εύρεση στοιχείων επιχειρηματικού ενδιαφέροντος.

Συνοψίζοντας λαμβάνοντας υπόψη το υπάρχον πλαίσιο, θα σχεδιαστούν ερωτήματα ακολουθώντας την ακόλουθη σειρά:

- Στοιχεία επιβάτη, τύπος πτήσης
- Απόδοση πτήσης (ακρίβεια αναχώρησης, ακύρωση πτήσης κτλ.)
- Μέτρηση γενικής ικανοποίησης για την αεροπορική εταιρεία
- Μέτρηση ικανοποίησης υπηρεσιών στο έδαφος
- Μέτρηση ικανοποίησης υπηρεσιών στην πτήση

Για να ποσοτικοποιηθεί η απάντηση στα ερωτήματα ικανοποίησης δημιουργείται μια κλίμακα από το 1(λίγο) μέχρι το 5(πολύ).
Η μορφή του ερωτηματολογίου που δίνεται στους επιβάτες φαίνεται στο Παράρτημα 1.

ΚΕΦΑΛΑΙΟ 4: ΑΝΑΠΤΥΞΗ ΘΕΜΑΤΟΣ

4.1: ΕΙΣΑΓΩΓΗ

Λαμβάνοντας υπόψη τις βιβλιογραφικές αναφορές και τη μεθοδολογία που περιγράφηκε παραπάνω, θα κατασκευαστεί ένα μοντέλο που εκτιμάει την ικανοποίηση του επιβάτη μιας αεροπορικής εταιρείας. Συγκεκριμένα, θα ληφθούν απαντήσεις σε ερωτήσεις σχετικές με την ικανοποίηση των επιβατών, από δείγμα επιβατών, προκειμένου να κατασκευαστεί το σύνολο δεδομένων που θα χρησιμοποιηθεί για τη μοντελοποίηση.

Θα γίνει εύρεση των σημαντικότερων παραγόντων που επιδρούν στην ικανοποίηση του επιβάτη. Αφενός για τους παράγοντες που είναι αριθμητικές μεταβλητές θα εφαρμοστεί F-ANOVAtest για τη μελέτη συσχετίσεων με την απαντητική κατηγορική μεταβλητή (ικανοποιημένος επιβάτης ή δυσαρεστημένος επιβάτης). Αφετέρου για τους παράγοντες που είναι κατηγορικές μεταβλητές θα εφαρμοστεί X^2 test για τη μελέτη συσχετίσεων με την απαντητική κατηγορική μεταβλητή (ικανοποιημένος επιβάτης ή δυσαρεστημένος επιβάτης).

Κατόπιν, εφόσον έχουν επιλεγεί οι παράγοντες και έχει μετασχηματιστεί το σύνολο δεδομένων θα δημιουργηθεί μοντέλο που θα προβλέπει τον ικανοποιημένο επιβάτη. Για το σκοπό αυτό, θα εφαρμοστούν 2 αλγόριθμοι της Μηχανικής Μάθησης, ο αλγόριθμος του Δέντρου Απόφασης και ο αλγόριθμος Τυχαίο Δάσος που αποτελεί επέκταση του αλγορίθμου Δέντρου Απόφασης (Prajwala, 2015). Τέλος, θα συγκριθούν τα αποτελέσματα και η ικανότητα πρόβλεψης των 2 μοντέλων με χρήση κατάλληλων μετρικών.

Θα παρουσιαστούν αναλυτικά τα παραπάνω βήματα ακολουθώντας τη μεθοδολογία CRISP-DM που περιγράφηκε στο προηγούμενο κεφάλαιο:

4.2: ΒΗΜΑ 1ο – ΚΑΤΑΝΟΗΣΗ ΑΝΑΓΚΩΝ ΕΠΙΧΕΙΡΗΣΗΣ

Μια αεροπορική εταιρεία στα πλαίσια βελτίωσης του προϊόντος της, ενδιαφέρεται να μελετήσει την ικανοποίηση των πελατών της.

Για να υλοποιηθεί αυτό αρχικά χρειάζεται σχεδιασμός ερωτηματολογίου που θα περιλαμβάνει χαρακτηριστικά του ταξιδιού και διάφορες προσφερόμενες υπηρεσίες προς αξιολόγηση από τους επιβάτες.

Αφότου ληφθούν απαντήσεις από αρκετούς επιβάτες, θα χρησιμοποιηθούν τα δεδομένα αυτά για τις εξής ανάγκες:

- Εύρεση των σημαντικότερων παραγόντων που επηρεάζουν την ικανοποίηση του επιβάτη προκειμένου να γίνει επένδυση για βελτίωσή τους
- Εύρεση μοτίβων και στατιστικών στοιχείων με επιχειρηματικό ενδιαφέρον όπως το ποσοστό των ικανοποιημένων επιβατών
- Πρόβλεψη της ικανοποίησης του πελάτη με βάση τα χαρακτηριστικά του ταξιδιού ή και της αξιολόγησης των υπηρεσιών από τον επιβάτη

4.2: ΒΗΜΑ 2ο – ΚΑΤΑΝΟΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Μέσω των διαφόρων καναλιών επικοινωνίας, δόθηκαν στους επιβάτες που ταξίδεψαν πρόσφατα με την εταιρεία ερωτηματολόγια και ελήφθησαν συνολικά 129.880 απαντήσεις, (βλ.ιστοσελίδα: www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction, δεδομένα του χρήστη kaggle TJ Klein). Για κάθε απάντηση, συγκεντρώθηκαν δεδομένα για στοιχεία της πτήσης όπως η διάρκεια της και η καθυστέρηση της, το προφίλ του επιβάτη όπως το φύλο και η πιστότητά του και η βαθμολόγηση του επιβάτη για κάθε μία από τις αξιολογούμενες υπηρεσίες του αεροπορικού ταξιδιού. Επίσης υπάρχει και η απάντηση σχετικά με την ικανοποίηση από την ταξιδιωτική εμπειρία. Τα δεδομένα αυτά αρκούν για να δημιουργηθεί το μοντέλο πρόβλεψης της κατηγορίας του επιβάτη (ικανοποιημένος ή δυσαρεστημένος) με χρήση κάποιων από των προαναφερθέντων παραγόντων έπειτα από κατάλληλη προετοιμασία τους που θα εξηγηθεί στη συνέχεια.

Gender	Customer Type	Age	Type of Travel	Class	Fli
Female 51%	Loyal Customer 82%	7 – 85	Business travel 69%	Business 48%	31
Male 49%	disloyal Customer 18%		Personal Travel 31%	Eco 45%	
				Other (1917) 7%	
Female	Loyal Customer	52	Business travel	Eco	160
Female	Loyal Customer	36	Business travel	Business	2863
Male	disloyal Customer	20	Business travel	Eco	192
Male	Loyal Customer	44	Business travel	Business	3377
Female	Loyal Customer	49	Business travel	Eco	1182
Male	Loyal Customer	16	Business travel	Eco	311
Female	Loyal Customer	77	Business travel	Business	3987
Female	Loyal Customer	43	Business travel	Business	2556
Male	Loyal Customer	47	Business travel	Eco	556
Female	Loyal Customer	46	Business travel	Business	1744

Εικόνα 2: Απεικόνιση δεδομένων στην ιστοσελίδα

Τα δεδομένα που χρειάζονται βρίσκονται σε ένα αρχείο data_airline_satisfaction.xlsx. Αυτά τοποθετούνται σε μια δομή δεδομένων που παρέχεται από τη βιβλιοθήκη pandas της Python, που λέγεται Πλαίσιο Δεδομένων (Snehkunjetal., 2022). Έτσι μπορούν να μετατραπούν ή να καθαριστούν τα δεδομένα για να πραγματοποιηθεί η ανάλυση που χρειάζεται. Αυτές οι ενέργειες θα αναπτυχθούν στη συνέχεια. Περισσότερες λεπτομέρειες σχετικά με τις δυνατότητες αυτής της γλώσσας προγραμματισμού που αφορούν όλες τις φάσεις ανάπτυξης του θέματος θα δοθούν στο Παράρτημα 2.

Στην παρακάτω εικόνα φαίνεται ο σχετικός κώδικας που χρειάζεται για να διαβαστούν τα δεδομένα και να δημιουργηθεί ένα πλαίσιο δεδομένων:

```
import pandas as pd

df=pd.read_excel('/kaggle/input/dataairlinesatisfactionv1/dataairlinesatisfactionv1.xlsx')
```

Εικόνα 3:Κώδικας στην Pythonγια ανάγνωση δεδομένων και δημιουργία ενός πλαισίου δεδομένων

Όπως αναφέρει και ο Lewis (2020), με τη χρήση κατάλληλων εντολών είναι εφικτό να γίνει εξερεύνηση των δεδομένων. Ως εκ τούτου, φαίνονται τα ονόματα των μεταβλητών και τα είδη τους (αριθμητική, συμβολοσειρά κτλ.). Παράλληλα, μπορούν να εξαχθούν στατιστικά δεδομένα (μέσος, διάμεσος, τυπική απόκλιση, μέγιστη/ελάχιστη τιμή κτλ.) για τις αριθμητικές μεταβλητές, αλλά και σύνολο πιθανών τιμών και εύρεση συχνότερης τιμής για τις κατηγορικές μεταβλητές. Στον παρακάτω πίνακα φαίνεται ο χωρισμός των μεταβλητών σε αριθμητικές και κατηγορικές ανάλογα με την κατηγορία ερωτήσεων:

Κατηγορία κριτηρίων/Είδος μεταβλητών	Συμβολοσειρά	Αριθμητική
Προφίλ επιβάτη	Φύλο	Ηλικία
	Πιστότητα επιβάτη	
	Λόγος Ταξιδιού	
Χαρακτηριστικά ταξιδιού	Είδος θέσης	Απόσταση πτήσης
		Καθυστέρηση αναχώρησης/άφιξης
Αξιολόγηση υπηρεσιών		Wifi εν πτήσει
		Ώρες αναχώρησης/άφιξης
		Διαδικασία onlineκράτησης
		Τοποθεσία εξόδου αναχώρησης
		Φαγητά/ποτά
		Επιβίβαση
		Άνεση θέσης
		Διασκέδαση εν πτήσει
		Υπηρεσίες στο αεροσκάφος
		Χώρος στα πόδια
		Χειρισμός αποσκευών
		Check-in
		Υπηρεσίες εν πτήσει
		Καθαριότητα
Γενική Ικανοποίηση	✓	

Πίνακας 1:Χωρισμός δεδομένων βάση το είδος ερώτησης και των τιμών των μεταβλητών

Κατανόηση των δεδομένων γίνεται επίσης μέσω ιστογραμμάτων για κάθε μία από τις αριθμητικές μεταβλητές. Έτσι για κάθε αξιολογούμενη υπηρεσία φαίνεται το εύρος των τιμών που παίρνει καθώς και ποιες τιμές είναι συχνότερες ή σπανιότερες. Όσον αφορά τις κατηγορικές μεταβλητές, χρησιμοποιούνται διαγράμματα μπάρας που δείχνουν τη συχνότητα κάθε πιθανής τιμής μιας κατηγορικής μεταβλητής όπως το είδος θέσης (οικονομική, διακεκριμένη κτλ.) και την ικανοποίηση επιβάτη (ευχαριστημένος/δυσαρεστημένος). Τέλος, ο συντελεστής συσχέτισης Pearsonδείχνει συσχέτιση μεταξύ των αριθμητικών μεταβλητών. (Caoetal., 2021)

Περισσότερες πληροφορίες, δίνονται στο παράρτημα 3 σχετικά με τις εντολές και τα δεδομένα που παρουσιάζονται. Επίσης, σημαντικά αποτελέσματα από την ανάλυση αυτή παρουσιάζονται σε επόμενη ενότητα.

4.3: ΒΗΜΑ 3ο – ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Για τη δημιουργία ενός Δέντρου Απόφασης, δαπανείται γενικά λιγότερος χρόνος στην προετοιμασία δεδομένων σε σχέση με άλλους αλγορίθμους. (Gupta, 2017) Τα αποτελέσματα είναι αξιόπιστα χωρίς κάποιου είδους κανονικοποίηση ή τυποποίηση καθώς τα δέντρα επιλέγουν διαφορετικά σημεία χωρισμού βάση της εκάστοτε μεταβλητής. Ο αλγόριθμος βασίζεται σε κανόνες και όχι σε αποστάσεις μεταξύ των σημείων. (Thenraj, 2020) Για τον ίδιο λόγο, δε χρειάζεται η αφαίρεση ακραίων τιμών. Όσον αφορά την προετοιμασία των δεδομένων για την εφαρμογή του Τυχαίου Δάσους δε διαφέρει από αυτή που γίνεται στο Δέντρο Απόφασης καθώς το Τυχαίο Δάσος αποτελείται από Δέντρα Απόφασης (Ζαγγανά, 2012).

Περισσότερες λεπτομέρειες σχετικά με τον κώδικα που χρειάζεται και επεξηγήσεις δίνονται στο παράρτημα 4. Παρακάτω, περιγράφονται συνοπτικά οι ενέργειες προετοιμασίας και τα αποτελέσματα που χρειάζονται για τη μετάβαση στο επόμενο βήμα της μοντελοποίησης.

Η προετοιμασία των δεδομένων για τους αλγόριθμους Δέντρο Απόφασης και Τυχαίο Δάσος περιλαμβάνει τις εξής ενέργειες (Goyal, 2021) :

- Χειρισμός των τιμών που λείπουν από το σύνολο δεδομένων: Εντοπίζονται οι τιμές και επιλέγεται μια στρατηγική αντικατάστασης των κενών τιμών. Στα δεδομένα αυτού του συνόλου, υπάρχει μια στήλη με ελλιπείς τιμές, αυτή της καθυστέρησης άφιξης σε λεπτά. Οι ελλιπείς τιμές αντικαθίστανται με τη διάμεσο.

- Επιλογή των πιο σημαντικών ιδιοτήτων: Ανάμεσα στα αντικείμενα της έρευνας είναι και οι σημαντικότεροι παράγοντες που επηρεάζουν την ικανοποίηση των επιβατών. Αυτός ο στόχος υλοποιείται σε αυτό το βήμα.. Πράγματι, η επιλογή των ιδιοτήτων επιτρέπει τη βελτιστοποίηση της ακρίβειας των κατηγοριοποιητών ως προς τις προβλέψεις, ελαττώνεται η προσπάθεια για

συλλογή και προετοιμασία δεδομένων, βελτιώνεται η ερμηνεία του μοντέλου και επιταχύνεται ο χρόνος πρόβλεψης. Στο πρόβλημα αυτό επιλέγονται οι 3 σημαντικότερες αριθμητικές μεταβλητές και 1 σημαντικότερη κατηγορική μεταβλητή.

Όσον αφορά τις αριθμητικές μεταβλητές, επιλέγονται όλες οι αξιολογούμενες υπηρεσίες και η καθυστέρηση αναχώρησης/ άφιξης και μελετάται η συσχέτιση σε κάθε μια από αυτές με την εξαρτημένη μεταβλητή του προβλήματος που είναι η κατηγορική μεταβλητή της ικανοποίησης του επιβάτη δηλαδή ευχαριστημένος ή δυσαρεστημένος. Με το κατάλληλο στατιστικό τεστ επιλέγονται οι 3 ιδιότητες που έχουν μεγαλύτερη εξάρτηση με την ικανοποίηση του επιβάτη. Το αποτέλεσμα δίνει **την επιβίβαση**, **την άνεση θέσης** και **τη διασκέδαση εν πτήσει** ως τους σημαντικότερους αριθμητικούς παράγοντες που επηρεάζουν την ικανοποίηση του επιβάτη.

Όσον αφορά τις κατηγορικές μεταβλητές, επιλέγονται η κατηγορία θέσης, η πιστότητα επιβάτη και το είδος ταξιδιού και μελετάται και πάλι η συσχέτιση σε κάθε μια από αυτές με την εξαρτημένη μεταβλητή του προβλήματος που είναι η κατηγορική μεταβλητή της ικανοποίησης του επιβάτη δηλαδή ευχαριστημένος ή δυσαρεστημένος. Με το κατάλληλο στατιστικό τεστ επιλέγεται η ιδιότητα που έχει μεγαλύτερη εξάρτηση με την ικανοποίηση του επιβάτη. Το αποτέλεσμα δίνει **την κατηγορία θέσης (οικονομική, οικονομική +, διακεκριμένη)** ως τον σημαντικότερο κατηγορικό παράγοντα που επηρεάζει την ικανοποίηση του επιβάτη.

- Κωδικοποίηση κατηγορικών μεταβλητών: Προκειμένου να υλοποιηθούν οι αλγόριθμοι του Δέντρου Απόφασης και Τυχαίο Δάσος κωδικοποιείται η κατηγορική μεταβλητή που επιλέχθηκε, αυτή της κατηγορίας θέσης ακολουθώντας τη λογική σειρά. Η Οικονομική θέση λαμβάνει την τιμή 0, η Οικονομική + λαμβάνει την τιμή 1 και η Διακεκριμένη θέση λαμβάνει την τιμή 2 που έχει το καλύτερο επίπεδο υπηρεσιών.

- Δημιουργία συνόλου εκπαίδευσης και τεστ: Για την υλοποίηση τέτοιου είδους αλγορίθμων, χρειάζεται να χωριστούν τα δεδομένα σε σύνολο εκπαίδευσης για

να δημιουργηθεί το μοντέλο και σύνολο τεστ για να αξιολογηθούν οι προβλέψεις του μοντέλου.

4.4: ΒΗΜΑ 4ο – ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Σε αυτό το βήμα θα αναπτυχθεί το μοντέλο για την πρόβλεψη της ικανοποίησης του επιβάτη δεδομένου της αξιολόγησης των αεροπορικών υπηρεσιών και τα χαρακτηριστικά του ταξιδιού. Θα χρησιμοποιηθούν 2 αλγόριθμοι για την υλοποίηση αυτού του σκοπού. Στα παρακάτω κεφάλαια θα συγκριθούν τα αποτελέσματα των 2 αλγορίθμων.

4.4.1: ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

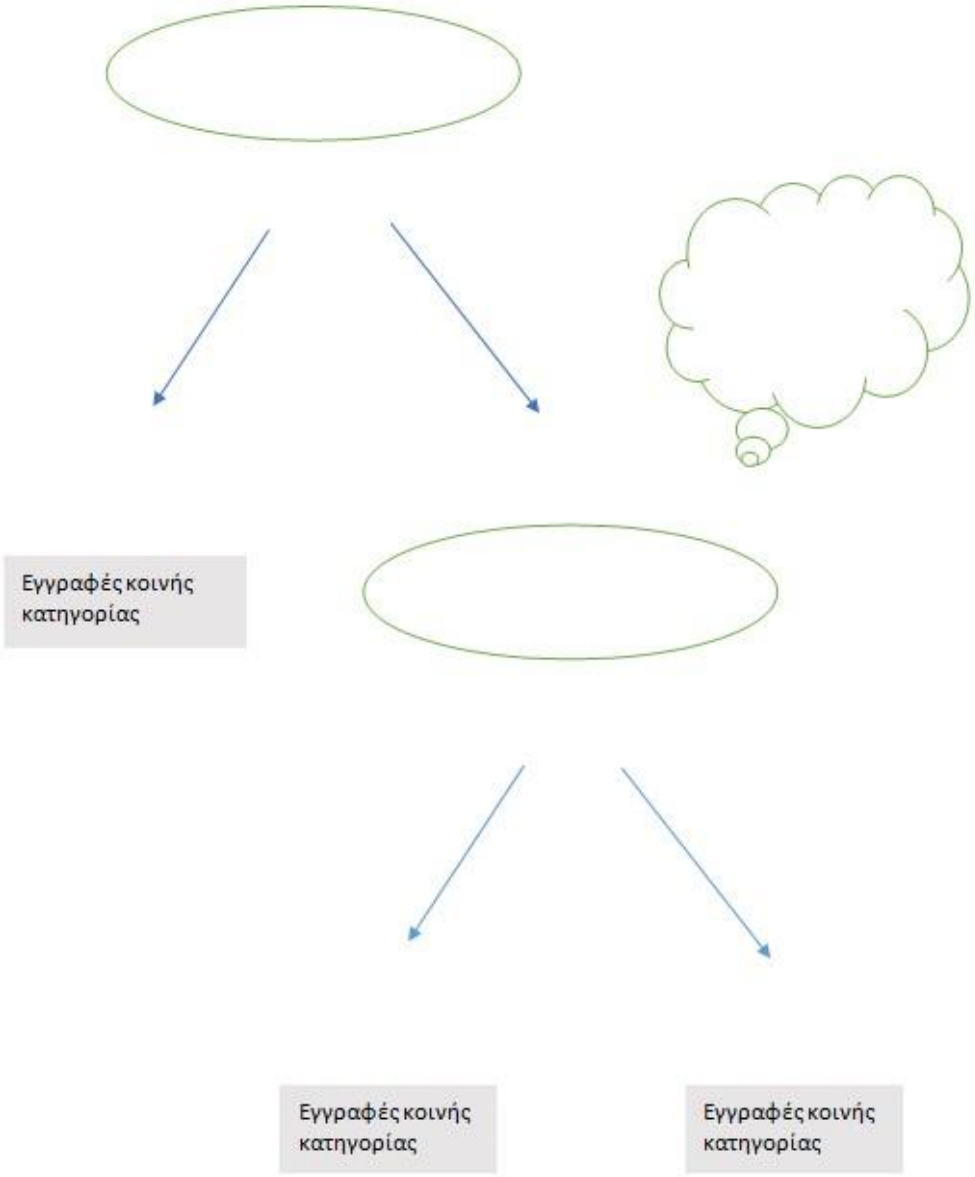
Τα **Δέντρα Απόφασης** ή **Δέντρα Ταξινόμησης** είναι από τα πιο χρησιμοποιούμενα μοντέλα της Επιβλεπόμενης Μηχανικής Μάθησης καθώς θεωρείται απλό μοντέλο κατηγοριοποίησης (SinghChauhan, 2022). Όταν επιλέγεται ο αλγόριθμος των Δέντρων Απόφασης, πρέπει να αποφασιστούν α) οι κατηγορίες ταξινόμησης στις οποίες θα καταλήξει ο διαχωρισμός αλλά και β) τα υπόλοιπα χαρακτηριστικά που πρέπει να παρατηρηθούν, τα οποία θα χρησιμοποιήσει ο αλγόριθμος σε κάθε βήμα διαχωρισμού ώστε να καταλήξει σε ένα σύνολο δειγμάτων με κοινά χαρακτηριστικά.

Ο αλγόριθμος των Δέντρων Απόφασης ξεκινάει με τη ρίζα του Δέντρου που αποτελείται από όλα τα δείγματα (σύνολο εκπαίδευσης). Κατόπιν επιχειρείται ο διαχωρισμός των δειγμάτων του συνόλου με την εισαγωγή ενός κόμβου. Ο κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού . Ο κόμβος έχει ακμές, κάθε μία ονοματίζεται με μια διαφορετική τιμή που μπορεί να πάρει το χαρακτηριστικό του κόμβου από το οποίο ξεκινάει. Κάθε ακμή αντιστοιχίζεται σε ένα φύλλο το οποίο αντιστοιχεί σε μια κατηγορία ταξινόμησης (Γεωργούλη, 2015).

Έπειτα μπορούν να προστεθούν κόμβοι αυξάνοντας το βάθος του δέντρου με ένα νέο χαρακτηριστικό που δεν έχει χρησιμοποιηθεί στο συγκεκριμένο κλαδί του δέντρου. Σε αυτούς τους κόμβους, επιχειρείται ο περαιτέρω διαχωρισμός των μη κατηγοριοποιημένων δειγμάτων του συνόλου εκπαίδευσης που έχουν συγκεντρωθεί εκεί. Όταν ο αλγόριθμος πετυχαίνει ένα πλήρη διαχωρισμό δειγμάτων τότε στα φύλλα του συγκεντρώνονται μόνο ομοιογενή δείγματα, δηλαδή δείγματα ίδιας κατηγορίας.(Sharma, 2020) Κάθε φύλλο με ομοιογενή δείγματα επιτρέπει την εξαγωγή συμπερασμάτων που μπορεί να εκφραστεί ως κανόνας που εκφράζει τον τρόπο προσδιορισμού μιας συγκεκριμένης κατηγορίας.

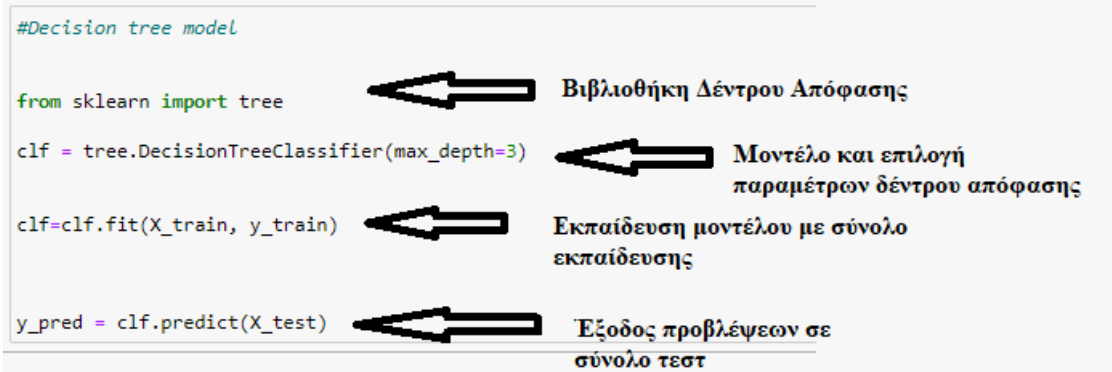
Βήμα 1: Ξεκινάει με έναν κόμβο που περιέχει όλες τις εγγραφές
Βήμα 2: Επιλέγει τυχαία ένα γνώρισμα και μια συνθήκη διαχωρισμού προκειμένου να διασπαστεί ο κόμβος και να μοιραστούν οι εγγραφές
Βήμα 3: Επαναλαμβάνεται το Βήμα 2 σε κάθε κόμβο μέχρι να επιτευχθεί ένας πλήρης διαχωρισμός των δειγμάτων ή μέχρι να ικανοποιηθεί μια συνθήκη που έχει προσδιοριστεί για γρηγορότερο τερματισμό του αλγόριθμου

Πίνακας 2:Ψευδοκώδικας του αλγορίθμου Δέντρου Απόφασης



Εικόνα 4: Διάγραμμα ροής του αλγορίθμου Δέντρου Απόφασης

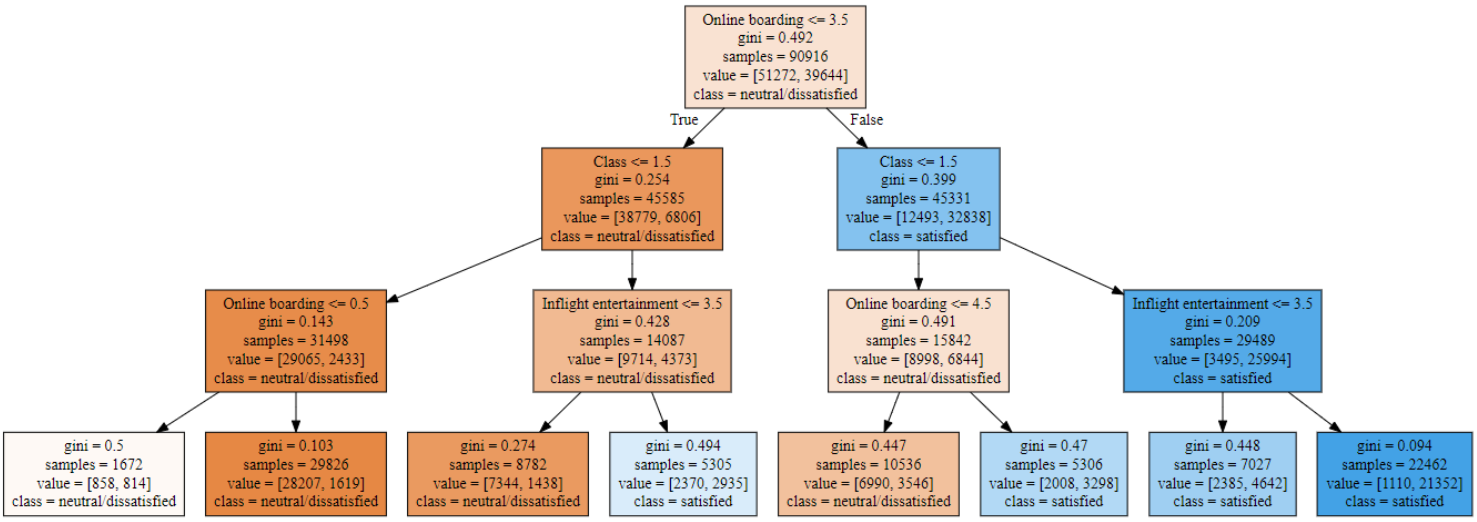
Στον αλγόριθμο Δέντρο Απόφασης χρησιμοποιείται ως υπερπαράμετρος το βάθος που μπορεί να φτάσει το δέντρο. Σημειώνεται ότι χαμηλός αριθμός οδηγεί σε χαμηλή ακρίβεια του μοντέλου ενώ ο πολύ μεγάλος αριθμός μπορεί να οδηγήσει στην υπερπροσαρμογή και να μη δουλεύει καλά στα δεδομένα του τεστ (Bramer, 2016). Ο αλγόριθμος δέχεται σαν είσοδο τις ανεξάρτητες μεταβλητές που επιλέχθηκαν, την εξαρτημένη μεταβλητή της κατηγοριοποίησης (πχ. Ικανοποιημένος ή δυσαρεστημένος επιβάτης) και το μέγιστο βάθος που μπορεί να φτάσει το δέντρο. Τα δεδομένα αυτά ανήκουν στο σύνολο εκπαίδευσης. Για να γίνει ο αλγόριθμος χρειάζεται ο αλγόριθμος tree από τη βιβλιοθήκη sklearn.(Lamrini, 2020) Ως έξοδος λαμβάνονται οι αποφάσεις που προτείνει το μοντέλο στο σύνολο τεστ.



Εικόνα 5: Κώδικας και επεξηγήσεις αλγορίθμου Δέντρου Απόφασης

Ο αλγόριθμος δέχεται σαν είσοδο τις 4 ανεξάρτητες μεταβλητές που επιλέχθηκαν στο προηγούμενο βήμα, την εξαρτημένη μεταβλητή της ικανοποίησης του επιβάτη και το μέγιστο βάθος που μπορεί να φτάσει το δέντρο που είναι 3 ώστε να αποφευχθεί τόσο η χαμηλή ορθότητα όσο και η υπερπροσαρμογή (Bramer, 2016). Το μοντέλο χτίζεται βάσει του συνόλου εκπαίδευσης. Ως έξοδος λαμβάνονται οι προβλέψεις που προτείνει το μοντέλο στο σύνολο τεστ.

Επίσης, παρουσιάζεται παρακάτω και το δέντρο απόφασης που δημιουργήθηκε με αυτά τα δεδομένα. Παρατηρείται ότι ξεκινάει τυχαία ο χωρισμός μιας ιδιότητας, αυτής της επιβίβασης και συνεχίζει ο χωρισμός βάσει της ιδιότητας της κατηγορίας θέσης και συνεχίζεται ο χωρισμός μέχρι να φτάσει το μέγιστο βάθος 3.



Εικόνα 6: Γραφική απεικόνιση Δέντρου Απόφασης

4.4.2: ΤΥΧΑΙΟ ΔΑΣΟΣ

Το Τυχαίο Δάσος είναι ένας αλγόριθμος εποπτευόμενης μηχανικής μάθησης και ένας από τους πλέον χρησιμοποιούμενους χάρη στην ακρίβεια, την απλότητα και την ευελιξία του. Το γεγονός ότι μπορεί να χρησιμοποιηθεί τόσο στην κατηγοριοποίηση όσο και στην παλινδρόμηση αλλά και η μη γραμμική φύση του τον καθιστά εύκολα προσαρμόσιμο σε ποικιλία δεδομένων και καταστάσεων. (Καραντζιάς, 2019) Λέγεται δάσος γιατί αποτελείται από πολλά Δέντρα Απόφασης. Τα δεδομένα αυτών των δέντρων ενώνονται με στόχο την εκτέλεση πιο ακριβών προβλέψεων. Το τυχαίο δάσος εξασφαλίζει ένα πιο ακριβές αποτέλεσμα χάρη στη μεγαλύτερη ποσότητα ομάδων και αποφάσεων σε αντίθεση με αυτή του μοναδικού δέντρου απόφασης. Επιπλέον το χαρακτηριστικό της τυχαιότητας που έχει το μοντέλο επιτρέπει την εύρεση του καλύτερου χαρακτηριστικού ανάμεσα στο τυχαίο υποσύνολο χαρακτηριστικών.(Breiman, 1999)

Διαλέγουμε τον αριθμό των δέντρων που θέλουμε να κατασκευάσουμε. Επιλέγονται τυχαία n αντικείμενα από το σύνολο εκπαίδευσης. Αυτά τα n αντικείμενα αποτελούν το σύνολο εκπαίδευσης για κάθε δέντρο. Όπως και στον αλγόριθμο των Δέντρων Απόφασης, πρέπει να αποφασιστούν α) οι κατηγορίες ταξινόμησης στις οποίες θα καταλήξει ο διαχωρισμός αλλά και β) τα υπόλοιπα χαρακτηριστικά που πρέπει να παρατηρηθούν, τα οποία θα χρησιμοποιήσει ο αλγόριθμος σε κάθε βήμα διαχωρισμού ώστε να καταλήξει σε ένα σύνολο δειγμάτων με κοινά χαρακτηριστικά.

Επιλέγεται τυχαία ένας αριθμός ανεξάρτητων μεταβλητών που θα εξετάζονται σε κάθε δέντρο ο οποίος παραμένει σταθερός καθ' όλη τη διάρκεια κατασκευής του δάσους. Σε κάθε δέντρο λοιπόν, γίνεται διαχωρισμός ως προς τις τιμές των μεταβλητών που έχουν επιλεχθεί και επιλέγεται ο καλύτερος από αυτούς. Αυτός θα χρησιμοποιηθεί ως συνθήκη στον κόμβο του δέντρου. Κάθε δέντρο αναπτύσσεται στο μεγαλύτερο δυνατό βαθμό χωρίς να πραγματοποιείται κλάδεμα. (Biau, 2012) Στην προηγούμενη ενότητα υπάρχουν περισσότερες πληροφορίες για την κατασκευή ενός Δέντρου Απόφασης. Κάνοντας τα παραπάνω βήματα έχει κατασκευαστεί το Τυχαίο Δάσος.

Βήμα 1: Επιλέγεται τυχαία ένα δείγμα από όλο το σύνολο δεδομένων.

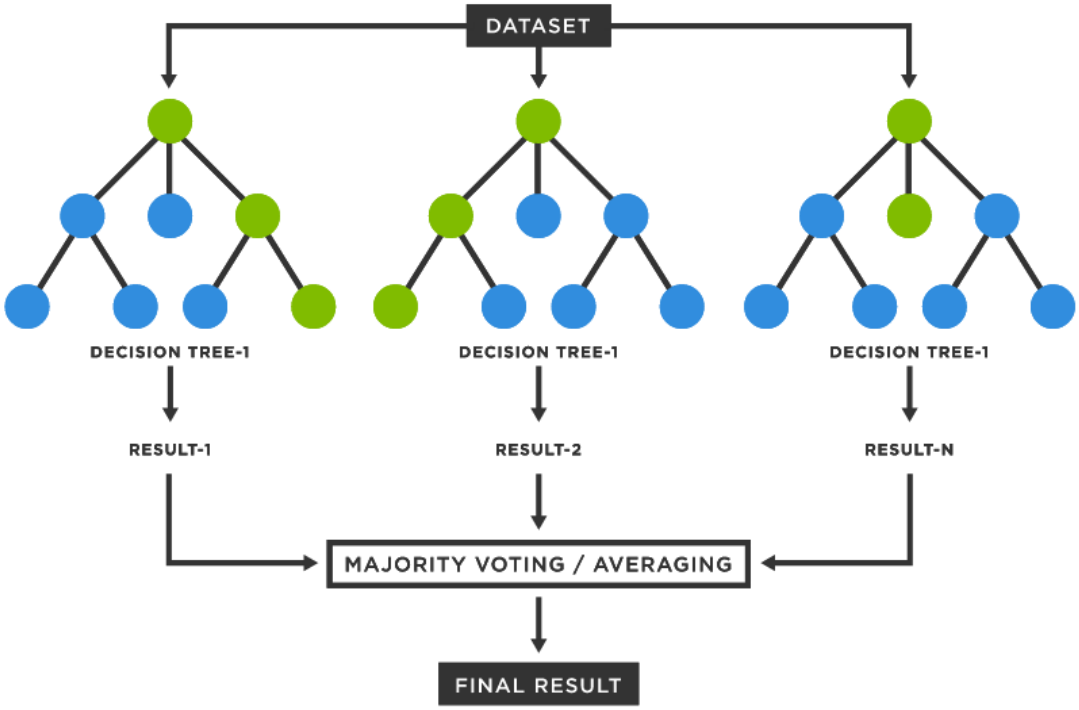
Βήμα 2: Κατασκευάζεται ένα δέντρο για κάθε δείγμα δεδομένων.

Βήμα 3: Προβλέπεται το αποτέλεσμα που λαμβάνεται από κάθε δέντρο αποφάσεων.

Βήμα 4: Υπολογίζονται οι «ψήφοι» για κάθε έναν στόχο από κάθε δέντρο.

Βήμα 5: Τελική πρόβλεψη του αλγορίθμου των Τυχαίων Δασών θεωρείται ο στόχος με τις περισσότερες «ψήφους».

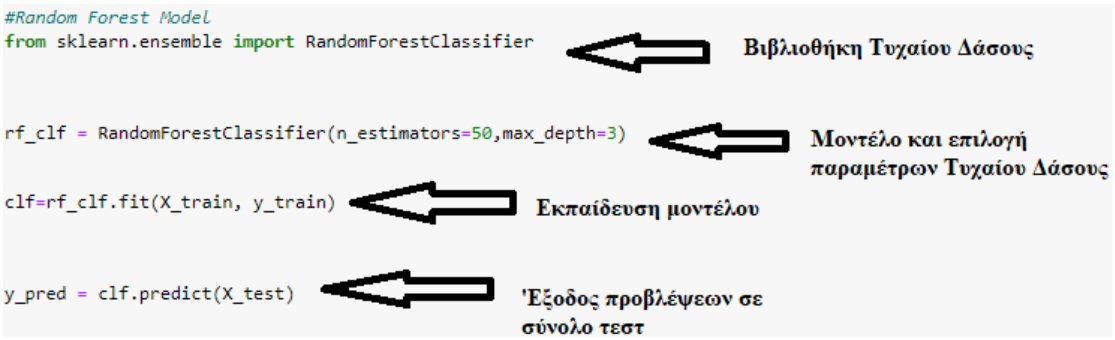
Πίνακας 3:Ψευδοκώδικας του αλγορίθμου Τυχαίο Δάσος



Εικόνα 7: Διάγραμμα ροής του αλγορίθμου Τυχαίο Δάσος

Στον αλγόριθμο Τυχαίο Δάσος χρησιμοποιείται ως υπερπαράμετρος ο αριθμός των δέντρων που θα δημιουργηθούν καθώς και το βάθος που μπορεί να φτάσει το δέντρο. Όσον αφορά τον αριθμό δέντρων, όσο μεγαλύτερος είναι δημιουργούνται πιο πολλά δέντρα απόφασης που μπορεί να οδηγήσουν σε πιο αξιόπιστα αποτελέσματα.(Mayumietal., 2012) Ως προς το βάθος του δέντρου, χαμηλός

αριθμός οδηγεί σε χαμηλή ακρίβεια του μοντέλου ενώ ο πολύ μεγάλος αριθμός μπορεί να οδηγήσει στην υπερπροσαρμογή και να μη δουλεύει καλά στα δεδομένα του τεστ (Bramer, 2016). Ο αλγόριθμος δέχεται σαν είσοδο τις ανεξάρτητες μεταβλητές που επιλέχθηκαν, την εξαρτημένη μεταβλητή της κατηγοριοποίησης (πχ. Ικανοποιημένος ή δυσαρεστημένος επιβάτης), τον αριθμό των δέντρων που θα δημιουργηθούν και το μέγιστο βάθος που μπορεί να φτάσει το δέντρο. Τα δεδομένα αυτά ανήκουν στο σύνολο εκπαίδευσης. Για να γίνει ο αλγόριθμος χρειάζεται ο αλγόριθμος RandomForestClassifier από τη βιβλιοθήκη sklearn.ensemble. Ως έξοδος λαμβάνονται οι αποφάσεις που προτείνει το μοντέλο στο σύνολο τεστ. (Koehrsen, 2018)



Εικόνα 8: Κώδικας και επεξηγήσεις αλγορίθμου Τυχαίο Δάσος

Στον αλγόριθμο Τυχαίο Δάσος χρησιμοποιείται ως επιπρόσθετη υπερπαράμετρος ο αριθμός των δέντρων που θα δημιουργηθούν, που όσο μεγαλύτερος είναι, δημιουργούνται πιο πολλά δέντρα απόφασης που μπορεί να οδηγήσουν σε πιο αξιόπιστα αποτελέσματα. Ο αλγόριθμος δέχεται σαν είσοδο τις 4 ανεξάρτητες μεταβλητές που επιλέχθηκαν στο προηγούμενο βήμα, την εξαρτημένη μεταβλητή της ικανοποίησης του επιβάτη και το μέγιστο βάθος που μπορεί να φτάσει το δέντρο που είναι 6, μεγαλύτερο από αυτό που χρησιμοποιήθηκε καθώς και τον αριθμό δέντρων που θα δημιουργηθούν που είναι 100 προκειμένου να αυξηθεί η ορθότητα του αλγορίθμου (Bharathidasonetal., 2014).Το μοντέλο χτίζεται βάσει του συνόλου εκπαίδευσης. Ως έξοδος λαμβάνονται οι προβλέψεις που προτείνει το μοντέλο στο σύνολο τεστ.

4.4: ΒΗΜΑ 5ο – ΑΞΙΟΛΟΓΗΣΗ

Σε αυτό το βήμα αξιολογούνται τα αποτελέσματα των μοντέλων χρησιμοποιώντας μετρικές που διαθέτουν οι τεχνικές εποπτευόμενης μάθησης. Στην ενότητα αυτή θα παρουσιαστούν συνοπτικά οι μετρικές αυτές ενώ τα αποτελέσματα τους θα αναλυθούν και θα αξιολογηθούν στην επόμενη ενότητα της Ανάλυσης Δεδομένων και Ερμηνείας Αποτελεσμάτων. Στο σχετικό Παράρτημα 7 υπάρχουν πληροφορίες σχετικά με το τι εκφράζουν οι μετρικές αυτές και τον κώδικα που χρειάζεται να συνταχθεί για την εκτέλεσή τους.

Παρακάτω παρουσιάζονται μετρικές για να αξιολογηθεί η απόδοση των μοντέλων (Barkved, 2022):

- Ορθότητα: η ορθότητα μπορεί να δείξει αν ένα μοντέλο εκπαιδεύτηκε σωστά και μπορεί να αποδώσει καλά.

Ωστόσο δεν είναι αντιπροσωπευτικό όταν υπάρχει ανισορροπία μεταξύ της απόδοσης του μοντέλου στις 2 διαφορετικές τάξεις που υπάρχουν. Γι΄αυτό, υπάρχουν και άλλες μετρικές που εστιάζουν σε κάθε μία τάξη ξεχωριστά.

- Πίνακας σύγχυσης: Η προβλεπόμενη κατηγοριοποίηση έναντι της πραγματικής κατηγοριοποίησης αναπαρίσταται σε ένα πίνακα που λέγεται πίνακας σύγχυσης.

- Αναφορά ταξινόμησης: Παρέχει πληροφορίες για κάθε μια τάξη (ικανοποίηση ή δυσαρέσκεια) χρησιμοποιώντας τους δείκτες ακρίβειας , ανάκλησης και F1-score.

Όσον αφορά την ακρίβεια, για κάθε τάξη ελέγχονται πόσες προβλέψεις της τάξης αυτής προβλέφθηκαν σωστά προς το σύνολο των πραγματικών τιμών αυτής της τάξης.

Όσον αφορά την ανάκληση, προκύπτει ως οι σωστές προβλέψεις μιας τάξης προς το σύνολο των προβλέψεων αυτής της τάξης.

Όσον αφορά το scoreF1, είναι μια μετρική που συνδυάζει ακρίβεια και ανάκληση.

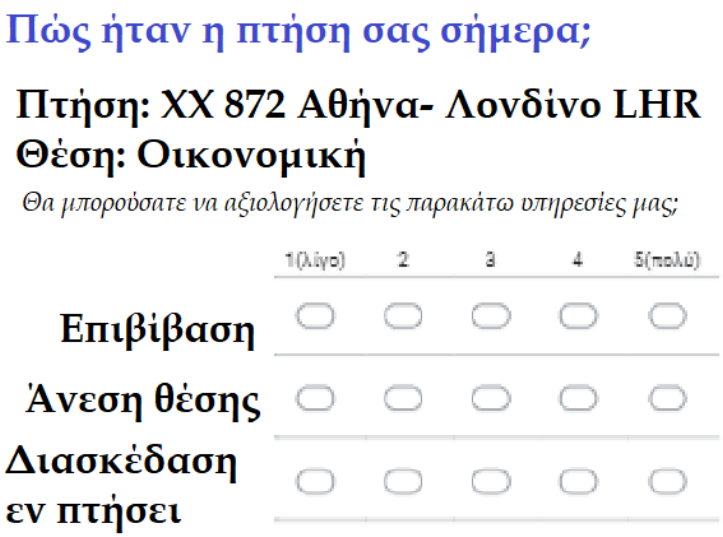
4.5: ΒΗΜΑ 6ο – ΕΦΑΡΜΟΓΗ

Σε αυτό το βήμα περιγράφονται οι ενέργειες εφαρμογής του μοντέλου.

Βάσει των αρχικών δεδομένων υλοποιείται ο αλγόριθμος - όπως παρουσιάστηκε στις προηγούμενες παραγράφους- από την ομάδα αναλυτών της αεροπορικής εταιρείας προκειμένου να βρεθούν οι σημαντικότεροι παράγοντες που επηρεάζουν την ικανοποίηση του επιβάτη και να δημιουργηθεί το μοντέλο πρόβλεψης.

Κατόπιν, θα είναι εφικτό να δίνονται σύντομα ερωτηματολόγια σε κάθε επιβάτη μετά το πέρας της πτήσης που θα επιτρέπουν στον επιβάτη να αξιολογήσει τις 3 υπηρεσίες που θεωρούνται ως οι σημαντικότεροι παράγοντες

που επηρεάζουν την ικανοποίηση του επιβάτη. Οι επιβάτες θα λαμβάνουν email ή SMS που έχουν δηλώσει. Στο ερωτηματολόγιο θα είναι αποθηκευμένες οι πληροφορίες του επιβάτη όπως το είδος θέσης ή το είδος δρομολογίου. Οπότε ένα μικρής έκτασης ερωτηματολόγιο που θα αποστέλλεται στο κινητό του επιβάτη είναι εύκολο να απαντηθεί και να συγκεντρώνεται μεγάλο δείγμα απαντήσεων.



Εικόνα 9: Μορφή σύντομου ερωτηματολογίου όπως θα αποστέλλεται στους επιβάτες

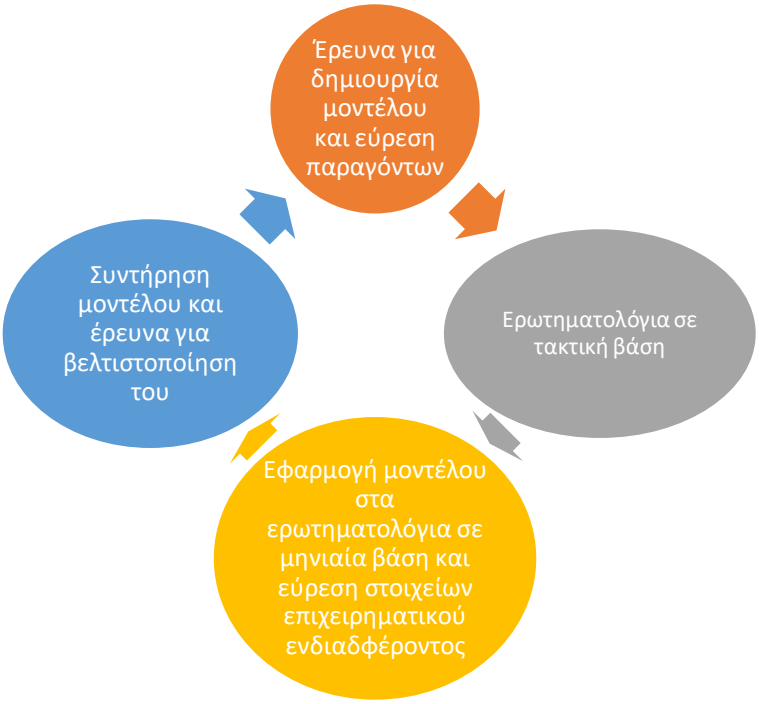
Σε μηνιαίο επίπεδο, θα μπορούν να συγκεντρώνονται όλες οι απαντήσεις των ερωτηματολογίων των πτήσεων του προηγούμενου μήνα. Έτσι, θα εξάγονται συμπεράσματα σχετικά με τα ποσοστά ικανοποίησης των επιβατών σε μηνιαίο επίπεδο αξιοποιώντας το μοντέλο. Ταυτόχρονα, θα μπορούν να βρεθούν στατιστικά στοιχεία βάση των δεδομένων σχετικά με κάθε υπηρεσία ξεχωριστά. Τη δημιουργία των ερωτηματολογίων, την αποστολή και τη συγκέντρωση των απαντήσεων θα τα αναλάβει το τμήμα πληροφορικής της εταιρείας.

Online boarding	Seat comfort	Inflight entertainment	Class	prediction
2	2	2	Eco	neutral or dissatisfied
2	3	3	Eco	neutral or dissatisfied
4	4	4	Business	satisfied
3	4	2	Eco Plus	neutral or dissatisfied
5	5	5	Business	satisfied
4	4	4	Business	satisfied
4	5	4	Business	satisfied
5	5	5	Business	satisfied
1	4	4	Eco Plus	neutral or dissatisfied
5	5	4	Business	satisfied
5	5	5	Business	satisfied
3	1	4	Business	satisfied
0	2	2	Business	neutral or dissatisfied
1	4	4	Eco	neutral or dissatisfied
3	3	3	Eco	neutral or dissatisfied
4	3	3	Business	satisfied

Εικόνα 10: Προβολή των απαντήσεων επιβατών και της πρόβλεψης που δίνει το μοντέλο σε Φύλλο εργασίας Excel

Βασικό στοιχείο της διαδικασίας είναι η συντήρηση της λύσης του προβλήματος. Αυτό μπορεί να γίνεται με χρήση αναλυτικών ερωτηματολογίων σε ετήσια βάση και σε ορισμένους επιβάτες, επανάληψη της διαδικασίας δημιουργίας μοντέλου πρόβλεψης καθώς και μελέτη νέων αλγορίθμων που θα μπορούσαν να δώσουν καλύτερα αποτελέσματα.

Ο αλγόριθμος μπορεί να εκτελείται με τη γλώσσα προγραμματισμού Python ενώ η εξαγωγή και οπτικοποίηση αποτελεσμάτων μπορεί να γίνει στο MicrosoftExcel.



Εικόνα 11: Σχηματική περιγραφή της διαδικασίας εφαρμογής του μοντέλου

ΚΕΦΑΛΑΙΟ 5: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΡΜΗΝΕΙΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στην ενότητα αυτή θα αναλυθούν τα δεδομένα του υπάρχοντος συνόλου δεδομένων ακολουθώντας τη διαδικασία που περιγράφηκε στην ενότητα 4. Συγκεκριμένα θα παρουσιαστούν στοιχεία και αποτελέσματα από την ανάλυση δεδομένων και όπου χρήζει θα δοθεί και ερμηνεία των αποτελεσμάτων.

5.1: ΠΡΩΤΗ ΕΞΕΡΕΥΝΗΣΗ ΔΕΔΟΜΕΝΩΝ

Αξιοποιώντας τα στατιστικά των δεδομένων και τις τιμές των συσχετίσεων προκύπτουν μερικά βασικά ευρήματα:

- Υπάρχουν συνολικά 129.880 απαντήσεις επιβατών σε ερωτήσεις σχετικά με την ικανοποίησή τους κατά τη διάρκεια της πτήσης. Όσον αφορά τις στήλες (μεταβλητές), υπάρχουν 5 κατηγορικές μεταβλητές (όπως φύλο και το είδος θέσης) και 17 αριθμητικές μεταβλητές όπως οι βαθμολογίες από την αξιολόγηση των υπηρεσιών και η καθυστέρηση αναχώρησης ή άφιξης σε λεπτά της ώρας.

- Το 43% των επιβατών που απάντησαν στο ερωτηματολόγιο δηλώνουν ικανοποιημένοι.

- Όσον αφορά τη μέση καθυστέρηση στην αναχώρηση της πτήσης υπολογίζεται στα 14,71 λεπτά ενώ όσον αφορά τη μέση καθυστέρηση στην άφιξη της πτήσης υπολογίζεται στα 15,09 λεπτά, λίγο παραπάνω από την αναχώρηση.

- Λαμβάνοντας υπόψη τη μέση βαθμολογία των υπηρεσιών από όλους τους επιβάτες μεγαλύτερη επίδοση φαίνεται να έχουν οι υπηρεσίες εν πτήσει, ο χειρισμός αποσκευών και η άνετη θέση και λαμβάνουν τιμή γύρω στο 3,5.

Αντίθετα, χαμηλότερες τιμές μικρότερο του 3, έχουν οι υπηρεσίες όπως wifi στο αεροσκάφος, τοποθεσία εξόδου αναχώρησης και η διαδικασία κράτησης διαδικτυακά.

- Πιο ισχυρή συσχέτιση μεταξύ των αριθμητικών μεταβλητών πρόκειται αυτής της καθυστέρησης ώρας αναχώρησης με αυτή της ώρας άφιξης που θεωρείται προφανής και λαμβάνει τιμή κοντά στο 1. Θετική συσχέτιση παρατηρείται με τιμές γύρω 0,7 μεταξύ καθαριότητας και άνεσης θέσης όπως επίσης μεταξύ καθαριότητας και διασκέδασης εν πτήσει. Όλες οι υπόλοιπες συσχετίσεις είναι ασθενείς, κοντά στο 0.

5.2: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΕΥΡΕΣΗ ΣΗΜΑΝΤΙΚΟΤΕΡΩΝ ΠΑΡΑΓΟΝΤΩΝ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΗΝ ΙΚΑΝΟΠΟΙΗΣΗ ΤΟΥ ΕΠΙΒΑΤΗ

Με βάση τα στατιστικά τεστ που πραγματοποιήθηκαν προκύπτουν κατηγορικές και αριθμητικές μεταβλητές που επηρεάζουν περισσότερο την ικανοποίηση του επιβάτη. Αυτές συνοψίζονται στον παρακάτω πίνακα:

Κατηγορικές Μεταβλητές	**Αριθμητικές μεταβλητές**
✓ Είδος Θέσης	✓ Επιβίβαση ✓ Άνεση Θέσης ✓ Διασκέδαση εν πτήσει

Πίνακας 4:Σημαντικότεροι Παράγοντες που επηρεάζουν την ικανοποίηση ενός επιβάτη μιας πτήσης

Αξίζει να σημειωθεί εδώ το είδος θέσης μπορεί να επηρεάσει σημαντικά την ικανοποίηση του επιβάτη. Πράγματι, όσο καλύτερο είναι το είδος θέσης π.χ. Διακεκριμένη, προσφέρονται καλύτερες υπηρεσίες και κατ' επέκταση είναι πιθανότερη η ικανοποίηση του επιβάτη.

5.3: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕΣΑ ΑΠΟ ΤΑ ΜΟΝΤΕΛΑ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Βάση τη δημιουργία του Δέντρου Απόφασης, διαπιστώνονται τα εξής αποτελέσματα σχετικά με την κατηγοριοποίηση ενός επιβάτη ως ευχαριστημένος ή δυσαρεστημένος:

- Επιβάτες της EcoPlus ή της Business,που έχουν αξιολογήσει με τουλάχιστον 4/5 τη διασκέδαση εν πτήσει όπως επίσης και την επιβίβαση είναι κατά βάση ικανοποιημένοι.

- Επιβάτες της Οικονομικής θέσης, που έχουν αξιολογήσει την επιβίβαση με βαθμό λιγότερο από 3/5 θεωρούνται κατά βάση δυσαρεστημένοι ή ουδέτεροι επιβάτες.

5.4: ΑΝΑΛΥΣΗ ΜΟΝΤΕΛΩΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΜΕΣΩ ΜΕΤΡΙΚΩΝ

Θα αναλυθούν τα αποτελέσματα των αλγορίθμων Δέντρο Απόφασης και Τυχαίο Δάσος καθενός ξεχωριστά ως προς τις μετρικές αξιολόγησης και κατόπιν θα γίνει σύγκριση τους. Παράλληλα θα ερμηνευθούν τα αποτελέσματα των μετρικών σε σχέση με επιχειρηματικές ανάγκες.

5.4.1: ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ

Οι τεχνικές εποπτευόμενης μάθησης διαθέτουν κατάλληλες μετρικές για να αξιολογηθεί η απόδοση του μοντέλου πρόβλεψης οι οποίες θα παρουσιαστούν παρακάτω:

- Ορθότητα

Το ποσοστό ορθότητας του μοντέλου είναι 83%, που δείχνει ότι οι περισσότερες προβλέψεις έγιναν σωστά στο σύνολο τεστ. Έτσι, φαίνεται ότι το μοντέλο εκπαιδεύτηκε σωστά και μπορεί να αποδώσει καλά.
Παρακάτω θα μελετηθεί η απόδοση του μοντέλου για καθεμία τάξη ξεχωριστά. Γι΄αυτό, θα χρησιμοποιηθούν άλλες μετρικές που εστιάζουν ξεχωριστά στην πρόβλεψη δυσαρεστημένων επιβατών και στην πρόβλεψη ευχαριστημένων επιβατών.

- Πίνακας σύγχυσης

Στον πίνακα σύγχυσης, φαίνεται ποιοι επιβάτες τοποθετήθηκαν σωστά και ποιοι όχι. Στη διαγώνιο του πίνακα φαίνονται οι επιβάτες που τοποθετήθηκαν σωστά.
Για παράδειγμα όσον αφορά τους δυσαρεστημένους επιβάτες, φαίνεται ότι 18760 επιβάτες τοποθετήθηκαν σωστά και είναι πραγματικά δυσαρεστημένοι ενώ 3420 τοποθετήθηκαν στους δυσαρεστημένους ενώ στην πραγματικότητα είναι ικανοποιημένοι. Όμως θα ληφθούν υπόψη ποσοστιαίες μετρικές που παρέχονται στην αναφορά ταξινόμησης παρακάτω.

- Αναφορά ταξινόμησης

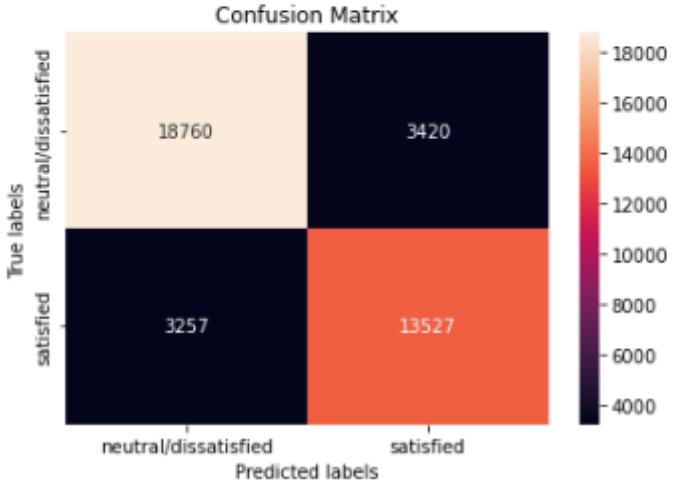
Όσον αφορά την ακρίβεια, για κάθε τάξη ελέγχονται πόσες προβλέψεις της τάξης αυτής προβλέφθηκαν σωστά προς το σύνολο των πραγματικών τιμών αυτής της τάξης. Είναι σημαντικό να αξιολογηθεί όταν το κόστος των λανθασμένων προβλέψεων της τάξης είναι μεγάλο. Στην περίπτωση αυτή θεωρώντας ότι δε θα έπρεπε να εμφανίζεται ως ικανοποιημένος επιβάτης ένας που είναι δυσαρεστημένος, χρειάζεται οι περισσότεροι πραγματικά δυσαρεστημένοι επιβάτες να έχουν προβλεφθεί σωστά, δηλαδή η ακρίβεια των δυσαρεστημένων επιβατών να είναι υψηλή. Πράγματι είναι, αφού αγγίζει ποσοστό 85%.

Όσον αφορά την ανάκληση, προκύπτει ως οι σωστές προβλέψεις μιας τάξης προς το σύνολο των προβλέψεων αυτής της τάξης. Αντίστοιχα με πριν, θεωρείται ότι το κόστος να εμφανίζεται ως ικανοποιημένος επιβάτης ένας που είναι δυσαρεστημένος είναι μεγάλο. Γι' αυτό εξετάζεται η τάξη των ικανοποιημένων επιβατών και συγκεκριμένα αυτών που έχουν τοποθετηθεί λανθασμένα, που θα πρέπει να είναι λίγοι. Αυτό επιτυγχάνεται με υψηλή τιμή ανάκλησης που αντίστοιχα εξασφαλίζεται αφού φτάνει το ποσοστό 81%.

Όσον αφορά το score F1, είναι μια μετρική που συνδυάζει ακρίβεια και ανάκληση. Για παράδειγμα ένα καλό F1 score για τους ικανοποιημένους επιβάτες συμβαίνει όταν οι επιβάτες που τοποθετήθηκαν λανθασμένα ως ικανοποιημένοι είναι λίγοι αλλά και οι επιβάτες που τοποθετήθηκαν ως δυσαρεστημένοι αλλά στην πραγματικότητα είναι ικανοποιημένοι είναι λίγοι. Έτσι συμβαίνει με το αποτέλεσμα 0.80.

	precision	recall	f1-score	support
neutral or dissatisfied	0.85	0.85	0.85	22180
satisfied	0.80	0.81	0.80	16784
accuracy			0.83	38964
macro avg	0.83	0.83	0.83	38964
weighted avg	0.83	0.83	0.83	38964

```
[Text(0, 0.5, 'neutral/dissatisfied'), Text(0, 1.5, 'satisfied')]
```



Εικόνα 12: Προβολή αποτελεσμάτων αξιολόγησης Δέντρου Απόφασης

5.4.2: ΤΥΧΑΙΟ ΔΑΣΟΣ

Το Τυχαίο Δάσος ως αλγόριθμος εποπτευόμενης μάθησης και βασιζόμενος στο Δέντρο Απόφασης θα αξιολογηθεί ως προς τις ίδιες μετρικές του Δέντρου Απόφασης:

- Ορθότητα

Για να φανεί αν το μοντέλο εκπαιδεύτηκε σωστά και μπορεί να αποδώσει καλά χρησιμοποιείται η ορθότητα όπως και στο Δέντρο Απόφασης. Πράγματι, προκύπτει ποσοστό ορθότητας 86%.

Και εδώ ελέγχεται όταν υπάρχει ανισορροπία μεταξύ της απόδοσης του μοντέλου στις 2 διαφορετικές τάξεις που υπάρχουν. Έτσι ελέγχονται και οι άλλες μετρικές που εστιάζουν σε κάθε μία τάξη ξεχωριστά: στην πρόβλεψη δυσαρεστημένων επιβατών και στην πρόβλεψη ευχαριστημένων επιβατών.

- Πίνακας σύγχυσης

Για κάθε τάξη, φαίνεται ποιοι επιβάτες τοποθετήθηκαν σωστά και ποιοι όχι.

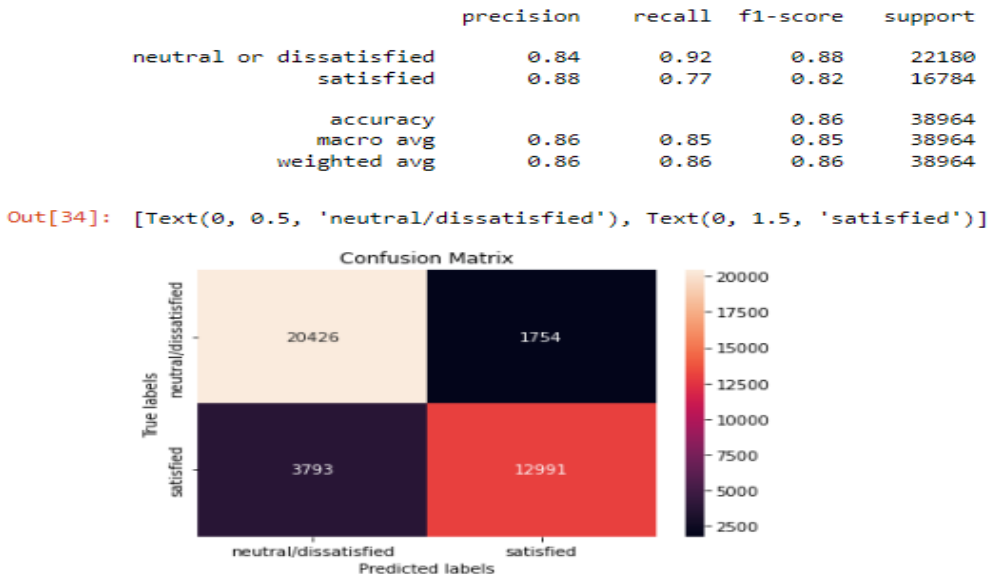
Εδώ, όσον αφορά τους δυσαρεστημένους επιβάτες, φαίνεται ότι 20426 επιβάτες τοποθετήθηκαν σωστά και είναι πραγματικά δυσαρεστημένοι ενώ 1754 τοποθετήθηκαν στους δυσαρεστημένους ενώ στην πραγματικότητα είναι ικανοποιημένοι. Όμως θα ληφθεί υπόψη και το αντίστοιχο ποσοστό τους με δείκτες που παρέχονται στην αναφορά ταξινόμησης παρακάτω.

- Αναφορά ταξινόμησης

Όσον αφορά την ακρίβεια, όπως και στο Δέντρο Απόφασης, θεωρείται ότι δε θα έπρεπε να θεωρούνται ικανοποιημένοι επιβάτες που δεν είναι στην πραγματικότητα. Έτσι χρειάζεται οι περισσότεροι πραγματικά δυσαρεστημένοι επιβάτες να έχουν προβλεφθεί σωστά, δηλαδή η ακρίβεια των δυσαρεστημένων επιβατών να είναι υψηλή. Πράγματι είναι, αφού αγγίζει ποσοστό 84%.

Όσον αφορά την ανάκληση, αντίστοιχα με πριν, θεωρείται ότι το κόστος τοποθέτησης ενός επιβάτη ως ικανοποιημένο που στην πραγματικότητα είναι δυσαρεστημένος, είναι μεγάλο. Γι' αυτό εξετάζονται οι ικανοποιημένοι επιβάτες και συγκεκριμένα αυτοί που έχουν τοποθετηθεί εκεί λανθασμένα, που θα πρέπει να είναι λίγοι. Αυτό επιτυγχάνεται με υψηλή τιμή ανάκλησης που αντίστοιχα εξασφαλίζεται αφού φτάνει το ποσοστό 77%.

Όσον αφορά το score F1, είναι καλό για τους ικανοποιημένους επιβάτες γιατί οι επιβάτες που θεωρήθηκαν ικανοποιημένοι λανθασμένα είναι λίγοι αλλά και οι επιβάτες που θεωρήθηκαν δυσαρεστημένοι αλλά είναι πραγματικά ικανοποιημένοι είναι λίγοι. Έτσι συμβαίνει με το αποτέλεσμα 0.82.



Εικόνα 13: Προβολή αποτελεσμάτων αξιολόγησης Τυχαίου Δάσους

5.4.3: ΣΥΓΚΡΙΣΗ

Για να συγκριθούν τα μοντέλα Δέντρο Απόφασης και Τυχαίο Δάσος θα παρουσιαστούν συνοπτικά οι τιμές των μετρικών ακρίβειας, ορθότητας και ανάκλησης στον παρακάτω πίνακα. Στο τέλος βάση των τιμών αυτών θα αποφασιστεί ποιο μοντέλο είναι πιο βέλτιστο.

ΜΕΤΡΙΚΕΣ	ΔΕΝΤΡΟ ΑΠΟΦΑΣΗΣ	ΤΥΧΑΙΟ ΔΑΣΟΣ
ΟΡΘΟΤΗΤΑ	83%	86%
ΑΚΡΙΒΕΙΑ	85%	84%
ΑΝΑΚΛΗΣΗ	81%	77%

Πίνακας 5: Σύγκριση αποτελεσμάτων των 2 μοντέλων

Παρατηρούνται μικροδιαφορές στις τιμές των 2 μοντέλων με τιμές άνω του 80% σχεδόν σε όλες τις περιπτώσεις. Ωστόσο καθώς δίνεται προτεραιότητα στην σωστή πρόβλεψη των δυσαρεστημένων επιβατών προκειμένου να βρεθούν τρόποι βελτίωσης των υπηρεσιών, επιλέγεται η μέθοδος του Δέντρου Απόφασης. Συγκεκριμένα, δίνεται έμφαση στις τιμές ακρίβειας των δυσαρεστημένων επιβατών και ανάκλησης των ικανοποιημένων επιβατών όπως εξηγήθηκαν στις παραγράφους 5.4.1 και 5.4.2, που λαμβάνουν μεγαλύτερες τιμές στο μοντέλο του Δέντρου Απόφασης σε σχέση με το μοντέλο του Τυχαίου Δάσους.

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ-ΠΡΟΤΑΣΕΙΣ

Στην τελευταία ενότητα θα παρουσιαστούν οι ενέργειες που έγιναν για την επίλυση του προβλήματος, θα αναπτυχθεί η γνώση που αναπτύχθηκε μέσω της επίλυσης αυτού του προβλήματος που έχει επιχειρηματικό ενδιαφέρον καθώς και θα αναφερθούν περαιτέρω ενέργειες που θα μπορούσαν να βελτιώσουν τη λύση του προβλήματος.

6.1: ΕΠΙΛΥΣΗ ΠΡΟΒΛΗΜΑΤΟΣ

Μελετώντας τις ανάγκες των αεροπορικών εταιρειών καθώς και αξιοποιώντας τις υπολογιστικές ικανότητες τις γλώσσας προγραμματισμού Pythonόπως επίσης και τα μοντέλα της Μηχανικής Μάθησης επιτεύχθηκαν τα ακόλουθα:

- Δημιουργία Ερωτηματολογίου που επιτρέπει την αξιολόγηση των υπηρεσιών από τους επιβάτες και συγκέντρωση των δεδομένων
- Επεξεργασία των δεδομένων αυτών με χρήση της γλώσσας προγραμματισμού Python
- Εύρεση στοιχείων με επιχειρηματικό ενδιαφέρον με χρήση κατάλληλων συναρτήσεων της Python
- Εφαρμογή στατιστικών τεστ για την εύρεση των σημαντικότερων παραγόντων που επηρεάζουν την ικανοποίηση του επιβάτη
- Υλοποίηση μοντέλων Μηχανικής Μάθησης για την πρόβλεψη της ικανοποίησης του επιβάτη δεδομένου των σημαντικότερων παραγόντων που επιλέχθηκαν στο προηγούμενο βήμα
- Αξιολόγηση των μοντέλων και επιλογή του βέλτιστου
- Δημιουργία σύντομων ερωτηματολογίων με 3 μόνο ερωτήσεις προς απάντηση από τους επιβάτες και εφαρμογή του μοντέλου για την πρόβλεψη αν είναι ο επιβάτης ικανοποιημένος ή όχι

6.2: ΕΞΑΓΩΓΗ ΓΝΩΣΗΣ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΕΝΔΙΑΦΕΡΟΝΤΟΣ

Γενικά, παρατηρείται ότι το ποσοστό ικανοποίησης επιβατών είναι χαμηλό κάτω από το 50%. Αυτό δείχνει ότι κάποιες υπηρεσίες της αεροπορικής εταιρείας χρήζουν βελτίωσης καθώς η ικανοποίηση του επιβάτη είναι ιδιαίτερης σημασίας αυτήν την περίοδο που ο ανταγωνισμός αυξάνεται με την είσοδο νέων εταιρειών στην αγορά (Scheffler, 2018).

Από τα δεδομένα, μπορούν να κατανοηθούν σε ποιες υπηρεσίες πρέπει να βρει τρόπους η εταιρεία για να τις αναβαθμίσει. Αυτό φαίνεται από τις χαμηλές βαθμολογίες. Συγκεκριμένα, πρέπει να βελτιωθεί η χρήση του wifi στην πτήση πράγμα που σημαίνει ότι είτε μπορεί να υπολειτουργεί, είτε να μην παρέχεται στα αεροσκάφη. Παράλληλα, σε συνεργασία με τα αεροδρόμια πρέπει να βελτιωθεί η προσβασιμότητα στις εξόδους αναχώρησης. Η τοποθέτηση σχετικών ενδείξεων σε όλους τους χώρους του αεροδρομίου, η μείωση των αποστάσεων από την είσοδο στον αερόσταθμό μέχρι την έξοδο αναχώρησης όπως και η χρήση εσωτερικών λεωφορείων από τα σημεία ελέγχου επιβατών που να οδηγούν στις εξόδους αναχώρησης είναι τρόποι βελτίωσης της προσβασιμότητας στις εξόδους αναχώρησης. Τέλος, θα πρέπει να δημιουργηθεί ένας εύκολος και ασφαλής τρόπος κράτησης εισιτηρίων διαδικτυακά. Η εταιρεία μπορεί να αναζητήσει εύχρηστα συστήματα κρατήσεων από διάφορους παρόχους (Menze, 2022).

Όσον αφορά τα δυνατά σημεία της αεροπορικής εταιρείας είναι οι υπηρεσίες εν πτήσει, ο χειρισμός των αποσκευών και η άνεση των θέσεων. Σε όλα αυτά πρέπει να διατηρηθεί ή και να αναβαθμιστεί η ποιότητά τους. Λαμβάνοντας υπόψη τους σημαντικότερους παράγοντες που επηρεάζουν την ικανοποίηση της πτήσης, πρέπει να δίνεται προτεραιότητα στις εξής υπηρεσίες προτείνοντας καινοτομίες και βελτιώσεις : η επιβίβαση που μπορεί να βελτιωθεί με χρήση προτεραιοτήτων είδος θέσης (διακεκριμένη ή οικονομική), οικογένειες με παιδιά, τακτικοί επιβάτες, αριθμός σειράς καθίσματος κτλ (Kisiel, 2020), η θέση με επένδυση σε υψηλής ποιότητας καθίσματα και η διασκέδαση εν πτήσει με ανανέωση υλικού ταινιών, μουσικής και παιχνιδιών σε τακτική βάση.

6.3: ΤΡΟΠΟΙ ΒΕΛΤΙΩΣΗΣ

Για την επίλυση του προβλήματος ικανοποίησης των επιβατών μπορούν να γίνουν μικρότερες ομαδοποιήσεις ανάλογα με το προϊόν που παρέχεται.

Έτσι μπορούν να δημιουργηθούν ξεχωριστά μοντέλα για τους επιβάτες Οικονομικής θέσης και για τους επιβάτες Διακεκριμένης θέσης. Οπότε για κάθε κατηγορία θέσης, θα εξετάζονται διαφορετικοί παράμετροι που επηρεάζουν την ικανοποίηση του επιβάτη και θα εξάγονται διαφορετικά ευρήματα επιχειρηματικού ενδιαφέροντος.

Αυτή η ομαδοποίηση μπορεί να υποδιαιρεθεί περαιτέρω με προσθήκη το είδος πτήσης (πτήση εσωτερικού ή πτήση εξωτερικού), δεδομένου ότι οι ανάγκες είναι διαφορετικές. Για παράδειγμα, σε μια πτήση εσωτερικού διάρκειας 30 λεπτών, δεν χρειάζεται η προσφορά υπηρεσιών διασκέδασης εν πτήσει, ενώ σε μια πολύωρη πτήση κρίνεται απαραίτητο (Hawk, 2018).

Οπότε με αυτές τις ενέργειες θα μελετάται ξεχωριστά κάθε κατηγορία πτήσεων/θέσεων αντί όλες μαζί. Δηλαδή επιβάτες οικονομικής θέσης εσωτερικού, επιβάτες οικονομικής θέσης εξωτερικού, επιβάτες διακεκριμένης θέσης εσωτερικού, επιβάτες διακεκριμένης θέσης εξωτερικού κτλ.

ΒΙΒΛΙΟΓΡΑΦΙΑ/ΑΡΘΟΓΡΑΦΙΑ

Άρθρα σε περιοδικό:

1. Namukasa, J. (2013), "The influence of airline service quality on passenger satisfaction and loyalty: The case of Uganda airline industry", The TQM Journal, Vol. 25 No. 5, pp. 520-532.

2. Schröer, Christoph & Kruse, Felix & Marx Gómez, Jorge. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model, Procedia Computer Science, 181 pp. 526-534

3.B.S, Kevin & Lestari, Anggun & Pratiwi, Sekar. (2018). An Analysis of Airlines Customer Satisfaction by improving Customer Service Performance, Advances in Engineering Research (AER), volume 147

4. Shouvik Sanyal1, Mohammed Wamique Hisam (2016). An Analysis of the Impact of Service Quality and Passenger Satisfaction on Passenger Preferences for Airlines: A Study of the Indian Aviation Sector, International Review of Management and Marketing, 6(2), 354-357

5. Paulo T.C. Freitas, Lenice M. Silva, Marcus V. Nascimento, Giovanna M.R. Borille (2021). Passenger profile and its effects on satisfaction level in food and beverage establishments: Case study of major Brazilian airports, Case Studies on Transport Policy,Volume 9, Issue 3, Pages 1219-1224

6.T R, Prajwala. (2015), A Comparative Study on Decision Tree and Random Forest Using R Tool, IJARCCE, 196-199

7. Rupal Snehkunj, Khushboo Vachiyatwala. (2022), Data Analysis Using Pandas Library of Python, Acta Scientific COMPUTER SCIENCES, Volume 4 Issue 3

8. Shengjia Cao, Yunhan Zeng, Shangru Yang and Songlin Cao (2021), Research on Python Data Visualization Technology, Journal of Physics: Conference Series, Volume 1757

9. Bramer, Max. (2016). Avoiding Overfitting of Decision Trees,Principles of Data Mining

10.Bharathidason, S. & Venkateswaran, Jothi (2014), Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees, International Journal of Computer Applications, 101, 26-30

11. Tomasz Kisiel (2020), Resilience of passenger boarding strategies to priority fares offered by airlines, Journal of Air Transport Management, Volume 87, 101853

12. Sebastian Raschka, ORCID,Joshua Patterson and Corey Nolet (2020), Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence, Machine Learning in Python, 11(4), 193

13.Lamrini Bouchra (2020), Contribution to Decision Tree Induction with Python: A Review,Data Mining - Methods, Applications and Systems

14.Thais Mayumi Oshiro, Pedro Santoro Perez & José Augusto Baranauskas (2012), How Many Trees in a Random Forest, Lecture Notes in Computer Science,LNAI 7376, pp. 154–168

15. Gerard Biau (2012), Analysis of a Random Forests Model, Journal of Machine Learning Research 13,1063-1095

16.Potdar, Kedar &Pardawala, Taher & Pai, Chinmay (2017), A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications, 175, 7-9

17.Giles Hooker (2007), Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables, Journal of Computational and Graphical Statistics, 16:3, 709-732

18. Korean J Anesthesiol(2013), The prevention and handling of the missing data, Korean Society of Anesthesiologists, 64(5), 402–406

Βιβλία

1.Γεωργούλη, Κ. (2015) Τεχνητή Νοημοσύνη, Μια εισαγωγική προσέγγιση, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Εθνικό Μετσόβιο Πολυτεχνείο.

2. Πρίντεζη, Γ.Φ. (1995) Αεροπορικό Μάρκετινγκ, Εκδόσεις Έλλην.

Ιστοσελίδες

1. TIBCO (2022) What is a random forest? [Online] Available from: https://www.tibco.com/es/reference-center/what-is-a-random-forest#:~:text=Un%20bosque%20aleatorio%20elegir%C3%A1%20caracter%C3%ADsticas,un%20%C3%A1rbol%20de%20decisi%C3%B3n%20individual. [Accessed: 15 October 2022].

2. SKILLX (2020) Advantages and Disadvantages of Decision Tree Learning. [Online] Available from: https://skilllx.com/advantages-and-disadvantages-of-decision-tree-learning/ [Accessed: 10 October 2022].

3. THE PROFESSIONALS POINT (2019) Advantages and Disadvantages of Random Forest Algorithm in Machine Learning. [Online] Available from: http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html#:~:text=Random%20Forest%20is%20based%20on,and%20therefore%20improves%20the%20accuracy

[Accessed: 10 October 2022].

4. EXPLORIUM (2019) The Complete Guide to Decision Tree Analysis. [Online] Available from: https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/ [Accessed: 10 October 2022].

5. COURSERA (2022) What Is Python Used For? A Beginner's Guide. [Online] Available from: https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python [Accessed: 20 October 2022].

6. PCSTEPS (2017) Τι Είναι τα Προγράμματα Ανοικτού Κώδικα / OpenSource. [Online] Available from: https://www.pcsteps.gr/192308-%CF%80%CF%81%CE%BF%CE%B3%CF%81%CE%AC%CE%BC%CE%BC%CE%B1%CF%84%CE%B1-%CE%B1%CE%BD%CE%BF%CE%B9%CE%BA%CF%84%CE%BF%CF%8D-%CE%BA%CF%8E%CE%B4%CE%B9%CE%BA%CE%B1-open-source/ [Accessed: 20 October 2022].

7. JUPYTER NOTEBOOK BEGINNER GUIDE (2015) What is the Jupyter Notebook? [Online] Available from: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html [Accessed: 20 October 2022].

8. JUNI LEARNING (2019) Explore the Role of Coding in Data Science & Analytics [Online] Available from: https://junilearning.com/blog/guide/coding-and-data-science/ [Accessed: 20 October 2022].

9. TOWARDS DATA SCIENCE (2020) Using Python to create static web pages — the easy way! [Online] Available from: https://towardsdatascience.com/using-python-to-create-static-web-pages-the-easy-way-6eb16c997571 [Accessed: 20 October 2022].

10. MEDIUM (2021) Use Python to Send Outlook Emails [Online] Available from: https://medium.com/mlearning-ai/use-python-to-send-outlook-emails-d673ce9e33e4 [Accessed: 20 October 2022].

11. KAGGLE (2019) Airline Passenger Satisfaction [Online] Available from: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction [Accessed: 15 September 2022].

12. NEAR EAST UNIVERSITY (2019) The importance of the roles of law in aviation safety[Online] Available from: http://docs.neu.edu.tr/library/6364947557.pdf [Accessed: 15 September 2022].

13. DIVA PORTAL (2018) The Relationship of Service Quality and Customer Satisfaction in the Airline Industry and the Moderating Effect of the Airline Type [Online] Availablefrom:

http://www.diva-portal.org/smash/get/diva2:1232279/FULLTEXT01.pdf [Accessed: 15 September 2022].

14. MEDIUM (2018) An overview of correlation measures between categorical and continuous variables [Online] Available from: https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365 [Accessed: 10 October 2022].

15. MACHINE LEARNING MASTERY (2020) Train-Test Split for Evaluating Machine Learning Algorithms [Online] Available fromhttps://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/ [Accessed: 10 October 2022].

16. INDEED (2022) Guide to the Types of Decision Trees in Machine Learning [Online] Available from: https://www.indeed.com/career-advice/career-development/types-of-decision-trees-machine-learning [Accessed: 10 October 2022].

17. KD NUGGETS (2022) Decision Tree Algorithm, Explained [Online] Available from: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html [Accessed: 10 October 2022].

18. UNIVERSITY OF VIRGINIA LAB (2020) Getting Started with pandas in Python [Online] Available from: https://data.library.virginia.edu/getting-started-with-pandas-in-python/ [Accessed: 10 October 2022].

19. PATHMIND (2020) Evaluation Metrics for Machine Learning - Accuracy, Precision, Recall, and F1 Defined. [Online] Available from: http://wiki.pathmind.com/accuracy-precision-recall-f1 [Accessed: 15 November 2022].

20. OBVIOUSLY (2022) How to Know if Your Machine Learning Model Has Good Performance. [Online] Available from: https://www.obviously.ai/post/machine-learning-model-performance [Accessed: 15 November 2022].

21. TOWARDS DATA SCIENCE (2018) Metrics to Evaluate your Machine Learning Algorithm [Online] Available from: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 [Accessed: 15 November 2022].

22. TOWARDS DATA SCIENCE (2017) Decision Trees in Machine Learning. [Online] Available from: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed: 15 November 2022].

23. TOWARDS DATA SCIENCE (2020) Do Decision Trees need Feature Scaling?[Online] Available from: https://towardsdatascience.com/do-decision-trees-need-feature-scaling-97809eaa60c6 [Accessed: 15 November 2022].

24. ANALYTICS VIDHYA (2020) 4 Simple Ways to Split a Decision Tree in Machine Learning

[Online] Available from: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/ [Accessed: 15 November 2022].

25. ΙΔΡΥΜΑΤΙΚΟ ΑΠΟΘΕΤΗΡΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΑΤΡΩΝ (2012) Αναγνώριση και κατάταξηονομάτωνοντοτήτωνσε ελληνικά κείμενα με χρήση Τυχαίων Δασών [Online]

Availablefrom: https://nemertes.library.upatras.gr/server/api/core/bitstreams/1df79e06-bd65-4191-a22b-baee1b1d9386/content [Accessed: 15 November 2022].

26. MACHINE LEARNING MASTERY (2020) How to Perform Feature Selection With Numerical Input Data. [Online] Available from: https://machinelearningmastery.com/feature-selection-with-numerical-input-data/ [Accessed: 20 October 2022].

27. UPGRAD (2021) Data Preprocessing in Machine Learning: 7 Easy Steps To Follow [Online] Available from: https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/ [Accessed: 20 October 2022].

28. MEDIUM (2020) Categorical Feature Selection using Chi- Squared Test. [Online] Available from: https://medium.com/analytics-vidhya/categorical-feature-selection-using-chi-squared-test-e4c0d0af6b7e [Accessed: 15 November 2022].

29. BUILT IN (2022) A Step-by-Step Explanation of Principal Component Analysis (PCA). [Online] Available from: https://builtin.com/data-science/step-step-explanation-principal-component-analysis [Accessed: 15 October 2022].

30. ANALYTICS VIDHYA (2022) Decision Tree Machine Learning Algorithm Using Python. [Online] Available from: https://www.analyticsvidhya.com/blog/2022/03/decision-tree-machine-learning-using-python/ [Accessed: 15 November 2022].

31. ANALYTICS VIDHYA (2022) Understanding Random Forest. [Online] Available from: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ [Accessed: 15 November 2022].

32. TOWARDS DATA SCIENCE (2018) An Implementation and Explanation of the Random Forest in Python. [Online] Available from: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76 [Accessed: 15 November 2022].

33. STATISTICS DEPARTMENT UNIVERSITY OF CALIFORNIA (1999) Random Forests—Random Features. [Online] Available from: https://www.stat.berkeley.edu/~breiman/random-forests.pdf [Accessed: 15 November 2022].

34. ΔΙΩΝΗ ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ (2019) Εφαρμογή αλγορίθμων Μηχανικής Μάθησης σε σύνολα δεδομένων και αποτίμηση αποτελεσμάτων [Online] Availablefrom:

https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/11873/Karantzias_me1607.pdf?sequence=3&isAllowed=y [Accessed: 15 November 2022].

35. ANALYTICS VIDHYA (2021) Logistic Regression- Supervised Learning Algorithm for Classification. [Online] Available from: https://www.analyticsvidhya.com/blog/2021/05/logistic-regression-supervised-learning-algorithm-for-classification/ [Accessed: 15 November 2022].

36. DATAASPIRANT (2021) HOW CATBOOST ALGORITHM WORKS IN MACHINE LEARNING. [Online] Available from: https://dataaspirant.com/catboost-algorithm/ [Accessed: 1 November 2022].

37. WISEGEEK (2022) What Factors Affect Airline Customer Satisfaction? [Online] Available from: https://www.wise-geek.com/what-factors-affect-airline-customer-satisfaction.htm [Accessed: 1 November 2022].

38. STATISTA (2022) Air transportation - statistics & facts [Online] Available from: https://www.statista.com/topics/1707/air-transportation/#topicOverview [Accessed: 1 November 2022].

39. IATA (2018) Aviation Safety [Online] Available from: https://www.iata.org/en/youandiata/travelers/aviation-safety/ [Accessed: 1 November 2022].

40. P PORTO (2008) KDD, SEMMA and CRISP-DM: a parallel overview [Online] Available from: https://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf [Accessed: 1 November 2022].

41. K D NUGGETS (2017) Four Problems in Using CRISP-DM and How to Fix Them [Online] Available from: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html [Accessed: 1 November 2022].

42. FOCUS WIRE (2022) Smooth online booking key to attracting repeat airline customers [Online] Available from: https://www.phocuswire.com/smooth-online-booking-key-to-attracting-repeat-airline-customers [Accessed: 1 November 2022].

43. ANHALT UNIVERSITY OF APPLIED SCIENCE (2018) Dynamic airline in-flight entertainment systems using predictive analysis[Online]Available from: http://opendata.uni-halle.de/bitstream/1981185920/12361/1/Masterarbeit%20Elena%20Hawk.pdf [Accessed: 1 November 2022].

ΠΑΡΑΡΤΗΜΑ 1: ΜΟΡΦΗ ΕΡΩΤΗΜΑΤΟΛΟΓΙΟΥ

Προφίλ Επιβάτη Χαρακτηριστικά ταξιδιού

1. Φύλο

- ◯ Άνδρας
- ◯ Γυναίκα

2. Ηλικία

3. Πιστότητα επιβάτη

- ◯ Τακτικός
- ◯ Σπάνιος

4. Λόγος Ταξιδιού

- ◯ Επαγγελματικός
- ◯ Προσωπικός

5. Απόσταση ταξιδιού

6. Καθυστέρηση αναχώρησης (σε λεπτά)

7. Καθυστέρηση άφιξης (σε λεπτά)

8. Είδος θέσης

- ◯ Economy
- ◯ Economy Plus
- ◯ Business

Αξιολόγηση υπηρεσιών

9. Αξιολογήστε τις ακόλουθες υπηρεσίες

	1(λίγο)	2	3	4	5(πολύ)
Υπηρεσία Wifi (αν υπάρχει)	◯	◯	◯	◯	◯
Χρόνοι αναχώρησης και άφιξης	◯	◯	◯	◯	◯
Διαδικασία κράτησης εισιτηρίου	◯	◯	◯	◯	◯
Τοποθεσία εξόδου αναχώρησης	◯	◯	◯	◯	◯
Σνακ, ποτά	◯	◯	◯	◯	◯
Επιβίβαση	◯	◯	◯	◯	◯
Άνεση θέσης	◯	◯	◯	◯	◯
Διασκέδαση εν πτήσει	◯	◯	◯	◯	◯
Υπηρεσίες στο αεροσκάφος	◯	◯	◯	◯	◯
Χώρος για τα πόδια	◯	◯	◯	◯	◯
Χειρισμός αποσκευών	◯	◯	◯	◯	◯
Check-in	◯	◯	◯	◯	◯
Υπηρεσίες εν πτήσει	◯	◯	◯	◯	◯
Καθαριότητα	◯	◯	◯	◯	◯

10. Ικανοποίηση με την αεροπορική εταιρεία

- ◯ Ικανοποιημένος
- ◯ Ουδέτερος ή δυσαρεστημένος

ΠΑΡΑΡΤΗΜΑ 2: ΓΛΩΣΣΑ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ PYTHON

Ένα σημαντικό εργαλείο για την ανάλυση ενός συνόλου δεδομένων είναι μια γλώσσα προγραμματισμού (Bharati, 2019). Για την υλοποίηση των μοντέλων των τεχνικών ταξινόμησης και ομαδοποίησης θα χρησιμοποιηθεί η γλώσσα προγραμματισμού «Python» .

A. Python

Η γλώσσα αυτή γράφτηκε από τον Ολλανδό προγραμματιστή GuidoVanRossumστα τέλη της δεκαετίας 1980-90. Η Python είναι μια γλώσσα προγραμματισμού με απλό συντακτικό, εξαιρετική αναγνωσιμότητα και φορητότητα. (Κομηνέας, 2017) Είναι μια γλώσσα υψηλού επιπέδου, δηλαδή ο κώδικας της πρέπει να μετατραπεί σε γλώσσα μηχανής για να εκτελεστεί από τον Η/Υ. Στην περίπτωση της Python, η επεξεργασία αυτή γίνεται από διερμηνευτές (interpreters). Εκτός από τα παραπάνω, η Python καθίσταται κατάλληλη καθώς είναι μια γλώσσα ανοικτού κώδικα που επιτρέπει την ανταλλαγή ιδεών, συνεργασία όσον αφορά τον πηγαίο κώδικα και διαθέτει αρκετά πακέτα υποστήριξης (βιβλιοθήκες).

B. Χρήσεις της Python

Η Python έχει ποικίλες χρήσεις από αυτοματοποίηση διαδικασιών μέχρι την υλοποίηση καθημερινών εργασιών. Η Python χρησιμοποιείται επιπλέον των προγραμματιστών από λογιστές, επιστήμονες ή επιχειρηματίες. Κάποιες από τις χρήσεις της είναι η ανάπτυξη ιστοσελίδων υλοποιώντας εργασίες όπως επικοινωνία με βάσεις δεδομένων και ασφάλεια. (Ellis, 2020) Επίσης, χρησιμοποιείται για την αυτοματοποίηση διαδικασιών όπως ο έλεγχος για λάθη σε ποικίλα αρχεία ή η μετατροπή αρχείων ή ο προγραμματισμός αποστολής μηνυμάτων ηλεκτρονικού ταχυδρομείου.(Wong, 2021) Όσον αφορά τις καθημερινές εργασίες από μη

προγραμματιστές θα μπορούσε να ήταν πρόγραμμα για αποστολή υπενθύμισης να πάρει ο χρήστης ομπρέλα σε περίπτωση βροχής ή πρόγραμμα για τον έλεγχο αποθεμάτων ενός καταστήματος.

Κλείνοντας εστιάζουμε στη χρήση της Python στην περίπτωση που πραγματεύεται αυτή η εργασία, δηλαδή την ανάλυση δεδομένων και τη μηχανική μάθηση. Οι επιστήμονες δεδομένων χρησιμοποιούν αυτή τη γλώσσα για να υλοποιήσουν περίπλοκους στατιστικούς υπολογισμούς, να δημιουργήσουν γραφήματα με τα δεδομένα, να χτίσουν αλγορίθμους μηχανικής μάθησης, να χειριστούν την ανάλυση δεδομένων ή άλλες εργασίες σχετικές με δεδομένα.(Raschkaetal., 2020) Η Python μπορεί να χτίσει ένα μεγάλο εύρος από απεικονίσεις δεδομένων όπως γραφήματα με γραμμές ή μπάρες, διάγραμμα πίτα, ιστόγραμμα ή 3D plots. Επιπρόσθετα, έχει βιβλιοθήκες που επιτρέπουν τη συγγραφή προγραμμάτων σχετικών με την ανάλυση δεδομένων και τη μηχανική μάθηση εύκολα και αποδοτικά.

ΠΑΡΑΡΤΗΜΑ 3: ΕΞΕΡΕΥΝΗΣΗ ΔΕΔΟΜΕΝΩΝ

Γίνεται πρώτη εξερεύνηση των δεδομένων μέσω των συναρτήσεων info() και describe(). Μέσω της info εντοπίζεται ο αριθμός των μεταβλητών-στηλών και εγγραφών- γραμμών του συνόλου καθώς και πληροφορίες για κάθε μία από αυτές όπως το όνομα και το είδος μεταβλητής(αριθμητική, συμβολοσειρά κτλ.). Μέσω της describe φαίνονται οι στατιστικές τιμές για κάθε μεταβλητή όπως η μέση τιμή, η τυπική απόκλιση, η μέγιστη/ελάχιστη τιμή κτλ. (Lewis, 2020)Στην περίπτωση αυτή υπάρχουν τόσο αριθμητικές όσο και κατηγορικές μεταβλητές. Οπότε, η εντολή describe χρησιμοποιείται 2 φορές ξεχωριστά. Όσον αφορά τις κατηγορικές μεταβλητές παραλείπονται μέσες τιμές, τυπικές αποκλίσεις και τα τεταρτημόρια.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 129880 entries, 0 to 129879
Data columns (total 24 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   id                                129880 non-null  int64
 1   Gender                            129880 non-null  object
 2   Customer Type                     129880 non-null  object
 3   Age                               129880 non-null  int64
 4   Type of Travel                    129880 non-null  object
 5   Class                             129880 non-null  object
 6   Flight Distance                   129880 non-null  int64
 7   Inflight wifi service             129880 non-null  int64
 8   Departure/Arrival time convenient 129880 non-null  int64
 9   Ease of Online booking            129880 non-null  int64
 10  Gate location                     129880 non-null  int64
 11  Food and drink                    129880 non-null  int64
 12  Online boarding                   129880 non-null  int64
 13  Seat comfort                      129880 non-null  int64
 14  Inflight entertainment            129880 non-null  int64
 15  On-board service                  129880 non-null  int64
 16  Leg room service                  129880 non-null  int64
 17  Baggage handling                  129880 non-null  int64
 18  Checkin service                   129880 non-null  int64
 19  Inflight service                  129880 non-null  int64
 20  Cleanliness                       129880 non-null  int64
 21  Departure Delay in Minutes        129880 non-null  int64
 22  Arrival Delay in Minutes          129487 non-null  float64
 23  satisfaction                      129880 non-null  object
dtypes: float64(1), int64(18), object(5)
```

Εικόνα 44: Κώδικας και outputτης εντολής info

```
In [4]: df.describe()
```

Out[4]:

	id	Age	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comf
count	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000000	129880.000
mean	64940.500000	39.427957	1190.316392	2.728696	3.057599	2.756876	2.976925	3.204774	3.252633	3.441
std	37493.270818	15.119360	997.452477	1.329340	1.526741	1.401740	1.278520	1.329933	1.350719	1.319
min	1.000000	7.000000	31.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	32470.750000	27.000000	414.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000
50%	64940.500000	40.000000	844.000000	3.000000	3.000000	3.000000	3.000000	3.000000	3.000000	4.000
75%	97410.250000	51.000000	1744.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	5.000
max	129880.000000	85.000000	4983.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000

Εικόνα 15: Κώδικας και outputτης εντολής describeγια αριθμητικές μεταβλητές

```
In [5]: df.describe(include='object')
```

Out[5]:

	Gender	Customer Type	Type of Travel	Class	satisfaction
count	129880	129880	129880	129880	129880
unique	2	2	2	3	2
top	Female	Loyal Customer	Business travel	Business	neutral or dissatisfied
freq	65899	106100	89693	62160	73452

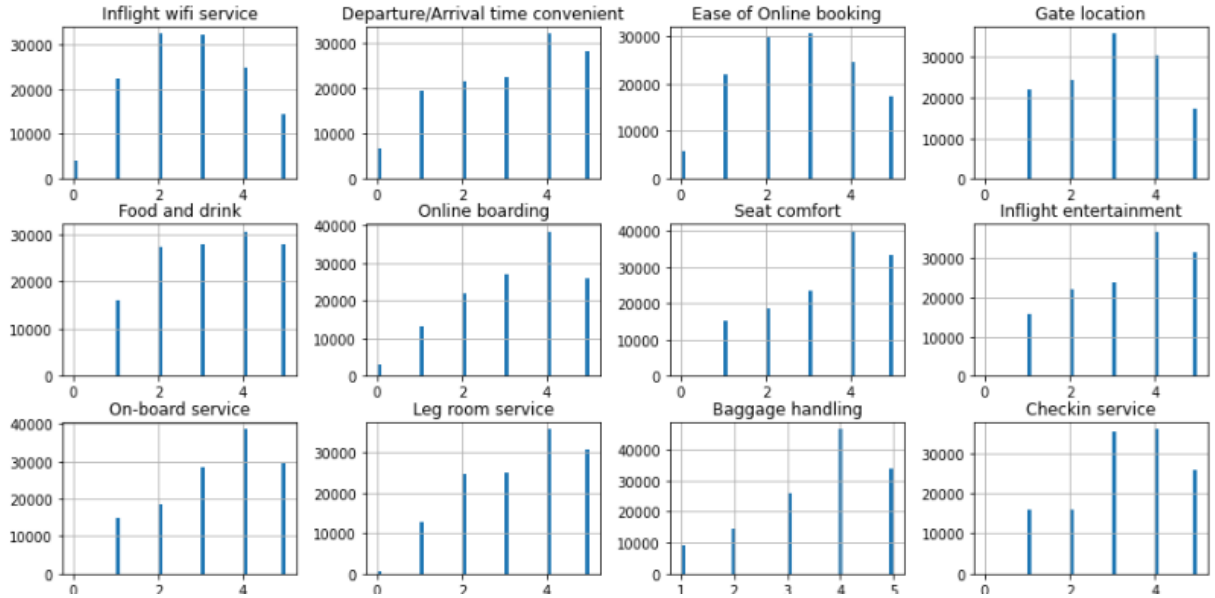
Εικόνα 16: Κώδικας και outputτης εντολής describeγια κατηγορικές μεταβλητές

Ένας άλλος τρόπος για να έχουμε μια γενική εικόνα των δεδομένων είναι μέσω ιστογραμμάτων για κάθε μία από τις αριθμητικές μεταβλητές. Φαίνεται δηλαδή για κάθε μεταβλητή το εύρος των τιμών που παίρνει καθώς και ποιες τιμές είναι συχνότερες ή σπανιότερες. Όσον αφορά τις κατηγορικές μεταβλητές, χρησιμοποιούνται διαγράμματα μπάρας που δείχνουν τη συχνότητα κάθε πιθανής τιμής μιας κατηγορικής μεταβλητής. Τέλος χρησιμοποιείται ο συντελεστής συσχέτισης Pearson που δείχνει συσχέτιση μεταξύ των αριθμητικών μεταβλητών.

```
import matplotlib.pyplot as plt

df.iloc[:,7:21].hist(bins=50, figsize=(15,10))

plt.show()
```



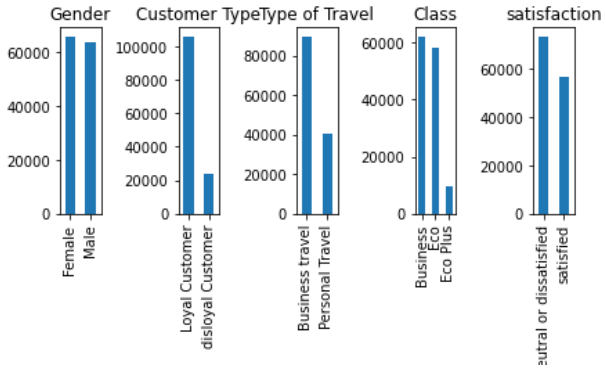
Εικόνα 17: Κώδικας και outputγια δημιουργία προβολή ιστογραμμάτων

```
In [15]: df2=df[['Gender','Customer Type','Type of Travel','Class','satisfaction']]

fig, ax = plt.subplots(1, 5)

for i, categorical_feature in enumerate(df2):
    df2[categorical_feature].value_counts().plot(kind="bar",ax=ax[i]).set_title(categorical_feature)

fig.tight_layout()
plt.show()
```



Εικόνα 18: Κώδικας και outputγια τη δημιουργία προβολή διαγραμμάτων μπάρας για τις κατηγορικές μεταβλητές

```
In [16]: df.corr()
```

Out[16]:

	id	Age	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	
id	1.000000	0.020322	0.095504	-0.023096	-0.002192	0.013400	-0.000113	-0.000510	0.055538	0.052164	0.001620	0.055502	0
Age	0.020322	1.000000	0.099459	0.016116	0.036960	0.022565	-0.000398	0.023194	0.207572	0.159136	0.074947	0.057078	0
Flight Distance	0.095504	0.099459	1.000000	0.006701	-0.018914	0.065165	0.005520	0.057066	0.214825	0.157662	0.130507	0.111194	0
Inflight wifi service	-0.023096	0.016116	0.006701	1.000000	0.344915	0.714807	0.338573	0.132214	0.457445	0.121513	0.207802	0.119928	0
Departure/Arrival time convenient	-0.002192	0.036960	-0.018914	0.344915	1.000000	0.437620	0.447510	0.000687	0.072287	0.008666	-0.008380	0.067297	0
Ease of Online booking	0.013400	0.022565	0.065165	0.714807	0.437620	1.000000	0.460041	0.030514	0.404866	0.028561	0.046564	0.039064	0
Gate location	-0.000113	-0.000398	0.005520	0.338573	0.447510	0.460041	1.000000	-0.002872	0.002756	0.002788	0.002741	-0.029019	-0
Food and drink	-0.000510	0.023194	0.057066	0.132214	0.000687	0.030514	-0.002872	1.000000	0.233500	0.575846	0.623461	0.057404	0
Online boarding	0.055538	0.207572	0.214825	0.457445	0.072287	0.404866	0.002756	0.233500	1.000000	0.419253	0.283922	0.154242	0
Seat comfort	0.052164	0.159136	0.157662	0.121513	0.008666	0.028561	0.002788	0.575846	0.419253	1.000000	0.611837	0.130545	0
Inflight entertainment	0.001620	0.074947	0.130507	0.207802	-0.008380	0.046564	0.002741	0.623461	0.283922	0.611837	1.000000	0.418574	0
On-board service	0.055502	0.057078	0.111194	0.119928	0.067297	0.039064	-0.029019	0.057404	0.154242	0.130545	0.418574	1.000000	0
Leg room service	0.044088	0.039119	0.134533	0.160317	0.010617	0.109450	-0.005181	0.033173	0.123225	0.104272	0.300397	0.357721	1

Εικόνα 19: Κώδικας και outputγια τη δημιουργία προβολή συσχετίσεων μεταξύ αριθμητικών μεταβλητών

ΠΑΡΑΡΤΗΜΑ 4: ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Η προετοιμασία των δεδομένων για τον αλγόριθμο Δέντρου Απόφασης περιλαμβάνει τις εξής ενέργειες:

- Χειρισμός των τιμών που λείπουν από το σύνολο δεδομένων: Χρειάζεται να συμπληρωθούν τιμές που λείπουν από το σύνολο δεδομένων που μπορεί να είναι είτε αριθμητικές είτε κατηγορικές. Η διαδικασία περιλαμβάνει να προσδιοριστούν από ποιες στήλες λείπουν δεδομένα. Έπειτα θα πρέπει να δοθεί μια τιμή στα κελιά τα οποία είναι κενά.(Anesthesiol, 2013) Ακολουθείται διαφορετική διαδικασία για τις κατηγορικές μεταβλητές και διαφορετική διαδικασία για τις αριθμητικές. Επιλέγεται μια στρατηγική αντικατάστασης των κενών τιμών. Όσον αφορά τις στρατηγικές, στις κατηγορικές μεταβλητές χρησιμοποιείται η πιο συχνή τιμή, ενώ στις αριθμητικές χρησιμοποιείται η διάμεσος. Στην ανάλυση αυτή χρειάστηκε μόνο η τεχνική συμπλήρωσης μιας αριθμητικής μεταβλητής.

```
#define columns with at least one na value
lst=df.columns[df.isna().any()].to_list()
print('The columns with na values are:')
print(lst)
```

```
The columns with na values are:
['Arrival Delay in Minutes']
```

```
#replace missing values of numeric variables with median
df['Arrival Delay in Minutes']=df['Arrival Delay in Minutes'].fillna(df['Arrival Delay in Minutes'].median())
```

Εικόνα 20: Κώδικας για χειρισμό τιμών που λείπουν

- Επιλογή των πιο σημαντικών ιδιοτήτων: Παρακάτω περιγράφονται οι διαδικασία για την επιλογή των3 σημαντικότερων αριθμητικών μεταβλητών και της σημαντικότερης κατηγορική μεταβλητής προκειμένου να επιτευχθεί η εύρεση των σημαντικότερων παραγόντων που επηρεάζουν την ικανοποίηση των επιβατών.

Όσον αφορά τις αριθμητικές μεταβλητές μελετάται η συσχέτιση κάθε μια από αυτές με την εξαρτημένη μεταβλητή του προβλήματος. Εφαρμόζεται F-ANOVA τεστ,

υπολογίζονται οι τιμές F για κάθε συσχέτιση και επιλέγονται οι 3 μεγαλύτερες τιμές δηλαδή οι 3 ιδιότητες που έχουν μεγαλύτερη εξάρτηση με την εξαρτημένη μεταβλητή. Για το σκοπό αυτό χρησιμοποιούνται τα f_classif και SelectKBest της βιβλιοθήκης sklearn.feature_selection. Η παράμετρος του SelectKBest είναι 3 όσες οι ιδιότητες που θέλουμε. Το αποτέλεσμα του τεστ δίνει τις 3 μεταβλητές που είναι σημαντικότεροι παράγοντες.(Hooker, 2012)

```
#numerical features selection
import numpy as np

df1=df.iloc[:,7:]

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif

X=df1.iloc[:,:-1]
y=np.ravel(df1[['satisfaction']])

fvalue_Best = SelectKBest(f_classif, k=3)
X_kbest = fvalue_Best.fit_transform(X, y)

f = X.columns[fvalue_Best.get_support(indices=True)]
```

```
f
```

```
Index(['Online boarding', 'Seat comfort', 'Inflight entertainment'], dtype='object')
```

Εικόνα 21: Κώδικας για την επιλογή 3 σημαντικότερων αριθμητικών μεταβλητών και εμφάνιση αυτών

Όσον αφορά τις κατηγορικές μεταβλητές μελετάται η συσχέτιση κάθε μια από αυτές με την εξαρτημένη μεταβλητή του προβλήματος. Εφαρμόζεται X^2 τεστ, υπολογίζονται οι τιμές της συνάρτησης X^2για κάθε συσχέτιση και επιλέγεται η μεγαλύτερη τιμή της, δηλαδή η ιδιότητα που έχει μεγαλύτερη εξάρτηση με την εξαρτημένη μεταβλητή. Για το σκοπό αυτό χρησιμοποιείται το chi2_contingency της βιβλιοθήκης scipy.stats. Έπειτα αποθηκεύεται η τιμή X^2κάθε μεταβλητής σε ένα λεξικό και τοποθετούνται οι ιδιότητες σε φθίνουσα σειρά ως προς το X^2. Η ιδιότητα με τη μεγαλύτερη τιμή δηλώνει και το σημαντικότερο παράγοντα που επηρεάζει την εξαρτημένη μεταβλητή. (Brownlee, 2020)

```
#categorical features selection

df2=df[['Customer Type','Type of Travel','Class','satisfaction']]
from scipy.stats import chi2_contingency
dict={}

for i in range(0,3):
    contingency = pd.crosstab(df2.iloc[:,i], df2['satisfaction'])
    c,p, dof, expected = chi2_contingency(contingency)
    dict[df2.columns[i]]=c

sorted(dict.items(), key=lambda x:x[1],reverse=True)
```

```
[('Class', 32906.17185866312),
 ('Type of Travel', 26282.520993423812),
 ('Customer Type', 4493.188803283598)]
```

Εικόνα 22: Κώδικας για έλεγχο συσχετίσεων κατηγορικών μεταβλητών

- Κωδικοποίηση κατηγορικών μεταβλητών: Προκειμένου να υλοποιηθεί οι αλγόριθμοι κατηγοριοποίησης Δέντρο Απόφασης και Τυχαίο Δάσος χρειάζεται να κωδικοποιηθούν οι κατηγορικές μεταβλητές που επιλέχθηκαν. Στην περίπτωση που είναι δίτιμη η μεταβλητή (ναι ή όχι) όπως πιστός επιβάτης μπορούν εύκολα να τοποθετηθούν οι τιμές 0(όχι) και 1(ναι). Στην περίπτωση που οι μεταβλητή παίρνει περισσότερες τιμές, για να γίνει καλύτερη ταξινόμηση χρειάζεται οι τιμές που θα δοθούν να ακολουθούν μια λογική σειρά ανάλογα με αυτό που αντιπροσωπεύουν.(Potdaretal., 2017) Γι' αυτό δίνονται τιμές από το 0 μέχρι το συνολικό αριθμό τιμών ανάλογα με το πόσο μοιάζουν οι τιμές μεταξύ τους.

```
#encoding of categorical variables
df["Class"]=df["Class"].map({"Eco":0,"Eco Plus":1,"Business":2})
```

Εικόνα 23: Κώδικας για κωδικοποίηση κατηγορικών μεταβλητών

- Δημιουργία συνόλου εκπαίδευσης και τεστ: Για την υλοποίηση τέτοιου είδους αλγορίθμων, χρειάζεται να χωριστούν τα δεδομένα σε σύνολο εκπαίδευσης για να δημιουργηθεί το μοντέλο και σύνολο τεστ για να αξιολογηθούν οι προβλέψεις του μοντέλου. Στην περίπτωση αυτή επιλέγεται τυχαία ένα 30% του συνόλου για τεστ μέσω της εντολής train_test_split της βιβλιοθήκης sklearn.model_selection.(Brownlee, 2020)

```
: #define selected independent variables(factors) and the dependent variable
X=df[['Online boarding','Seat comfort','Inflight entertainment','Class']]
y=df[['satisfaction']]

from sklearn.model_selection import train_test_split


# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) # 70% training and 30% test
```

Εικόνα 24: Κώδικας για χωρισμό του συνόλου

ΠΑΡΑΡΤΗΜΑ 5: ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Οι τεχνικές εποπτευόμενης μάθησης διαθέτουν κατάλληλες μετρικές για να αξιολογηθεί η απόδοση του μοντέλου πρόβλεψης οι οποίες θα παρουσιαστούν παρακάτω:

- Ορθότητα

Δείχνει αν ένα μοντέλο εκπαιδεύτηκε σωστά και μπορεί να αποδώσει καλά. Υπολογίζεται ως οι σωστές προβλέψεις που έγιναν προς το σύνολο των προβλέψεων. Χρειάζεται το accuracy_score από τη βιβλιοθήκη sklearn.metrics.

```
#Evaluation of test set


from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.metrics import classification_report
import seaborn as sns
import matplotlib.pyplot as plt



from sklearn.metrics import accuracy_score

# Model Accuracy, how often is the classifier correct?
print("Accuracy of test set:",round(metrics.accuracy_score(y_test, y_pred),2)*100,'%')
```

Εικόνα 25: Κώδικας υπολογισμού ορθότητας

Όσον αφορά τις μετρικές που εστιάζουν σε κάθε μία τάξη ξεχωριστά:

- Πίνακας σύγχυσης

Για κάθε τάξη, φαίνεται ποιες εγγραφές τοποθετήθηκαν σωστά και ποιες όχι. Στη διαγώνιο του πίνακα φαίνονται οι εγγραφές που τοποθετήθηκαν σωστά.

Λαμβάνεται υπόψη και το αντίστοιχο ποσοστό τους με δείκτες που παρέχονται στην αναφορά ταξινόμησης παρακάτω.

- Αναφορά ταξινόμησης: Παρέχει πληροφορίες για κάθε μια τάξη (απόρριψη ή έγκριση) χρησιμοποιώντας τους δείκτες ακρίβειας , ανάκλησης και F1-score.

Όσον αφορά την ακρίβεια, για κάθε τάξη ελέγχονται πόσες προβλέψεις της τάξης αυτής προβλέφθηκαν σωστά προς το σύνολο των πραγματικών τιμών αυτής της τάξης. Είναι σημαντικό να αξιολογηθεί όταν το κόστος των λανθασμένων προβλέψεων της τάξης είναι μεγάλο.

Όσον αφορά την ανάκληση, προκύπτει ως οι σωστές προβλέψεις μιας τάξης προς το σύνολο των προβλέψεων αυτής της τάξης.

Όσον αφορά το scoreF1, είναι μια μετρική που συνδυάζει ακρίβεια και ανάκληση. (Mishra, 2018)

```
#classification report to have a look  at precision,recall, f1-score
print(classification_report(y_test, y_pred))

#confusion matrix
q=confusion_matrix(y_test, y_pred)
ax= plt.subplot()
sns.heatmap(q, annot=True, fmt='g', ax=ax)
ax.set_xlabel('Predicted labels');ax.set_ylabel('True labels')
ax.set_title('Confusion Matrix')
```

Εικόνα 26: Κώδικας εύρεσης πίνακα σύγχυσης και αναφοράς ταξινόμησης