ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
TECHNICAL UNIVERSITY OF CRETE

*Σχολή Μηχανικών Παραγωγής και Διοίκησης*

*Μεταπτυχιακό Πρόγραμμα Σπουδών στη*

*Διοίκηση Επιχειρήσεων*

# «Αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη τιμών ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης»

**Χρύσα Αντρία**

**Διπλ. Πολιτικός Μηχανικός Α.Π.Θ.**

**Α.Μ. 2022019010**

**Επιβλέπων καθηγητής: κ. Τσαφαράκης Στέλιος**

**Τριμελής εξεταστική επιτροπή**

**κ. Ατσαλάκης Γεώργιος**

**κ. Δούμπος Μιχαήλ**

**κ. Τσαφαράκης Στέλιος**

**Δεκέμβριος 2023**

## *Ευχαριστίες*

*Με το πέρας της μεταπτυχιακής μου εργασίας, θα ήθελα πρωτίστως να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κύριο Στέλιο Τσαφαράκη, για την πολύτιμη βοήθεια, καθοδήγηση, αλλά και την άμεση ανταπόκρισή του σε οποιαδήποτε δυσκολία παρουσιάστηκε καθ' όλη τη διάρκεια εκπόνησής της. Επίσης, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον κύριο Κωνσταντίνο Ζερβουδάκη για την ουσιαστική συμβολή του και την πολύτιμη υποστήριξή του κατά την πορεία της εργασίας, καθώς και για την προθυμία του να με καθοδηγήσει σε κρίσιμα σημεία της μελέτης μου. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την ανιδιοτελή αγάπη και στήριξή τους.*

# Περίληψη

Η ραγδαία ανάπτυξη της αγοράς βραχυχρόνιας μίσθωσης καταλυμάτων μέσω πλατφορμών όπως το Airbnb έχει δημιουργήσει την ανάγκη για ακριβέστερη και πιο αξιόπιστη εκτίμηση των τιμών ενοικίασης. Η παρούσα εργασία διερευνά την εφαρμογή αλγορίθμων μηχανικής μάθησης με στόχο την εκτίμηση της τιμής ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης, με πεδίο μελέτης την περιοχή της Θεσσαλονίκης.

Αρχικά πραγματοποιήθηκε συλλογή και προεπεξεργασία δεδομένων από την πλατφόρμα Inside Airbnb. Στη συνέχεια, αναπτύχθηκαν και αξιολογήθηκαν δύο μοντέλα πρόβλεψης: ένα βασισμένο στα Τεχνητά Νευρωνικά Δίκτυα και ένα στον αλγόριθμο Τυχαία Δάση. Η σύγκριση των αποτελεσμάτων πραγματοποιήθηκε μέσω δεικτών απόδοσης όπως ο συντελεστής προσδιορισμού $R^2$ και το Root Mean Squared Error (RMSE). Τα αποτελέσματα έδειξαν ότι ο αλγόριθμος Τυχαία Δάση μπορεί να δώσει καλύτερες εκτιμήσεις καθώς είναι πιο αποδοτικός με συντελεστή $R^2 = 0,898$ και μέσο σφάλμα RMSE = 13,442.

Τα ευρήματα καταδεικνύουν τη χρησιμότητα της μηχανικής μάθησης ως εργαλείο υποστήριξης αποφάσεων στην τιμολόγηση ακινήτων βραχυχρόνιας μίσθωσης. Παράλληλα, αναδεικνύονται οι δυνατότητες και οι προοπτικές εξέλιξης μέσω της ενσωμάτωσης εξωτερικών παραμέτρων και πιο σύνθετων αλγορίθμων σε μελλοντικές εφαρμογές.

# Abstract

The rapid expansion of the short-term rental market through platforms such as Airbnb has created a growing need for more accurate and reliable rental price prediction models. This study explores the application of machine learning algorithms for forecasting rental prices of short-term accommodation, with a focus on the city of Thessaloniki.

Following data collection and preprocessing using information from the Inside Airbnb platform, two predictive models were developed and evaluated: one based on Artificial Neural Networks and another on the Random Forest algorithm. The performance of both models was assessed using key metrics, including the coefficient of determination ($R^2$) and the Root Mean Squared Error (RMSE). Results indicated that the Random Forest algorithm can provide better estimations, as it is more efficient with an $R^2$ coefficient of 0,898 and a mean error (RMSE) of 13,442.

These findings highlight the potential of machine learning as a decision-support tool in the pricing strategy of short-term rental properties. At the same time, they point to future opportunities for improvement through the incorporation of external variables and the application of more advanced algorithmic approaches.

# Περιεχόμενα

# 1. Εισαγωγή

Την τελευταία δεκαετία ο τομέας των βραχυχρόνιων μισθώσεων έχει γνωρίσει τεράστια άνθιση με τους ταξιδιώτες να προτιμούν αυτού του τύπου τα καταλύματα ολοένα και περισσότερο για τη διαμονή τους. Αυτό οφείλεται σε διάφορους λόγους, βασικότεροι από τους οποίους είναι:

- **Ποικιλία στους τύπους καταλυμάτων.**
  Οι ενοικιαστές μπορούν να επιλέξουν από μικρά διαμερίσματα μέχρι μεγάλα σπίτια/εξοχικά που είναι ιδανικά για αυτούς, αναλόγως με το εάν ταξιδεύουν μόνοι τους ή με κάποια μεγαλύτερη ομάδα ατόμων (π.χ. οικογένεια).
- **Ευελιξία.**
  Τα καταλύματα βραχυχρόνιας διαμονής προσφέρουν ευελιξία στις περιόδους διαμονής, με αποτέλεσμα να καθίσταται εύκολη η ενοικίαση κάποιου για μία ημέρα, μια εβδομάδα ή ακόμα και για μεγαλύτερο διάστημα, ανάλογα με τις ανάγκες του εκάστοτε ενοικιαστή.
- **Εξοικονόμηση χρημάτων.**
  Καθώς είναι πιο οικονομικά από τα παραδοσιακά ξενοδοχεία, ιδίως όταν πρόκειται να εκμισθωθούν από μεγαλύτερες ομάδες.
- **Πλήρης εξοπλισμός.**
  Τα καταλύματα αυτού του τύπου συνήθως παρέχονται πλήρως εξοπλισμένα (πλυντήριο, κουζίνα κ.α.) που κάνουν τη διαμονή πολλή πιο άνετη.

Η ολοένα και αυξανόμενη λοιπόν ζήτηση που γνωρίζουν σήμερα τα καταλύματα βραχυχρόνιας μίσθωσης, που υποστηρίζεται από πλατφόρμες όπως το AIRBNB, καθιστά τη δυνατότητα της πρόβλεψης των τιμών ενοικίασής τους ζωτικής σημασίας τόσο για τους ενοικιαστές αλλά κυρίως για τους ιδιοκτήτες. Ωστόσο η πολυπλοκότητα και η ποικιλία των παραγόντων που την επηρεάζουν, όπως η ακριβής τοποθεσία, η εποχή, τα γενικά χαρακτηριστικά του εκάστοτε ακινήτου, οι τοπικές τάσεις κ.α. δημιουργούν προκλήσεις που υπερβαίνουν την ικανότητα πρόβλεψης των ανθρώπων.

Η Μηχανική Μάθηση και η Τεχνητή Νοημοσύνη, με τη βοήθεια των οποίων θα μπορούσαν να ξεπεραστούν αυτού του είδους τα προβλήματα έχουν εισχωρήσει εκτός των άλλων και στον συγκεκριμένο τομέα. Εταιρείες όπως η Zillow, έχουν ήδη αρκετά εργαλεία και λειτουργείες βασισμένες σε αυτές. Χαρακτηριστικό παράδειγμα αποτελεί

η πλατφόρμα Zestimate, μέσω της οποίας μπορεί οποιοσδήποτε κάτοικος της Αμερικής να λάβει μια εκτίμηση του ακινήτου του. Η εκτίμηση αυτή γίνεται όλο και πιο ακριβής όσο ο όγκος των δεδομένων των ακινήτων που εισάγονται στην πλατφόρμα αυξάνονται.

Στη χώρα μας όμως τεχνολογίες όπως αυτή βρίσκουν μικρή εφαρμογή σε προβλήματα κοστολόγησης κατοικιών και ακόμα μικρότερη στην κοστολόγηση καταλυμάτων βραχυχρόνιας μίσθωσης. Το παραπάνω, σε συνδυασμό με την αξία της ακριβούς εκτίμησης του μισθώματος των συγκεκριμένων καταλυμάτων, δημιουργεί την ανάγκη για περαιτέρω έρευνα και προσπάθεια πρακτικής εφαρμογής μεθόδων μηχανικής μάθησης με σκοπό την αυτοματοποίησή της. Η παρούσα διπλωματική εργασία επικεντρώνεται ακριβώς σε αυτήν την προσπάθεια.

## 1.1 Η πλατφόρμα AIRBNB

Τα δεδομένα των καταλυμάτων που χρησιμοποιήθηκαν στην παρούσα αντλήθηκαν από την πλατφόρμα AIRBNB κρίνεται λοιπόν σκόπιμη μια γενική αναφορά σε αυτή.

Όλα ξεκίνησαν το 2007 κατά τη διάρκεια του Συνεδρίου Βιομηχανικού Σχεδίου στο San Francisco, όταν οι Brian Chesky και Joe Gebbia αποφάσισαν να μετατρέψουν το σπίτι τους σε τουριστικό κατάλυμα εξαιτίας της πληρότητας όλων των δωματίων των ξενοδοχείων της πόλης. Εξόπλισαν λοιπόν το σπίτι τους με μόλις τρία επιπλέον φουσκωτά στρώματα και μέσω του ιστοσελίδας airbedandbreakfast.com την οποία δημιούργησαν, βρήκαν τους πρώτους τρείς φιλοξενούμενούς τους, οι οποίοι και πλήρωσαν 80 δολάρια έκαστος. Εκείνη ήταν η στιγμή που εντόπισαν ένα κενό και μια πολύ μεγάλη ευκαιρία που υπήρχε στην συγκεκριμένη αγορά.

Έτσι οι δύο 27χρονοι τότε Chesky και Gebbia βρήκαν τον προγραμματιστή Nathan Blecharczyk, ο οποίος θα τους έφτιαχνε την σημερινή πλατφόρμα. Οι ρυθμοί ανάπτυξης της εταιρείας ήταν ιλιγγιώδεις με αποτέλεσμα η πλατφόρμα να έχει ολοκληρωθεί το 2008 με 800 καταχωρήσεις χώρων προς ενοικίαση σε αυτή αμέσως.

Στη συνέχεια, η έκταση της εταιρείας πήρε παγκόσμιες διαστάσεις με την AIRBNB να δραστηριοποιείται σήμερα σε περισσότερες από 220 χώρες και περιοχές του κόσμου. Τα κύρια χαρακτηριστικά της πλατφόρμας που την έχουν καθιερώσει ως ένα από τα μεγαλύτερα ονόματα στη βραχυχρόνια μίσθωση καταλυμάτων είναι τα εξής:

- **Ποικιλία στους τύπους διαμονής.**

  Οι ταξιδιώτες έχουν τη δυνατότητα να επιλέξουν τον καταλληλότερο για αυτούς κάθε φορά τύπο καταλύματος. Από βίλες, μονοκατοικίες, διαμερίσματα μέχρι ένα μόνο δωμάτιο σε σπίτι που ήδη κατοικείται από κάποιον άλλον, φορητά τροχόσπιτα κ.α.

- **Κριτικές.**

  Οι χρήστες μπορούν να αφήσουν κριτικές για τα καταλύματα, τις οποίες μπορεί να διαβάσει ο κάθε υποψήφιος εκμισθωτής, καθιστώντας έτσι πιο εύκολη τη επιλογή του κατάλληλου καταλύματος από τους ταξιδιώτες.

- **Ασφάλεια.**

  Μέσω μηχανισμών που παρέχει το AIRBNB ελέγχεται τόσο η ταυτότητα των οικοδεσποτών όσο και των ενοικιαστών. Επιπλέον, υπάρχουν κανόνες και πολιτικές σχετικά με την ασφάλεια και συμπεριφορά των χρηστών.

- **Εύκολη πληρωμή.**

  Μέσω της πλατφόρμας οι πληρωμές μπορούν να πραγματοποιηθούν εύκολα και κυρίως με ασφάλεια και για τους δύο εμπλεκόμενους.

## 1.2 Ορισμοί – Βασικές έννοιες

Με σκοπό την καλύτερη κατανόηση των όσων πρόκειται να αναλυθούν και εφαρμοστούν στη συνέχεια, στην υπό-ενότητα αυτή παρουσιάζονται οι βασικοί ορισμοί και οι έννοιες οι οποίες πρόκειται να μας απασχολήσουν.

Η Μηχανική Μάθηση, η οποία και θα μας απασχολήσει επί το πλείστο στην παρούσα εργασία, ανήκει στην ευρύτερη κατηγορία της **Τεχνητής Νοημοσύνης (Artificial Intelligence).** Ως τεχνητή νοημοσύνη νοείται η δυνατότητα των μηχανών να αναπαράγουν τη γνωστική αντίληψη ενός ανθρώπου, όπως είναι η μάθηση, η σχεδίαση και η δημιουργικότητα καθιστώντας τες ικανές να «αντιλαμβάνονται» το περιβάλλον τους, να δίνουν λύσεις σε προβλήματα και να ενεργούν με σκοπό την επίτευξη ενός ορισμένου στόχου. Η μηχανή (υπολογιστής) δέχεται δεδομένα, τα επεξεργάζεται και τέλος ανταποκρίνεται βάσει αυτών. Πολλοί είναι οι επιστήμονες οι οποίοι κατά καιρούς έχουν αποπειραθεί να ορίσουν το τί εστί Τεχνητή Νοημοσύνη. Σύμφωνα λοιπόν με τον ***John McCarthy (1956)***, τον ιδρυτή του συγκεκριμένου τομέα, η Τεχνητή Νοημοσύνη ορίζεται ως *«Ο τομέας της επιστήμης που ασχολείται με τη δημιουργία μηχανών που θα εκτελούν λειτουργίες που απαιτούν νοημοσύνη, όπως το μάθημα, η*

*διανόηση, η αντίληψη, η γλώσσα και η σχεδίαση».* Ενώ οι ***Rich & Knight (1990)*** την ορίζουν ως τη *«Μελέτη του πώς να κάνουμε τους υπολογιστές ικανούς να κάνουν πράγματα στα οποία προς το παρόν οι άνθρωποι τα καταφέρνουν καλύτερα».* Πολλές, διαφορετικές ως προς το πεδίο και τις εφαρμογές στις οποίες εστιάζει, είναι οι υπό-κατηγορίες που ανήκουν στον ευρύτερο κλάδο της τεχνητής νοημοσύνης. Στις βασικότερες από αυτές συγκαταλέγονται οι: Εξόρυξη Δεδομένων, Μηχανική Μάθηση, Βαθιά Μάθηση, Επεξεργασία Φυσικής Γλώσσας (NLP), Αναγνώριση και Κατηγοριοποίηση Προτύπων, Αυτόματη Σχεδίαση, Ρομποτική, Αυτόνομα Συστήματα και πολλές άλλες.

Ο τομέας της **Εξόρυξης Δεδομένων (Data Mining)** ασχολείται με την ανάλυση μεγάλων όγκων δεδομένων και την εξαγωγή χρήσιμων πληροφοριών και προτύπων που «κρύβονται» πίσω από αυτά. Ο ***Palace (1996)*** όρισε την εξόρυξη δεδομένων ως τη *«Διαδικασία εύρεσης συσχετίσεων ή μοτίβων ανάμεσα σε δεκάδες δεδομένα σε μεγάλες σχεσιακές βάσεις δεδομένων».* Ένας ακόμα ορισμός που έχει δοθεί από τους ***Mal et al (2000)*** είναι: *«Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής τεχνικών τεχνητής νοημοσύνης (όπως προηγμένη μοντελοποίηση και διέγερση κανόνων) σε ένα μεγάλο σύνολο δεδομένων για τον προσδιορισμό των προτύπων στα δεδομένα».*

Με σκοπό να εντοπιστούν οι κρυμμένες σχέσεις και τα πρότυπα μεταξύ των χαρακτηριστικών των καταλυμάτων βραχυχρόνιας μίσθωσης και της τιμής του ενοικίου αυτών χρησιμοποιήθηκαν μέθοδοι **Μηχανικής Μάθησης (Machine Learning)**, που σύμφωνα με τον ***Arthur Samuel (1959)***, ορίζεται ως *«Ένα πεδίο της μελέτης της τεχνητής νοημοσύνης που ασχολείται με το σχεδιασμό και την ανάπτυξη συστημάτων που μπορούν να μάθουν από τα δεδομένα.»*. Στην ουσία αυτό που επιτυγχάνεται με τη βοήθεια της Μηχανικής Μάθησης είναι η δημιουργία συστημάτων τα οποία είναι ικανά να προβλέπουν, κατανοούν και να λαμβάνουν αποφάσεις αφού εκπαιδευτούν με τη βοήθεια κατάλληλων αλγορίθμων και μεγάλου όγκου παρελθοντικών δεδομένων.

## 1.3 Ιστορική εξέλιξη Μηχανικής Μάθησης

Η πορεία της μηχανικής μάθησης, η οποία ξεκινά πολλά χρόνια πριν και συνεχίζει μέχρι και σήμερα είναι εξαιρετικά ενδιαφέρουσα και αξίζει να γίνει μια επιγραμματική αναφορά σε αυτή.

- **1940-1950:** Με την εμφάνιση των πρώτων υπολογιστών, κατά τη διάρκεια του 2ου Παγκόσμιου Πολέμου, εμφανίστηκαν και οι πρώτες ιδέες που αφορούν τη μηχανική μάθηση. Συγκεκριμένα ο Alan Turing μέσα από το άρθρο του "Computing Machinery and Intelligence" πρότεινε το διάσημο «Τεστ Turing», ως μια μέθοδο αξιολόγησης της νοημοσύνης που μπορεί να έχει μια μηχανή.
- **1950-1960:** Κατά τη διάρκεια αυτής της δεκαετίας, η προσπάθεια επικεντρώθηκε στη συμβολική μηχανική μάθηση, με την οποία η ανάλυση και επεξεργασία των δεδομένων πραγματοποιείται με τη βοήθεια υπολογιστικών αλγορίθμων και κανόνων λογικής.
- **1960-1970:** Τα πρώτα νευρωνικά δίκτυα, τα οποία είναι εμπνευσμένα από τη δομή του ανθρώπινου εγκεφάλου, αναπτύσσονται κατά τη δεκαετία του 1960.
- **1970-1980:** Κατά τη δεκαετία του 1980, οι μέθοδοι μηχανικής μάθησης στηρίζονται σε κανόνες, εξαιτίας των περιορισμών υπολογιστικής ισχύος.
- **1980-1990:** Τα νευρωνικά δίκτυα επανέρχονται στο προσκήνιο με την διάδοση της νευρωνικής δικτύωσης και την εισαγωγή νέων αλγορίθμων εκπαίδευσης.
- **2000-2010:** Η αύξηση της υπολογιστικής ισχύος και η χρήση των γραφικών καρτών (GPUs) είχε ως αποτέλεσμα της ευρεία διάδοση της εφαρμογής μεθόδων μηχανικής μάθησης. Η περίοδος αυτή είναι και η εποχή ανάπτυξης νέων τεχνικών, όπως αυτές των Convolutional Neural Networks (CNNs) και Recurrent Neural Networks (RNNs).
- **2010-σήμερα:** Πλέον η μηχανική μάθηση εφαρμόζεται ευρέως και γνωρίζει μεγάλη επιτυχία.

Η ανάπτυξη της μηχανικής μάθησης είναι αποτέλεσμα της συνεισφοράς πλήθους ερευνητών. Στη συνέχεια αναφέρονται ορισμένοι από τους πολλούς.

- **Alan Turing:** Ο Turing, για τον οποίο έγινε μια αναφορά και προηγουμένως, αν και δεν ασχολήθηκε με τη μηχανική μάθηση αυτή καθ' εαυτή, έχει παράγει σημαντικό έργο πάνω στη θεωρία του υπολογισμού και της μηχανικής

νοημοσύνης. Ενώ με το άρθρο του “Computing Machinery and Intelligence”, στο οποίο περιλαμβάνεται το Turing Test, άλλαξε τον τρόπο που αντιλαμβανόμαστε την εκμάθηση και της επεξεργασία της πληροφορίας.

- **Arthur Samuel:** Είναι αυτός που επινόησε τον όρο “machine learning” το 1950. Ο Samuel ανέπτυξε επίσης ένα πρόγραμμα για το παιχνίδι Checkers, μέσω του οποίου η εμπειρία μπορούσε να βελτιώσει την απόδοση.
- **Frank Rosenblatt:** Ο οποίος δημιούργησε το 1950 το Perceptron, το πρώτο μοντέλο νευρωνικού δικτύου. Αν και το συγκεκριμένο είχε περιορισμένη εφαρμογή, αποτέλεσε ένα πολύ σημαντικό βήμα για την ανάπτυξη σύγχρονων νευρωνικών δικτύων.
- **Geoffrey Hinton, Yoshua Bengio και Yann LeCun:** Η συμβολή των τριών αυτών επιστημόνων στην ανάπτυξη των νευρωνικών δικτύων και των βαθιών μοντέλων μάθησης ήταν καθοριστική. Ο Hinton ασχολείται με αλγόριθμους εκπαίδευσης νευρωνικών δικτύων, ενώ οι Bengio και LeCun έχουν εστιάσει στη θεωρητική και πρακτική πλευρά της μηχανικής μάθηση.

Οι προαναφερθέντες είναι μόνο λίγοι από τους πολλούς ερευνητές οι οποίοι συνέβαλαν ουσιαστικά στην ανάπτυξη της μηχανικής μάθησης.

## 1.4 Υπάρχουσα Βιβλιογραφία

Εδώ και αρκετά χρόνια πολλοί είναι οι ερευνητές που έχουν στρέψει το ενδιαφέρον τους στο συγκεκριμένο ζήτημα που πραγματεύεται και η παρούσα. Ιδίως τα τελευταία χρόνια, με την μηχανική μάθηση να εισχωρεί σε ολοένα και περισσότερους τομείς, τέτοιες τεχνικές επιστρατεύονται πλέον από το πλείστο των ερευνητών με σκοπό την επίλυση του ζητήματος της εκτίμησης των τιμών των ακινήτων γενικότερα αλλά και των καταλυμάτων βραχυχρόνιας μίσθωσης ειδικότερα. Ορισμένες μελέτες οι οποίες έχουν ήδη πραγματοποιηθεί πάνω στο συγκεκριμένο ζήτημα παρουσιάζονται στη συνέχεια.

Το 2019 οι Tiancheng Cai, Kevin Han και Han Wu, χρησιμοποιώντας δεδομένα που υπάρχουν στην πλατφόρμα Inside Airbnb, αναπτύσσουν ένα μοντέλο πρόβλεψης τιμών των καταλυμάτων Airbnb στην Μελβούρνη της Αυστραλίας. Κατά την έρευνά τους μελετούν και συγκρίνουν διάφορα μοντέλα πρόβλεψης τιμής, με βάση τα νευρωνικά δίκτυα, καθώς επίσης και παραδοσιακές μεθόδους μηχανικής μάθησης, όπως είναι οι

τεχνικές παλινδρόμησης, τα τυχαία δάση και η ενίσχυση κλίσης. Τελικά αξιολογώντας όλα τα παραπάνω καταλήγουν στο ότι η μέθοδος με την καλύτερη απόδοση είναι αυτή της παλινδρόμησης με ενίσχυση κλίσης.

Το 2020 οι Ang Zhu, Rong Li και Zehao Xie αναλύοντας ένα δείγμα 48.896 ακινήτων στη Νέα Υόρκη από την πλατφόρμα AIRBNB, δημιούργησαν ένα μοντέλο πρόβλεψης των τιμών των ακινήτων βραχυχρόνιας μίσθωσης χρησιμοποιώντας τεχνικές μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας. Με σκοπό τη δημιουργία του τελικού μοντέλου, εξετάστηκαν διάφοροι μέθοδοι όπως: γραμμική παλινδρόμηση, τυχαία δάση και Νευρωνικά Δίκτυα, XGBoost καθώς και συνδυασμός όλων των παραπάνω. Τελικά διαπιστώθηκε ότι οι αλγόριθμοι XGBoost, τυχαία δάση αλλά και ο συνδυασμός τους επέφερε τα καλύτερα αποτελέσματα με την μεγαλύτερη απόδοση.

Το 2021 οι Pouya Rezazadeh Kalehbasti, Liubov Nikolenko και Hoormazd επιδίωξαν να δημιουργήσουν ένα μοντέλο το οποίο θα προβλέπει τις τιμές των καταλυμάτων βραχυχρόνιας μίσθωσης και θα βασίζεται τόσο στα χαρακτηριστικά των ακινήτων όσο και στις κριτικές αυτών. Πολλοί αλγόριθμοι χρησιμοποιήθηκαν για το σκοπό αυτό, όπως γραμμικής παλινδρόμησης, K-means, Support Vector Machine, νευρωνικά δίκτυα και εκπαιδεύτηκαν με ένα σετ δεδομένων από ακίνητα της πλατφόρμας AIRBNB στη Νέα Υόρκη. Επίσης πραγματοποιήθηκε ανάλυση συναισθήματος στις κριτικές των πελατών, οι οποίες βοήθησαν στο να βελτιωθεί η απόδοση του μοντέλου. Τελικά, αποδείχθηκε πως ο αλγόριθμος Support Vector Machine έδωσε τα αποτελέσματα με τη μεγαλύτερη ακρίβεια.

Το 2023 οι Jinwen Tang, Jinlin Cheng, Min Zhang κατανοώντας το πόσο κρίσιμη είναι η ακριβής και ορθή κοστολόγηση των καταλυμάτων βραχυχρόνιας μίσθωσης αλλά και η εξακρίβωση των παραγόντων που την επηρεάζουν, εφάρμοσαν μεθόδους μηχανικής μάθησης με σκοπό τη δημιουργία ενός αλγόριθμου πρόβλεψής τους. Συνέλλεξαν, λοιπόν, δεδομένα από καταχωρήσεις του AIRBNB στο Σίδνεϊ της Αυστραλίας και χρησιμοποίησαν 10 Αλγόριθμους Μηχανικής Μάθησης. Συνολικά, εντόπισαν 35 μεταβλητές, μαζί με αυτή της τιμής. Για την αξιολόγηση των αλγόριθμων που χρησιμοποίησαν, εφάρμοσαν το Student’s t-test, το τετραγωνικό μέσο σφάλμα και την $R^2$ τιμή. Κατέληξαν στο ότι ο αλγόριθμος CatBoostRegressor είχε την καλύτερη απόδοση, και σύμφωνα με τον αλγόριθμο αυτό οι βασικοί παράγοντες που επηρεάζουν

την τιμολόγηση είναι ο μέγιστος αριθμός επισκεπτών, ο αριθμός των υπνοδωματίων και το εάν το δωμάτιο είναι ιδιωτικό ή όχι. Βάσει αυτών των αποτελεσμάτων, οι εγγεγραμμένοι οικοδεσπότες μπορούν να αποκτήσουν έγκαιρες πληροφορίες σχετικά με την αγορά ενοικίασης κατοικιών για να καθορίσουν λογικές τιμές.

## 1.5 Στόχοι και πεδίο έρευνας

Όπως έχει αναφερθεί ήδη και παραπάνω, η ανάγκη πρόβλεψης της πιθανής τιμής ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης λόγω της ραγδαίας ανάπτυξης που γνωρίζει ο συγκεκριμένος κλάδος σήμερα, αποτελεί πλέον μια επιτακτική ανάγκη για τους ταξιδιώτες – ενοικιαστές και κυρίως για τους ιδιοκτήτες. Αυτός ακριβώς είναι και ο στόχος της παρούσας.

Συγκεκριμένα γίνεται προσπάθεια δημιουργίας ενός μοντέλου πρόβλεψης της τιμής ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης με τη μεγαλύτερη δυνατή ακρίβεια. Για το σκοπό αυτό δημιουργούνται 2 μοντέλα, Νευρωνικών Δικτύων & Τυχαίων Δασών, των οποίων τελικά η απόδοση συγκρίνεται έτσι ώστε να διαπιστωθεί ποιο δίνει τα πιο ακριβή αποτελέσματα. Αυτό γίνεται με τη χρήση τεχνικών τεχνητής νοημοσύνης και μεθόδων μηχανικής μάθησης. Τα μοντέλα εκπαιδεύονται με πλήθος δεδομένων από καταλύματα βραχυχρόνιας μίσθωσης που έχουν καταχωρηθεί στην πλατφόρμα AIRBNB με σκοπό την εύρεση των «κρυφών» σχέσεων που υπάρχουν μεταξύ των χαρακτηριστικών τους και της τιμής εκμίσθωσης αυτών.

## 1.6 Δομή Εργασίας

Στην 1η Ενότητα γίνεται μια εισαγωγή τόσο στο θέμα που πρόκειται να αναλυθεί όσο και στις μεθόδους οι οποίες πρόκειται να χρησιμοποιηθούν. Στη 2η Ενότητα αναλύονται σε θεωρητικό επίπεδο όλες οι τεχνικές και μεθοδολογίες που πρόκειται να εφαρμοστούν. Στην 3η Ενότητα παρουσιάζεται αναλυτικά το λογισμικό Rapid Miner το οποίο πρόκειται να χρησιμοποιηθεί καθώς και όλα τα βήματα, σε θεωρητικό επίπεδο, που ακολουθούνται για την ανάπτυξη των μοντέλων. Στην 4η ενότητα αναλύονται πλήρως όλες οι εφαρμογές που ακολουθήθηκαν πρακτικά με σκοπό την δημιουργία των μοντέλων πρόβλεψης των τιμών ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης, παρουσιάζονται τα αποτελέσματα της πρακτικής εφαρμογής τους και πραγματοποιείται σύγκριση τις ακρίβειας των 2 μοντέλων με σκοπό την επιλογή του αποτελεσματικότερου. Τέλος, και στην 5η Ενότητα αναλύονται τα

συμπεράσματα όλων των παραπάνω και παρατίθενται ορισμένες προτάσεις για μελλοντική και ευρύτερη διερεύνηση του συγκεκριμένου ζητήματος.

# 2. Θεωρητικό υπόβαθρο

Στην ενότητα αυτή θα αναλυθεί σε θεωρητικό επίπεδο ο τρόπος και οι μεθοδολογίες που ακολουθούνται γενικότερα αλλά και ειδικότερα σε όσες χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία, με σκοπό να πραγματοποιηθεί η εκτίμηση των τιμών ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης με την χρήση αλγορίθμων μηχανικής μάθησης. Τέλος, παρουσιάζονται συνοπτικά τα διαθέσιμα λογισμικά τα οποία χρησιμοποιούνται για τους σκοπούς αυτούς.

## 2.1 Μεθοδολογίες

Με σκοπό να πραγματοποιηθεί Εξόρυξη Γνώσης από δεδομένα με αλγόριθμους μηχανικής μάθησης, ακολουθείται μια βασική μεθοδολογία που αποτελείται από τα εξής στάδια:

1. **Συλλογή Δεδομένων**

   Πρώτο βήμα αποτελεί η συλλογή των απαραίτητων δεδομένων από διάφορες πηγές (π.χ. βάσεις δεδομένων, αισθητήρες κ.λπ.). Στην παρούσα διπλωματική χρησιμοποιήθηκαν δεδομένα από την ιστοσελίδα insideairbnb, μια ανεξάρτητη ιστοσελίδα που παρέχει δεδομένα και αναλύσεις σχετικά με τις καταχωρίσεις της πλατφόρμας Airbnb σε διάφορες πόλεις παγκοσμίως.

2. **Προ-επεξεργασία Δεδομένων**

   Σε αυτό το στάδιο γίνεται καθαρισμός των δεδομένων με σκοπό να βελτιωθεί η ποιότητα και η ακρίβεια αυτών. Εδώ περιλαμβάνεται η απομάκρυνση σφαλμάτων, η διαχείριση ελλείψεων ή ασαφών δεδομένων, η εξομάλυνση των τιμών και η μετατροπή των δεδομένων στην κατάλληλη μορφή.

3. **Μετασχηματισμός Δεδομένων**

   Στη συνέχεια τα δεδομένα μετασχηματίζονται στην κατάλληλη μορφή. Η συγκεκριμένη επιλογή είναι πολύ κρίσιμη καθώς έτσι μπορεί να μειωθεί η πολυπλοκότητα και να βελτιωθεί η απόδοση των αλγόριθμων.

4. **Μοντελοποίηση**

   Στο στάδιο αυτό επιλέγεται η καλύτερη μαθησιακή μέθοδος του μοντέλου και εφαρμόζονται οι αντίστοιχοι αλγόριθμοι Μηχανικής Μάθησης με σκοπό την δημιουργία μοντέλων ικανών να προβλέψουν δεδομένα. Στο επόμενο υπό-

κεφάλαιο θα αναπτυχθούν αναλυτικά σε θεωρητικό επίπεδο όλες οι τεχνικές Μηχανικής Μάθησης.

5. **Αξιολόγηση μοντέλου**

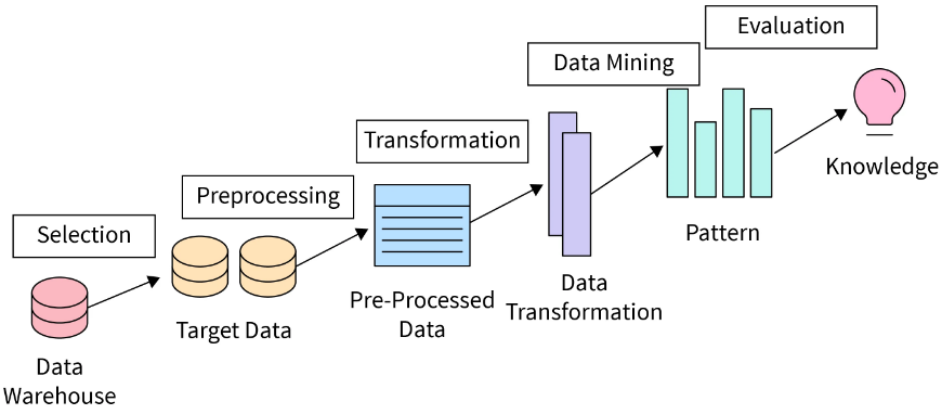
   Μετά τη δημιουργία των μοντέλων, αυτά θα πρέπει να αξιολογηθούν ως προς την ακρίβεια, τη σταθερότητα και την ικανότητα τους να γενικεύουν τα αποτελέσματα σε νέα δεδομένα. Η αξιολόγηση γίνεται μέσω δεικτών μετρικών αποδόσεων (όπως: Mean Squared Error, R-squared, precision, recall κ.λπ.). Οι μέθοδοι αυτοί αξιολόγησης θα αναλυθούν περαιτέρω σε επόμενο υπό-κεφάλαιο.

6. **Ερμηνεία των αποτελεσμάτων**

   Σε αυτό το στάδιο τα αποτελέσματα παρουσιάζονται με κατανοητό και χρήσιμο για το επιδιωκόμενο αποτέλεσμα τρόπο (π.χ. μέσω αναφορών, γραφημάτων κ.λπ.)

7. **Ανάπτυξη και χρήση μοντέλου**

   Τέλος, το μοντέλο αναπτύσσεται σε ένα λειτουργικό σύστημα ώστε να μπορεί να χρησιμοποιηθεί σε πραγματικό χρόνο.




*Σχήμα 2.1 Βήματα Διαδικασίας Εξόρυξης Δεδομένων (πηγή: KDD in Data Mining- Scaler Topics)*

## 2.2 Μηχανική Μάθηση

Με τη βοήθεια της Μηχανικής Μάθησης είναι εφικτός ο σχεδιασμός και η δημιουργία συστημάτων τα οποία μπορούν να μάθουν από δεδομένα. Ανάλογα τώρα με το είδος των δεδομένων που πρόκειται να επεξεργαστούν υπάρχουν 3 διαφορετικοί τρόποι μάθησης και αρκετοί διαφορετικοί αλγόριθμοι, σε κάθε μια από αυτές τις κατηγορίες, με τη βοήθεια των οποίων μπορεί αυτό να καταστεί εφικτό.

Οι 3 αυτές κατηγορίες της Μηχανικής Μάθησης είναι οι εξής:

- → ***Επιβλεπόμενη Μάθηση (Supervised Learning)***
- → ***Μη επιβλεπόμενη Μάθηση (Unsupervised Learning)***
- → ***Ενισχυτική Μάθηση (Reinforcement Learning)***

### *2.2.1 Επιβλεπόμενη Μάθηση (Supervised Learning)*

Στην κατηγορία αυτή το μοντέλο εκπαιδεύεται με επισημασμένα δεδομένα (labeled data). Αυτό πρακτικά σημαίνει πως το σετ δεδομένων εκπαίδευσης (training set) αποτελείται από ζεύγη και σε κάθε είσοδο αντιστοιχεί μια σωστή έξοδος [π.χ. ένα σετ δεδομένων εκπαίδευσης θα μπορούσε να περιλαμβάνει διάφορα χαρακτηριστικά από έναν αριθμό ακινήτων (τετραγωνικά, την περιοχή, έτος κατασκευής ακινήτου, όροφος κ.λπ.), που θα αποτελούσαν τα δεδομένα εισόδου και την τιμή πώλησής τους που θα αποτελούσε την αντίστοιχη έξοδο]. Με βάση τα δεδομένα αυτά, το μοντέλο εκπαιδεύεται, μαθαίνοντας τη σχέση μεταξύ των ζευγών εισόδου – εξόδου με σκοπό να μπορεί να γενικεύει και να κάνει προβλέψεις για νέες εισόδους, άγνωστες.

Η Επιβλεπόμενη Μάθηση μπορεί να λύσει βασικά 2 κατηγορίες προβλημάτων. Αυτή της ταξινόμησης και αυτή της παλινδρόμησης.

***<u>Προβλήματα ταξινόμησης</u>***

Με την ταξινόμηση το μοντέλο μπορεί να προβλέψει την κατηγορία στην οποία ανήκει η μεταβλητή εισόδου. Παραδείγματος χάριν μπορεί να ταξινομήσει τα email σε spam ή μη, να αναγνωρίσει και να ταξινομήσει εικόνες σε διάφορες κατηγορίες, να ανιχνεύσει πιθανή απάτη σε συναλλαγές κατηγοριοποιώντας τες ως «ύποπτες» ή «κανονικές» και πολλά άλλα.

Ορισμένοι από τους αλγόριθμους Μηχανικής Μάθησης που χρησιμοποιούνται με σκοπό την επίλυση προβλημάτων ταξινόμησης είναι οι: Logistic Regression, Naïve Bayes, Nearest Neighbor, Support Vector Machines, Decision Trees, Boosted Trees, Random Forest, Neural Networks.

***<u>Προβλήματα Παλινδρόμησης</u>***

Στα προβλήματα παλινδρόμησης το μοντέλο στοχεύει στην πρόβλεψη μιας συγκεκριμένης συνεχόμενης αριθμητικής τιμής ή οποία έχει προκύψει από τις πληροφορίες των ανεξάρτητων μεταβλητών εισόδου. Για παράδειγμα μπορεί να

προβλέπει την τιμή πώλησης ενός ακινήτου την θερμοκρασία, τον αριθμού πωλήσεων, την κατανάλωση ενέργειας και πολλά άλλα.

Ορισμένοι από τους αλγόριθμους Μηχανικής Μάθησης που χρησιμοποιούνται με σκοπό την επίλυση προβλημάτων παλινδρόμησης είναι οι: Neural Networks, Linear Regression, Logistic Regression, Clustering, K-means, Support Vector Machines, Decision Trees, Random Forest, Naïve Bayes.

Η παρούσα διπλωματική εργασία πραγματεύεται ένα πρόβλημα παλινδρόμησης καθώς πρόκειται να γίνει προσπάθεια πρόβλεψης μιας συγκεκριμένης, συνεχόμενης αριθμητικής τιμής (μίσθωμα καταλυμάτων).

### *2.2.2 Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)*

Σε αυτή την κατηγορία, ο αλγόριθμος εκπαιδεύεται από δεδομένα τα οποία δεν είναι επισημασμένα και στα οποία δεν υπάρχουν προκαθορισμένοι σωστοί έξοδοι. Στόχος είναι η ανακάλυψη συσχετίσεων, προτύπων και μοτίβων στα δεδομένα εισόδου χωρίς την καθοδήγηση κάποιου σωστού αποτελέσματος.

Στη συνέχεια αναλύονται ορισμένες από τις τεχνικές της Μη Επιβλεπόμενης Μάθησης.

***Ομαδοποίηση (Clustering)***

Εδώ σκοπός είναι η ομαδοποίηση των δεδομένων βάσει της μεταξύ τους ομοιότητας και συγκριτικά με τα δεδομένα που ανήκουν σε διαφορετικές ομάδες. Ορισμένοι από τους αλγόριθμους που επιδιώκουν να λύσουν το συγκεκριμένο πρόβλημα είναι οι: k-means, ιεραρχική ομαδοποίηση (Hierarchical Clustering) (SLINK, CLINK) και πυκνωτικού αλγόριθμου (DBSCAN, OPTICKS). Κάθε ένα από αυτούς προσπαθεί να λύσει το συγκεκριμένο πρόβλημα βασιζόμενος σε εντελώς διαφορετική λογική.

***Μείωση Διαστάσεων (Dimensionality Reduction)***

Ο στόχος είναι η μείωση του αριθμού των χαρακτηριστικών (μεταβλητών) στα δεδομένα, διατηρώντας παράλληλα τις όσο το δυνατών περισσότερες και βασικότερες πληροφορίες. Μπορεί να επιφέρει χρήσιμα αποτελέσματα στην οπτικοποίηση των δεδομένων και στην εύρεση κρυφών δεσμών. Ορισμένοι αλγόριθμοι που εφαρμόζονται στην συγκεκριμένη τεχνική είναι οι: Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA), Ενσωμάτωση Στοχαστικών Γειτόνων με κατανομή t [t-

Distributed Stochastic Neighbor Embedding (t-SNE)] και UMAP (Uniform Manifold Approximation and Projection).

***Συσχετιστική Ανάλυση (Association Analysis)***

Στην συσχετική ανάλυση επιδιώκεται η εύρεση πιθανών σχέσεων μεταξύ χαρακτηριστικών μέσα στα δεδομένα. Ορισμένοι αλγόριθμοι που εφαρμόζονται στην συγκεκριμένη τεχνική είναι οι: Apriori και FP-Growth (Frequent Pattern Growth).

***Ανίχνευση Ανωμαλιών (Anomaly Detection)***

Με τη συγκεκριμένη τεχνική αναζητούνται δεδομένα τα οποία δεν συμφωνούν με το γενικό πρότυπο του συνόλου των δεδομένων. Εφαρμόζεται σε περιπτώσεις ανίχνευσης απάτης, αναγνώρισης σφαλμάτων και ασφάλειας δικτύων. Αλγόριθμοι που εφαρμόζονται εδώ είναι οι: Isolation Forest και One-Class SVM.

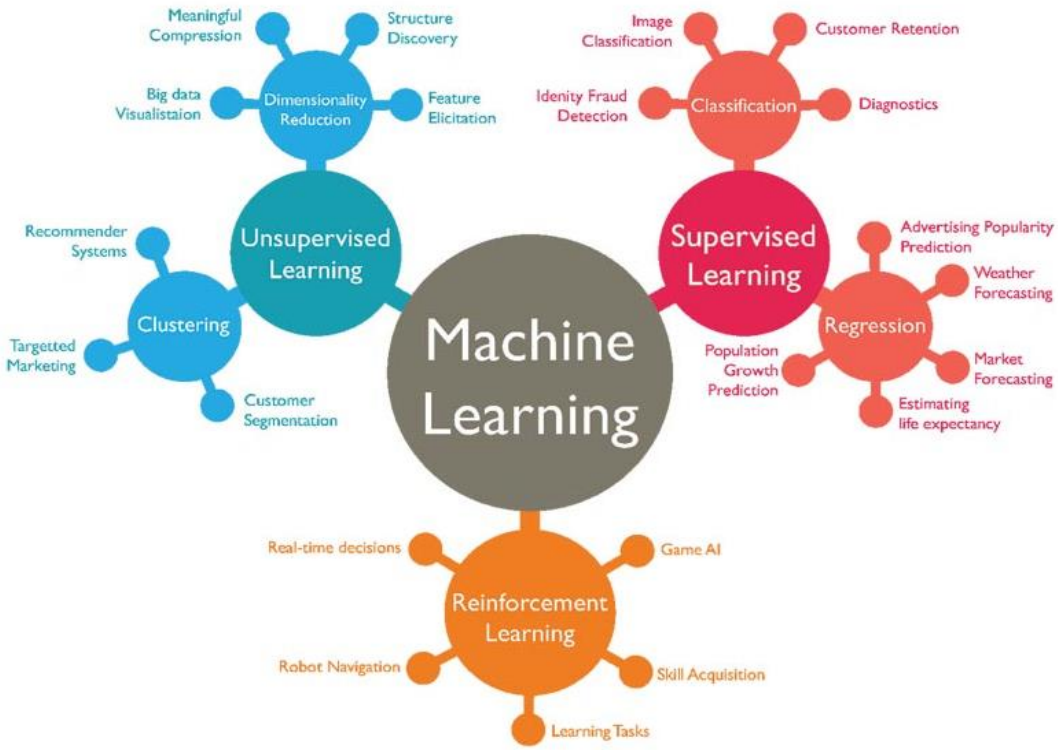
Η Μη Επιβλεπόμενη Μάθηση αποτελεί ένα ισχυρό εργαλείο στην κατανόηση και την εξερεύνηση δεδομένων, ειδικά όταν υπάρχει ασάφεια για το τι πρέπει να αναζητηθεί εκ των προτέρων.

### *2.2.3 Ενισχυτική Μάθηση (Reinforcement Learning)*

Η συγκεκριμένη είναι μια διαδικασία μάθησης όπου το μοντέλο λαμβάνει αποφάσεις αλληλεπιδρώντας με το περιβάλλον του. Λαμβάνοντας ανατροφοδότηση με τη μορφή ανταμοιβών ή ποινών, το μοντέλο εκπαιδεύεται ώστε να παίρνει αποφάσεις με σκοπό τη μεγιστοποίηση της συνολικής ανταμοιβής με την πάροδο του χρόνου. Η συγκεκριμένη τεχνική αφορά μια πολύπλοκη προσέγγιση που συνδυάζει στοιχεία από τη θεωρία παιχνιδιών, τον έλεγχο και τη βελτιστοποίηση, και είναι ιδανική για προβλήματα όπου οι ενέργειες έχουν επιπτώσεις που επηρεάζουν μελλοντικές αποφάσεις.

Γενικά και για να συνοψίσουμε, η Μηχανική Μάθηση περιλαμβάνει μια αλληλουχία μεθοδολογιών, όπως απεικονίζονται στο παρακάτω *Σχήμα 2.1*.

Κάτι που αξίζει να σημειωθεί στο σημείο αυτό είναι πως ορισμένα μοντέλα ενδέχεται να ανήκουν σε περισσότερες από μια κατηγορίες και ορισμένοι αλγόριθμοι μπορούν να προσαρμοστούν σε διαφορετικά είδη μάθησης, με την κατάλληλη πάντα προσαρμογή των δεδομένων εισόδου – εξόδου.




*Σχήμα 2.2 Αλληλουχία Μεθοδολογιών Μηχανικής Μάθησης (πηγή: nowmag.gr)*

## 2.3 Τεχνητά Νευρωνικά Δίκτυα

Εδώ αναλύεται σε θεωρητικό επίπεδο ο τρόπος λειτουργίας του αλγόριθμου των Τεχνητών Νευρωνικών Δικτύων (Neural Networks) γενικότερα αλλά και ειδικότερα όσον αφορά την εφαρμογή του στο ζήτημα που πραγματεύεται η παρούσα διπλωματική.

Τα Τεχνητά Νευρωνικά Δίκτυα (Neural Networks) αποτελούν ένα μαθηματικό μοντέλο το οποίο είναι εμπνευσμένο και προσομοιώνει τη δομή και λειτουργία του ανθρώπινου εγκεφάλου. Όπως τα νευρωνικά δίκτυα του εγκεφάλου λοιπόν έτσι και τα τεχνητά αφορούν ένα σύστημα διασυνδεδεμένων νευρώνων (κόμβων), που έχουν τη δυνατότητα να ανταποκρίνονται σε ερεθίσματα που δέχονται στην είσοδό τους και να προσαρμόζονται στο περιβάλλον τους. Επομένως στη γενική μορφή τους τα νευρωνικά δίκτυα αποτελούν μια «μηχανή» η οποία έχει σχεδιαστεί να μοντελοποιεί τον τρόπο με τον οποίο ο ανθρώπινος εγκέφαλος εκτελεί μια συγκεκριμένη λειτουργία.

Τα νευρωνικά δίκτυα αποτελούν το θεμέλιο της βαθιάς μάθησης (deep learning) και ορισμένες κατηγορίες προβλημάτων στις οποίες έχουν αποδειχθεί εξαιρετικά ισχυρά

εργαλεία είναι η δημιουργία μοντέλων για προβλέψεις μέσω προσεγγίσεων μιας άγνωστης συνάρτησης, η αναγνώριση προτύπων, η επεξεργασία φυσικής γλώσσας και πολλά άλλα.

### *2.3.1 Δομή & Λειτουργία Τεχνητών Νευρώνων*

Όπως και στα βιολογικά νευρωνικά δίκτυα έτσι και στα τεχνητά, θεμελιώδη μονάδα επεξεργασίας της πληροφορίας αποτελεί ο τεχνητός Νευρώνας ή αλλιώς Κόμβος, ο οποίος προσομοιώνει τη λειτουργία του βιολογικού νευρώνα του εγκεφάλου.

Ένας νευρώνας έχει τα εξής χαρακτηριστικά:

***Είσοδοι (Inputs):*** Κάθε νευρώνας δέχεται πληροφορίες από μια ή περισσότερες εισόδους οι οποίες αντιπροσωπεύουν χαρακτηριστικά των δεδομένων που μπορεί να προέρχονται είτε από το εξωτερικό περιβάλλον, είτε από νευρώνες σε προηγούμενα επίπεδα του δικτύου. Για παράδειγμα και στην προκειμένη περίπτωση μελέτης κάθε νευρώνας στο επίπεδο εισόδου μπορεί να αντιστοιχεί στο εμβαδόν του καταλύματος, τον αριθμό των δωματίων, την τοποθεσία, την ημερομηνία ενοικίασης, τις κριτικές, σε άλλες πιθανές παροχές (π.χ. Wi-Fi) και πολλά άλλα.

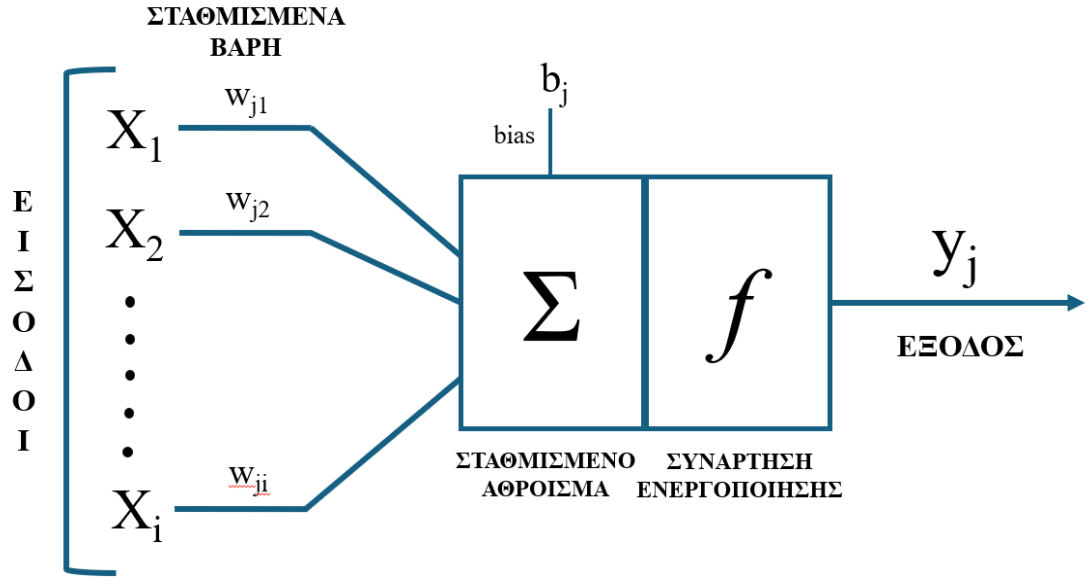
***Σταθμισμένο Άθροισμα (Weighted SUM):*** Κάθε είσοδος του νευρώνα πολλαπλασιάζεται με το σταθμισμένο βάρος που της αντιστοιχεί και που ως στόχο έχει την αύξηση ή την μείωση της ισχύος του εισερχόμενου σήματος. Στη συνέχεια οι τιμές αυτές αθροίζονται και σε αυτές προστίθεται και μια εξωτερική παράμετρος μετατόπισης (bias) με σκοπό την αύξηση της ακρίβειας του τελικού Σταθμισμένου Αθροίσματος. Εδώ ένα παράδειγμα συναφές με την περίπτωση μελέτης που πραγματεύεται η παρούσα θα μπορούσε να είναι ο πολλαπλασιασμός των χαρακτηριστικών των υπό εξέταση καταλυμάτων (εμβαδόν, αριθμός υπνοδωματίων κ.λπ.) με ένα συντελεστή (σταθμισμένο βάρος) ανάλογα με την σημαντικότητα του χαρακτηριστικού και στη συνέχεια η πρόσθεση όλων αυτών με σκοπό την εξαγωγή του σταθμισμένου αθροίσματος.

***Συνάρτηση Ενεργοποίησης (ή Μεταφοράς) (Activation Function):*** Στη συνέχεια το σταθμισμένο άθροισμα περνάει από τη Συνάρτηση Ενεργοποίησης η οποία προσδίδει τη μη γραμμικότητα στα νευρωνικά δίκτυα, το κάνει πιο ικανό να αντιμετωπίσει πολύπλοκα μοτίβα στα δεδομένα και μετασχηματίζει το σταθμισμένο άθροισμα σε τιμή

εξόδου του νευρώνα. Οι πιο κοινές συναρτήσεις ενεργοποίησης είναι οι: Συνάρτηση Βηματικής Ενεργοποίησης (Step Function), Σιγμοειδής Συνάρτηση (Sigmoid Function), Συνάρτηση ReLU (Rectified Linear Unit) και Συνάρτηση tanh (Hyperbolic Tangent). Στην υπό μελέτη περίπτωση της εκτίμησης της τιμής ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης, η συνάρτηση ενεργοποίησης επιτρέπει στο νευρωνικό δίκτυο να εξετάσει μη γραμμικές σχέσεις όπως π.χ. το πως η τοποθεσία και το μέγεθος μπορεί να επηρεάζουν την τιμή ενός καταλύματος με πολύπλοκο τρόπο που δεν είναι απλά προσθετικός ή γραμμικός.

***Έξοδος (Output):*** Η έξοδος του τεχνητού νευρώνα αφορά το αποτέλεσμα της συνάρτησης ενεργοποίησης και αυτό μπορεί είτε να συνδεθεί και να περάσει στον επόμενο νευρώνα μέσω συνάψεων που υλοποιούνται με τη μορφή των σταθμισμένων βαρών (βαρών συνάψεων), είτε να είναι το τελικό αποτέλεσμα του δικτύου. Για παράδειγμα στην υπό μελέτη περίπτωση, η έξοδος αφορά μια συνεχόμενη τιμή, δηλαδή την εκτιμώμενη τιμή ενοικίασης του εκάστοτε καταλύματος βραχυχρόνιας μίσθωσης.

Στο παρακάτω *Σχήμα 2.2* φαίνεται η δομή ενός τεχνητού νευρώνα με όλα τα χαρακτηριστικά που αναφέραμε προηγουμένως.




*Σχήμα 2.3 Σχηματική αναπαράσταση της Δομής ενός Τεχνητού Νευρών (πηγή: Ιδία Επεξεργασία)*

### *2.3.2 Δομή & Λειτουργία Τεχνητών Νευρωνικών Δικτύων*

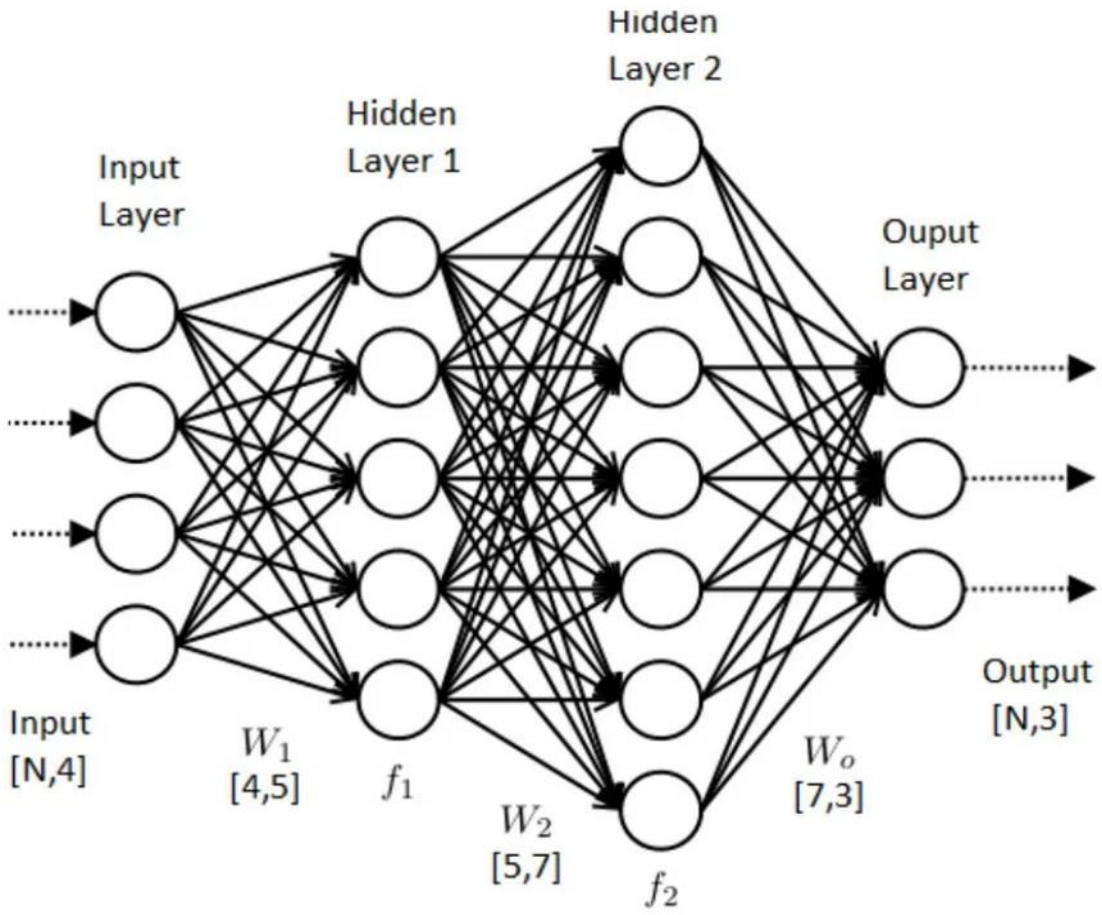
Τα Τεχνητά Νευρωνικά Δίκτυα αποτελούνται από πολλούς τέτοιους παράλληλα διατεταγμένους Τεχνητούς Νευρώνες. Κάθε ένα τέτοιο σύμπλεγμα αποτελεί μια στοιβάδα - στρώμα (Layer). Ένα τυπικό νευρωνικό δίκτυο περιλαμβάνει μια ακολουθία στοιβάδων – στρωμάτων συνδεδεμένα είτε μερικώς είτε πλήρως μεταξύ τους. Τα επίπεδα από τα οποία μπορεί να αποτελούνται τα Τεχνητά Νευρωνικά Δίκτυα είναι τα εξής:

***Επίπεδο Εισόδου (Input Layer):*** Το επίπεδο αυτό περιλαμβάνει τους νευρώνες που αντιστοιχούν στα χαρακτηριστικά των δεδομένων εισόδου. Αυτό είναι το πρώτο στρώμα με το οποίο το δίκτυο επικοινωνεί με το χρήστη. Στο επίπεδο αυτό δεν πραγματοποιείται καμία επεξεργασία και τα δεδομένα εισέρχονται απλά στο δίκτυο.

***Κρυφά Επίπεδα (Hidden Layers):*** Στα κρυφά επίπεδα συγκαταλέγονται τα συμπλέγματα τεχνητών νευρώνων οι οποίοι επεξεργάζονται τα δεδομένα και εκτελούν τις απαραίτητες υπολογιστικές διαδικασίες. Σε κάθε νευρώνα που ανήκει στα επίπεδα αυτά εφαρμόζονται τόσο το σταθμισμένο άθροισμα όσο και η συνάρτηση ενεργοποίησης με σκοπό την εξαγωγή προτύπων και σχέσεων από τα δεδομένα. Τα κρυφά επίπεδα δεν υπάρχουν σε κάθε νευρωνικό δίκτυο και αποτελούν τα επίπεδα στα οποία το δίκτυο δεν έρχεται καθόλου σε επαφή με το χρήστη. Εδώ οι συνδέσεις είναι πλήρης με το επόμενο επίπεδο, δηλαδή κάθε νευρώνας αυτού του επιπέδου συνδέεται με το επόμενο.

***Επίπεδο Εξόδου (Output Layer):*** Αποτελεί το τελευταίο επίπεδο του δικτύου και περιλαμβάνει νευρώνες που παράγουν το τελικό αποτέλεσμα και δίνει την ανταπόκριση του δικτύου στις δεδομένες εισόδους. Ο αριθμός των νευρώνων που υπάρχουν στο επίπεδο αυτό εξαρτάται από το είδος του προβλήματος. Το επίπεδο αυτό αποτελεί τη $2^{η}$ και τελευταία επαφή του χρήστη με το δίκτυο.

Στο παρακάτω *Σχήμα 2.4* αναπαρίσταται η δομή ενός τυπικού Τεχνητού Νευρωνικού Δικτύου με 2 κρυφά επίπεδα.




*Σχήμα 2.4 Τυπική Δομή Τεχνητού Νευρωνικού Δικτύου με 2 κρυφά επίπεδα (πηγή: medium.com)*

Υπάρχουν Τεχνητά Νευρωνικά Δίκτυα τα οποία δεν περιέχουν καμία κρυφή στιβάδα και τα οποία ονομάζονται “single layer networks”. Τα δίκτυα αυτά είναι ικανά να λύσουν μόνο γραμμικά και απλά προβλήματα. Υπάρχουν όμως και Δίκτυα τα οποία περιέχουν μία ή περισσότερες κρυφές στοιβάδες που ονομάζονται “multilayer networks” και τα οποία είναι πολύ πιο ισχυρά και ικανά να μάθουν σύνθετες, μη γραμμικές σχέσεις από τα δεδομένα καθιστώντας τα κατάλληλα για πιο απαιτητικά προβλήματα. Σε αυτά ανήκει το πρόβλημα που πραγματεύεται η παρούσα, η εκτίμηση δηλαδή της τιμής ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης. Τέλος, αξίζει να σημειωθεί ότι τα νευρωνικά δίκτυα τα οποία αποτελούνται από πολλά κρυφά επίπεδα συχνά ονομάζονται ως βαθιά νευρωνικά δίκτυα (Deep Neural Networks).

### *2.3.3 Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων*

Η εκπαίδευση των Τεχνητών Νευρωνικών Δικτύων δεν αφορά στον κανονισμό σαφών κανόνων λειτουργίας αλλά στην εύρεση κατάλληλων συντελεστών βάρους και

σφαλμάτων μεροληψίας (bias), που επιτρέπουν στο δίκτυο να λειτουργεί σαν να γνώριζε τους κανόνες.

***Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων με Επιβλεπόμενη Μάθηση***

Με σκοπό να ολοκληρωθεί ορθά η εκπαίδευση ενός Τεχνητού Νευρωνικού Δικτύου με τη μέθοδο της Επιβλεπόμενης Μάθησης ακολουθούνται τα παρακάτω στάδια.

Αρχικά τα δεδομένα εισόδου εισέρχονται στο Δίκτυο και τα αρχικά βάρη και σφάλματα μεροληψίας (bias) αυτού αρχικοποιούνται με τυχαίες τιμές. Στη συνέχεια και σε κάθε κύκλο εκπαίδευσης τα δεδομένα προωθούνται μέσα σε όλο το δίκτυο (forward propagation) ενώ πολλαπλασιάζονται με τα βάρη κάθε σύνδεσης (συνάψεων) και υπολογίζεται το σταθμισμένο άθροισμα για κάθε νευρώνα. Το αποτέλεσμα περνά από τη συνάρτηση ενεργοποίησης, η οποία αποφασίζει εάν και κατά πόσο θα ενεργοποιηθεί ο κάθε νευρώνας ο οποίος με τη σειρά του το προωθεί στους επόμενους μέχρι τελικά να φτάσει στο επίπεδο εξόδου.

Το αποτέλεσμα που έχει προωθηθεί στην έξοδο συγκρίνεται με την πραγματική τιμή (label) που θέλαμε να προβλέψουμε και ακολουθεί ο υπολογισμός του σφάλματος μέσω μιας συνάρτησης κόστους (cost function / loss function) όπως η Mean Squar Error (MSE) για προβλήματα παλινδρόμησης και η Cross-Entropy Loss για προβλήματα ταξινόμησης. Αυτή η συνάρτηση υπολογίζει το πόσο απέχει η έξοδος του δικτύου από την πραγματική έξοδο.

Ακολουθεί η οπισθοδρόμηση (Backpropagation), μια κρίσιμη για την εκπαίδευση του δικτύου διαδικασία κατά την οποία το υπολογιζόμενο σφάλμα μεταδίδεται προς τα πίσω, ξεκινώντας από το επίπεδο εξόδου προς τα κρυφά επίπεδα και καταλήγει στο επίπεδο εισόδου. Κατά την οπισθοδρόμηση ενημερώνονται τα βάρη και τα biases του δικτύου, μέσω μιας μεθόδου βελτιστοποίησης όπως ο Gradient Descent, έτσι ώστε να μειωθεί το σφάλμα.

Η όλη διαδικασία που περιγράφεται παραπάνω επαναλαμβάνεται για πολλούς κύκλους εκπαίδευσης – εποχές με το ίδιο σύνολο δεδομένων και σταματάει όταν το ελάχιστο τετραγωνικό σφάλμα στο σετ δεδομένων εκπαίδευσης είναι αρκετά μικρό ή όταν έχει περάσει ένας συγκεκριμένος αριθμός επαναλήψεων τον οποίον και έχουμε εμείς ορίσει. Επίσης υπάρχουν ακόμα 2 παράγοντες οι οποίοι καθορίζονται εξωτερικά και

επηρεάζουν την ταχύτητα της εκπαίδευσης και την εγκυρότητα του παραγόμενου μοντέλου. Αυτά είναι ο ρυθμός μάθησης (learning rate) που ελέγχει το πόσο γρήγορα μαθαίνει το δίκτυο και το μέγεθος παρτίδας (batch size), δηλαδή το πλήθος των δεδομένων «παραδειγμάτων» που περνάνε ταυτόχρονα μέσα από το δίκτυο.

Κρίσιμο σε αυτό το σημείο είναι να σημειωθεί πως θα πρέπει να ληφθεί υπόψιν και να γίνει προσπάθεια αποφυγής της υπερβολικής εξειδίκευσης ενός μοντέλου στα δεδομένα εκπαίδευσης (υπερπροσαρμογή / overfitting), που έχει σαν αποτέλεσμα την αδυναμία του μοντέλου να γενικεύει και να αποδίδει καλά σε νέα, άγνωστα δεδομένα. Αυτό μπορεί να γίνει με την εφαρμογή ορισμένων τεχνικών όπως η κανονικοποίηση, το dropout, η διασταυρούμενη επικύρωση (cross validation), η μείωση της πολυπλοκότητας του μοντέλου κ.α.

***Εκπαίδευση Τεχνητών Νευρωνικών Δικτύων & Δεδομένα***

Με σκοπό την ορθή εκπαίδευση ενός Νευρωνικού Δικτύου απαιτείται μεγάλος όγκος δεδομένων. Αυτά χωρίζονται στις εξής ομάδες:

- → **Σετ εκπαίδευσης (Training Set):** Αφορά το σύνολο των δεδομένων που χρησιμοποιούνται για την εκπαίδευση του Δικτύου.
- → **Σετ επικύρωσης (Validation Set):** Αφορά ένα ξεχωριστό σύνολο δεδομένων το οποίο χρησιμοποιείται με σκοπό την παρακολούθηση της απόδοσης του δικτύου.
- → **Σετ δοκιμής (Test Set):** Το οποίο χρησιμοποιείται στο τέλος της εκπαίδευσης με σκοπό την αξιολόγηση της απόδοσης του δικτύου σε νέα-άγνωστα δεδομένα.

### *2.3.4 Πλεονεκτήματα εφαρμογής του αλγορίθμου Τεχνητών Νευρωνικών Δικτύων*

Τα πλεονεκτήματα εφαρμογής του αλγόριθμου Τεχνητών Νερωνικών Δικτύων είναι πολλά. Ορισμένα από αυτά δίνονται παρακάτω:

1. ***Διαχείριση πολύπλοκων και μη γραμμικών σχέσεων.***

   Τα νευρωνικά δίκτυα είναι ικανά να εκπαιδευτούν σε πολύπλοκες και μη γραμμικές σχέσεις, το οποίο καθίσταται δύσκολο για άλλους αλγόριθμους.

2. ***Αυτοεκμάθηση.***

   Τα νευρωνικά δίκτυα έχουν την ικανότητα να μάθουν αυτόματα τα σημαντικά χαρακτηριστικά των δεδομένων χωρίς να είναι απαραίτητη η

επιλογή αυτών από τον χρήστη. Το χαρακτηριστικό αυτό καθιστά το μοντέλο πιο ευέλικτο και ισχυρό.

3. ***Παράλληλη επεξεργασία.***

   Μπορεί να εκτελείται παράλληλη επεξεργασία δεδομένων, κάτι που οφείλεται στην κατανεμημένη φύση των νευρώνων και τους πολυάριθμους υπολογισμούς που μπορούν να γίνουν ταυτόχρονα σε διάφορους κόμβους (νευρώνες).

4. ***Διαχείριση μεγάλων δεδομένων.***

   Τα νευρωνικά δίκτυα έχουν αποδειχθεί ιδιαιτέρως ισχυρά στην επεξεργασία μεγάλου όγκου δεδομένων, τα οποία και δίνουν τη δυνατότητα στα δίκτυα να μάθουν πολύπλοκα μοτίβα και να βελτιώσουν την ποιότητα των προβλέψεών τους.

5. ***Ευελιξία & Προσαρμοστικότητα.***

   Προσαρμόζονται σε διάφορα είδη προβλημάτων (παλινδρόμηση, ταξινόμηση, αναγνώριση μοτίβων κ.λπ.) και είναι επίσης κατάλληλα για πιο εξειδικευμένες εφαρμογές όπως η αναγνώριση ομιλίας, η ανάλυση εικόνας κ.λπ.

Όπως μπορεί να γίνει αντιληπτό και από τα παραπάνω η εφαρμογή του αλγόριθμου των Τεχνητών Νευρωνικών Δικτύων για την επίλυση του προβλήματος που πραγματεύεται η παρούσα αποτελεί μια πολύ καλή επιλογή, η οποία μπορεί να επιφέρει ακριβή και έγκυρα αποτελέσματα.

## 2.4 Τυχαία Δάση

Τα τυχαία δάση αποτελούν ένα πολύ ισχυρό αλγόριθμο μηχανικής μάθησης, ο οποίος χρησιμοποιείται βασικά για προβλήματα παλινδρόμησης και ταξινόμησης. Ο αλγόριθμος αυτός αποτελείται ουσιαστικά από πολλαπλά δένδρα απόφασης που λειτουργούν ως ένα σύνολο. Κάθε δένδρο εκπαιδεύεται σε ένα διαφορετικό, τυχαίο υποσύνολο όλων των δεδομένων ώστε να παράγει μια πρόβλεψη. Η τελική πρόβλεψη του αλγόριθμου προκύπτει από τον μέσο όρο των παραπάνω αποτελεσμάτων. Με τον τρόπο αυτό μειώνεται ο κίνδυνος της υπερπροσαρμογής (overfitting) και αυξάνεται σημαντικά η ακρίβεια και η ισχύς του μοντέλου.

Παρακάτω αναλύεται σε θεωρητικό επίπεδο η δομή και ο τρόπος λειτουργίας τόσο του αλγόριθμου Δένδρα Απόφασης (Decision Trees) όσο και αυτού των Τυχαίων Δασών (Random Forests).

### 2.4.1 Δομή & Λειτουργία Δένδρων Απόφασης (Decision Trees)

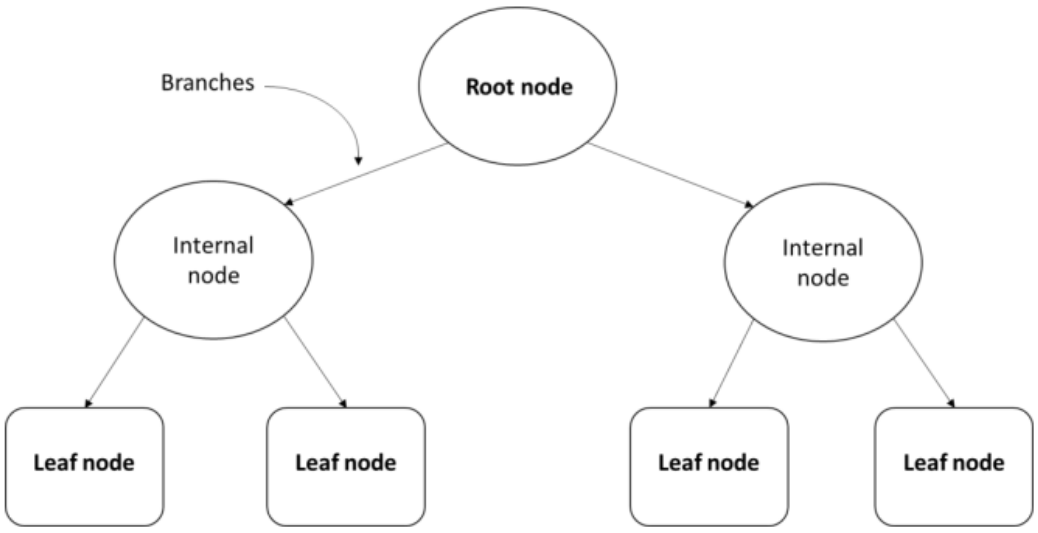
Όπως αναφέρθηκε και προηγουμένως, ο αλγόριθμος των Τυχαίων Δασών συνδυάζει πολλά δένδρα απόφασης με σκοπό την εξαγωγή μιας όσο το δυνατόν ακριβέστερης πρόβλεψης. Για το λόγο αυτό στο συγκεκριμένο σημείο κρίνεται χρήσιμη μια συνοπτική ανασκόπηση του τρόπου λειτουργίας της θεμελιώδους αυτής μονάδας για τον αλγόριθμο που εξετάζουμε, τα δένδρα απόφασης (Decision Trees).

Ειδικότερα, τα δένδρα απόφασης χωρίζονται σε 2 κατηγορίες. Έτσι έχουμε τα δένδρα ταξινόμησης (όταν η μεταβλητή-στόχος είναι κατηγορική) και τα δένδρα παλινδρόμησης (όταν η μεταβλητή-στόχος είναι συνεχής τιμή), όπως στην περίπτωσή μας. Η κεντρική λογική της εκπαίδευσης με τη χρήση του αλγόριθμου δένδρα παλινδρόμησης είναι η διαδοχική διάσπαση, βάσει κάποιου χαρακτηριστικού, των συνολικών παρατηρήσεων σε υποσύνολα, με σκοπό να ελαχιστοποιηθεί η διασπορά της μεταβλητής στόχου σε κάθε υποσύνολο. Η διαδικασία αυτή αναπαρίσταται ως ένα ανεστραμμένο δένδρο το οποίο περιλαμβάνει τη ρίζα, κόμβους, διακλαδώσεις αλλά και φύλλα. Η ρίζα είναι ο πρώτος κόμβος του «δένδρου» μας, έχει μόνο εξερχόμενες διακλαδώσεις και εκεί βρίσκεται συγκεντρωμένο το σύνολο των δεδομένων εκπαίδευσης στη συνέχει γίνεται διάσπαση με βάση κάποιο χαρακτηριστικό, ακολουθείται η αντίστοιχη διακλάδωση και το υποσύνολο περνάει στον επόμενο κόμβο. Συνήθως ως κριτήριο διαχωρισμού επιλέγεται το Mean Squared Error (MSE) ή το Mean Absolute Error (MAE), που υπολογίζουν το πόσο καλά ο διαχωρισμός μειώνει το συνολικό σφάλμα της πρόβλεψης. Στη συνέχεια επιλέγεται εκ νέου κάποιο χαρακτηριστικό με σκοπό τη διάσπαση του υποσύνολου σε 2 νέα έτσι ώστε αυτά μέσω των διακλαδώσεων να περάσουν στον επόμενο κόμβο. Αυτή η διαδικασία επαναλαμβάνεται με στόχο τη δημιουργία το δυνατόν ομοιογενέστερων ομάδων με τιμές-στόχους πολύ κοντινές μεταξύ τους, μειώνοντας έτσι το συνολικό σφάλμα. Το δένδρο συνεχίζει την παραπάνω διαδικασία μέχρι να επιτευχθεί κάποιος τερματικός όρος π.χ. μέγιστος αριθμός κόμβων, διαθέσιμος αριθμός δειγμάτων, ελάχιστος αριθμός δειγμάτων σε ένα φύλλο κ.λπ. Όταν γίνει αυτό, δημιουργούνται τα φύλλα του δένδρου,

τα οποία περιέχουν την προβλεπόμενη τιμή. Αυτή συνήθως προκύπτει από το μέσο όρο των αριθμητικών τιμών των δεδομένων που έφτασαν στον κόμβο.

Αφού εκπαιδευτεί το μοντέλο με τον παραπάνω τρόπο και όταν χρειαστεί να πραγματοποιηθεί μια νέα πρόβλεψη το δένδρο λαμβάνει μια νέα είσοδο η οποία ακολουθεί όλη τη διαδρομή από τη ρίζα προς τα φύλλα περνώντας από τους κόμβους του δένδρου. Τελικά η πρόβλεψη της νέας εισόδου είναι η τιμή του φύλλου. Στο παρακάτω *Σχήμα 2.6* βλέπουμε γραφικά τη δομή του αλγόριθμου.

Ένα πρόβλημα που μπορεί να προκύψει όταν επιλέγεται η εφαρμογή του αλγορίθμου Δένδρο Απόφασης είναι ότι στην περίπτωση που αυτό γίνει πολύ περίπλοκο και προσαρμοστεί υπερβολικά στα δεδομένα εκπαίδευσης, μπορεί να προκαλέσει overfitting και δυσκολία στο μοντέλο να προσαρμοστεί και να προβλέψει νέα δεδομένα. Για να αποφευχθεί το πρόβλημα αυτό, συχνά εφαρμόζεται η τεχνική του «κλαδέματος» (pruning), δηλαδή ελάττωση του βάθους του δένδρου, που μπορεί να γίνει πριν ή μετά την εκπαίδευση. Ωστόσο, και για την αποφυγή αυτού του προβήματος, συνήθη επιλογή φαίνεται να αποτελεί πλέον και ο αλγόριθμος Random Forest που οποία έχει τη βάση του στον αλγόριθμο Decision Tree καθώς δημιουργεί πολλά δένδρα απόφασης, συνδυάζοντας τις προβλέψεις τους για πιο ακριβή αποτελέσματα.




***Σχήμα 2.6 Τυπική Δομή του αλγόριθμου Δένδρα Απόφασης. (πηγή: Ιδία επεξεργασία)***

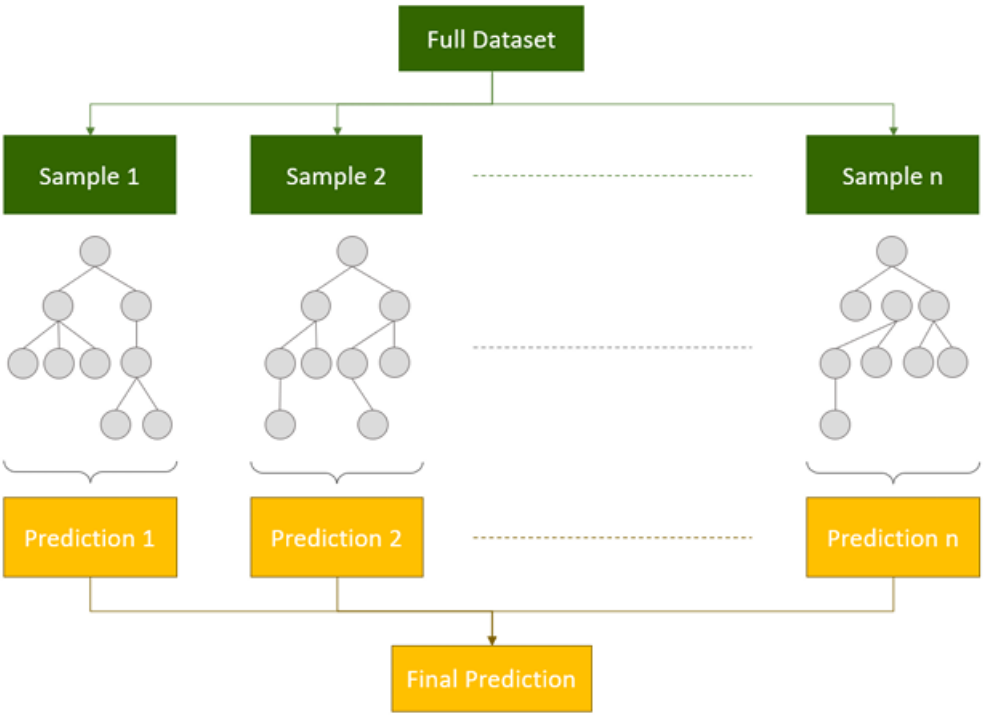
### 2.4.2 Τυχαία Δάση (Random Forests)

Ο αλγόριθμος Τυχαία Δάση είναι μια ισχυρή μέθοδος μηχανικής μάθησης που, όπως αναφέρθηκε και προηγουμένως, βασίζεται σε σύνολα δένδρων απόφασης και για το λόγο αυτό δίνει αποτελέσματα μεγάλης ακρίβειας.

Πιο συγκεκριμένα ο αλγόριθμος Random Forest δημιουργεί πολλά δένδρα απόφασης, καθένα από τα οποία εκπαιδεύεται βάσει ενός τυχαίου υποσύνολου του αρχικού συνόλου δεδομένων εκπαίδευσης (bagging). Κάθε δένδρο μαθαίνει διαφορετικά μοτίβα, δίνοντας σε κάθε κόμβο έμφαση σε διαφορετικές ανεξάρτητες μεταβλητές έτσι ώστε να αποφευχθεί η υπερβολική εξάρτηση από συγκεκριμένα μόνο χαρακτηριστικά. Τελικά κάθε δένδρο δίνει μια διαφορετική πρόβλεψη και στη συνέχεια ο αλγόριθμος Random Forest, αφού τις συνδυάσει, εξάγει το τελικό αποτέλεσμα, το οποίο στην περίπτωση προβλήματος παλινδρόμησης είναι ο Μέσος Όρος των προβλέψεων που προέκυψαν από το εκάστοτε δένδρο.

Ο αλγόριθμος Τυχαία Δάση σταματά την εκπαίδευση όταν έχουν δημιουργηθεί όλα τα δένδρα τα οποία τον αποτελούν (π.χ. 100) και όταν αυτά έχουν ολοκληρώσει την διαδικασία εκπαίδευσής τους, όπως αυτή αναλύθηκε στην ενότητα 2.4.1. Στο παρακάτω Σχήμα 2.7 δίνεται σχηματικά η δομή του αλγόριθμου.

Αφού ολοκληρωθεί η εκπαίδευση και με σκοπό ο αλγόριθμος να πραγματοποιήσει προβλέψεις σε νέα δεδομένα, τότε αυτά περνάνε από όλα τα δένδρα και κάθε δένδρο αποφασίζει με βάση τους διαχωρισμούς που έμαθε κατά την εκπαίδευση. Κάθε δένδρο δίνει μια ξεχωριστή πρόβλεψη και το τελικό αποτέλεσμα προκύπτει από το μέσο όρο αυτών.




*Σχήμα 2.7 Δομή τυπικού αλγόριθμου Τυχαία Δάση (πηγή: Σερβετάς Γεώργιος, Μεταπτυχιακή Διατριβή «Μηχανική Μάθηση στην Πρόβλεψη της τιμής ενοικίασης Airbnb στο Άμστερνταμ»)*

### 2.4.3 Πλεονεκτήματα Αλγόριθμου Τυχαία Δάση (Random Forest)

Τα σημαντικότερα πλεονεκτήματα εφαρμογής του συγκεκριμένου αλγόριθμου είναι τα εξής:

1. ***Μεγάλη ακρίβεια του παραγόμενου μοντέλου.***
   Τα τυχαία δάση δίνουν μεγάλη ακρίβεια στο μοντέλο αφού έχουμε το συνδυασμό πολλών δένδρων αποφάσεων.
2. ***Ανθεκτικότητα στην υπερπροσαρμογή.***
   Ο αλγόριθμος αυτός είναι πολύ λιγότερο ευάλωτος στην υπερπροσαρμογή σε σύγκριση με ένα απλό Δένδρο απόφασης και αυτό οφείλεται στην τεχνική bagging (που αναλύθηκε προηγουμένως), στην εξέταση τυχαίων υποσυνόλων των δεδομένων και όχι του συνόλου αυτών και στο γεγονός ότι το τελικό αποτέλεσμα προκύπτει από το συνδυασμό όλων των παραγόμενων δένδρων απόφασης.
3. ***Αποδοτική χρήση σε προβλήματα παλινδρόμησης και ταξινόμησης.***
   Ο αλγόριθμος μπορεί να λειτουργήσει εξίσου αποτελεσματικά τόσο σε προβλήματα παλινδρόμησης όσο και σε προβλήματα ταξινόμησης.

4. **Αποδοτική διαχείριση μεγάλου όγκου δεδομένων.**

   Η ικανότητα διαχείρισης μεγάλου όγκου δεδομένων από τον συγκεκριμένο αλγόριθμο οφείλεται κυρίως στο ότι αποτελείται από πολλά ανεξάρτητα δένδρα που μπορούν να εκπαιδευτούν ταυτόχρονα σε πολλούς υπολογιστικούς πυρήνες αλλά και στην τυχαία επιλογή δειγμάτων με τα οποία εκπαιδεύεται κάθε δένδρο με αποτέλεσμα την βελτίωση της ταχύτητας εκπαίδευσης του μοντέλου.

## 2.5 Λογισμικά

Πολλά είναι τα λογισμικά και τα εργαλεία τα οποία είναι κατάλληλα για την εφαρμογή τεχνικών Μηχανικής Μάθησης. Ορισμένα από αυτά ξεχωρίζουν για την φιλικότητά τους προς το χρήστη ενώ άλλα γιατί προσφέρουν ισχυρά χαρακτηριστικά για προχωρημένους χρήστες. Η επιλογή του κατάλληλου λογισμικού ανάμεσα από τα περίπου 200 που υπάρχουν διαθέσιμα στην αγορά εξαρτάται από τις ανάγκες του προς επίλυση προβλήματος, το επίπεδο εξειδίκευσης του εκάστοτε χρήστη και από το είδος των διαθέσιμων δεδομένων.

Μεταξύ των διαθέσιμων λογισμικών ορισμένα από τα πιο δημοφιλή στις εφαρμογές Μηχανικής Μάθησης είναι τα εξής: Scikit-learn, PyTorch, Tensorflow, Weka, KNIME, Colab, Apache Mahout, Accord.Net, Shogun, Keras.io, Rapid Miner, Google Cloud ML Engine, Amazon Machine Learning, NET, Oryx2, Google ML kit for cell, Big ML, OpenNN, Vertex AI, XGBOOST κ.α.

Για την επίλυση του προβλήματος που πραγματεύεται η παρούσα και τη δημιουργία ενός μοντέλου Μηχανικής Μάθησης το οποίο θα εκτιμά την τιμή ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης επιλέχθηκε το λογισμικό Rapid Miner, για το οποίο θα αναφερθούμε σε θεωρητικό επίπεδο στο παρακάτω *Κεφάλαιο 3*.

# 3. Rapid Miner

Στην ενότητα αυτή θα παρουσιαστεί σε θεωρητικό επίπεδο το λογισμικό Rapid Miner, η μεθοδολογία δημιουργίας μοντέλων μηχανικής μάθησης με τη βοήθεια αυτού καθώς και τα πλεονεκτήματα επιλογής του για εφαρμογές Μηχανικής Μάθησης.

## 3.1 Γενικά

Το Rapid Miner αποτελεί πλέον μια από τις πιο δημοφιλής πλατφόρμες ανοικτού κώδικα, ανάλυσης, εξόρυξης και διαχείρισης δεδομένων, δημιουργίας προβλέψεων και ανάπτυξης μοντέλων μηχανικής μάθησης. Αποτελεί ένα ευέλικτο και ισχυρό εργαλείο, ιδανικό τόσο για αρχάριους όσο και για προχωρημένους χρήστες και διαθέτει ένα πολύ φιλικό προς τον χρήστη περιβάλλον χωρίς την απαίτηση γραφής κώδικα.

Τα δεδομένα που μπορούν να εισαχθούν στο λογισμικό Rapid Miner δύναται να προέρχονται από πολλές διαφορετικές πηγές και βάσεις δεδομένων. Αυτά μπορεί να είναι:

- ***Δομημένα Δεδομένα****:* Δηλαδή δεδομένα που οργανώνονται σε στήλες και γραμμές, τα οποία συνήθως βρίσκουμε σε βάσεις δεδομένων και υπολογιστικά φύλλα. (π.χ. Excel, CSV, MySQL, PostgreSQL, Oracle, SQL Server, Google Sheets). Σε αυτή την κατηγορία ανήκουν και τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα και αντλήθηκαν από τη βάση δεδομένων της ιστοσελίδας Inside AIRBNB σε μορφή Excel.
- ***Μη Δομημένα Δεδομένα:*** Δηλαδή δεδομένα που δεν έχουν σαφή και προκαθορισμένη μορφή (π.χ. Text Data, Web Data)
- ***Ημιδομημένα Δεδομένα:*** Δηλαδή δομημένα δεδομένα τα οποία δεν οργανώνονται σε πίνακες αλλά περιέχουν οργανωμένα στοιχεία μέσω ειδικής σήμανσης (π.χ. XML, JSON, Logs).
- ***Εξωτερικές πηγές:*** Δηλαδή άλλες πηγές εξωτερικών δεδομένων με τις οποίες το Rapid Miner έχει τη δυνατότητα σύνδεσης (π.χ. APIs, Cloud Platforms, Big Data Platforms)

Το Rapid Miner παρέχει ισχυρά εργαλεία που βοηθούν στην προετοιμασία και επεξεργασία των παραπάνω δεδομένων, όπως το φιλτράρισμα, η κανονικοποίηση, η μετατροπή, η διαχείριση ελλιπών δεδομένων, η μείωση πολυδιαστατικότητας και πολλά άλλα μέσω του εκάστοτε τελεστή που θα χρησιμοποιηθεί κάθε φορά.

Περιλαμβάνει επίσης ένα ευρύ φάσμα αλγορίθμων παλινδρόμησης, ταξινόμησης ομαδοποίησης, ανάλυσης συσχέτισης και μείωσης διαστάσεων (όπως Neural Networks, Support Vector Machines, k-Means, Decision Trees, Random Forests και πολλούς ακόμα). Στο Rapid Miner υπάρχουν επίσης και οι μετά-τελεστές, που πρόκειται για τελεστές βελτιστοποίησης για το σχεδιασμό διαδικασιών π.χ. επαναλήψεις δεδομένων ή τεχνικές βελτιστοποίησης παραμέτρων. Υπάρχουν επίσης οι τελεστές αξιολόγησης των παραγόμενων μοντέλων με εργαλεία όπως η δημιουργία πινάκων σύγχυσης, ROC καμπυλών, και μετρικών όπως η ακρίβεια, η ανάκληση, και το F1-score. Όσον αφορά την παρουσίαση των αποτελεσμάτων, το Rapid Miner παρέχει εκτός από την μορφή των πινάκων και τη δυνατότητα οπτικοποίησης αυτών μέσω διαγραμμάτων (2D και 3D) και γραφημάτων.

Το λογισμικό Rapid Miner μπορεί να προσφέρει πολλά περισσότερα μέσω των επεκτάσεων του, οι οποίες ξεπερνούν τις 1.500 διευρύνοντας έτσι τις δυνατότητές του και καθιστώντας το ένα πανίσχυρο εργαλείο ανάλυσης δεδομένων. Οι επεκτάσεις προσφέρουν μεγάλη ευελιξία και επιτρέπουν στους χρήστες να εκμεταλλευτούν στον μέγιστο βαθμό τις δυνατότητες της πλατφόρμας για πιο εξειδικευμένες αναλύσεις. Ανάλογα με τις ανάγκες του εκάστοτε προβλήματος οι επεκτάσεις μπορεί να βοηθήσουν στη βελτίωση της απόδοσης, στην ανάλυση εξειδικευμένων δεδομένων, στην επεξεργασία κειμένου, στην εφαρμογή μεθόδων βαθιάς μηχανικής μάθησης και σε πολλά άλλα. Ορισμένες από τις πιο βασικές και χρήσιμες επεκτάσεις που προσφέρει το Rapid Miner είναι οι: Text Processing Extension (Επεξεργασία Κειμένου), Web Mining Extension (Εξόρυξη Ιστού), Time Series Extension (Χρονοσειρές), R & Python Scripting Extensions, Machine Learning Extensions (Αλγόριθμοι Μηχανικής Μάθησης), Big Data Extensions, Deep Learning Extension, Optimization Extension (Βελτιστοποίηση), Anomaly Detection Extension (Ανίχνευση Ανωμαλιών) και πολλές άλλες.

Επιπλέον το λογισμικό Rapid Miner αν και δεν απαιτεί τη γνώση ή την ανάπτυξη κάποιου κώδικα, επιτρέπει την ενοποίηση με γλώσσες προγραμματισμού όπως π.χ. Python και R, δίνοντας με τον τρόπο αυτό τη δυνατότητα σε πιο έμπειρους χρήστες να επεκτείνουν τη λειτουργικότητά του. Το Rapid Miner υποστηρίζει επίσης τη δυνατότητα AutoML [Auto Machine Learning (Αυτόματη Μηχανική Μάθηση)] που

βοηθάει τους χρήστες με την αυτόματη επιλογή του καλύτερου αλγόριθμου για το εκάστοτε σύνολο δεδομένων και πρόβλημα.

## 3.2 Πλεονεκτήματα & Μειονεκτήματα

Το RapidMiner είναι ένα ισχυρό, ευέλικτο και εύχρηστο εργαλείο που προσφέρει πολλά πλεονεκτήματα για όσους θέλουν να ασχοληθούν με τη μηχανική μάθηση και την ανάλυση δεδομένων. Ορισμένα από αυτά είναι:

- *Ευκολία στη χρήση.*
- *Ευχάριστο γραφικό περιβάλλον (GUI).*
- *Ευρεία υποστήριξη δεδομένων.*
- *Ισχυρή προεπεξεργασία δεδομένων.*
- *Πλούσια Βιβλιοθήκη Αλγορίθμων.*
- *Δυνατότητα επεξεργασίας μεγάλων δεδομένων.*
- *Επέκταση μέσω Plugins και ενσωμάτωση.*
- *Αυτοματοποιημένη μηχανική μάθηση (AutoML).*
- *Υποστήριξη για Ενσωμάτωση σε Επιχειρηματικά Περιβάλλοντα.*
- *Αξιοπιστία και Απόδοση.*

Εκτός από τα πολυάριθμα πλεονεκτήματα που διαθέτει, το Rapid Miner όπως κάθε λογισμικό έχει ορισμένους περιορισμούς και μειονεκτήματα τα οποία θα πρέπει να ληφθούν υπόψιν κατά την επιλογή ανάλογα με τις ανάγκες του προβλήματος και του χρήστη. Ορισμένα από αυτά είναι τα εξής:

- *Απαιτήσεις απόδοσης σε πολύ μεγάλα δεδομένα.*
- *Απαιτήσεις Μνήμης RAM.*
- *Περιορισμένη Υποστήριξη για Κάποιες Γλώσσες Προγραμματισμού.*
- *Μικρότερη Ελευθερία Εξατομίκευσης Αλγορίθμων.*

Το RapidMiner συνεπώς είναι ένα πολύτιμο εργαλείο για την ανάλυση δεδομένων και τη μηχανική μάθηση, ειδικά για χρήστες που θέλουν μια φιλική και εύχρηστη πλατφόρμα. Ωστόσο οι περιορισμοί του, κυρίως όσον αφορά την επεξεργασία μεγάλων δεδομένων και τις ανάγκες πόρων, μπορεί να το καθιστούν λιγότερο αποτελεσματικό σε ορισμένα περιβάλλοντα ή για πολύ απαιτητικές αναλύσεις.

Λαμβάνοντας λοιπόν υπόψιν τα παραπάνω και με βάση τις υπολογιστικές ανάγκες και τα δεδομένα της παρούσης, το Rapid Miner αποτελεί μια πολύ καλή επιλογή λογισμικού για τη δημιουργία του μοντέλου εκτίμησης της τιμής ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης με μεθόδους μηχανικής μάθησης.

## 3.3 Μεθοδολογία Rapid Miner

Σε αυτή την υπό-ενότητα θα αναλυθούν σε θεωρητικό επίπεδο όλα τα βήματα και οι διαδικασίες που θα πρέπει να ακολουθηθούν στο περιβάλλον του Rapid Miner με σκοπό τη δημιουργία του μοντέλου που θα εκτιμά την τιμή ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης.

Πριν την έναρξη δημιουργίας του επιθυμητού μοντέλου θα πρέπει να προηγηθεί μια αρχική επεξεργασία των δεδομένων που πρόκειται να χρησιμοποιηθούν. Έτσι από το σύνολο των δεδομένων θα πρέπει να ελεγχθούν και να αφαιρεθούν ή να αντικατασταθούν τυχόν ελλιπή δεδομένα και πιθανές ακραίες τιμές οι οποίες μπορεί να οφείλονται ακόμα και σε λάθος κατά την εισαγωγή των δεδομένων. Τέλος, τα δεδομένα θα πρέπει να διαχωριστούν σε υποσύνολα εκπαίδευσης και δοκιμής.

Το Rapid Miner μας επιτρέπει αυτή τη δυνατότητα πρώιμης επεξεργασίας του συνόλου δεδομένων μας μέσω τελεστών όπως οι “Replace Missing Values”, “Detect Outlier”, “Filter Examples” και “Split Data”. Ας δούμε λίγο πιο αναλυτικά όμως πως λειτουργεί ο κάθε ένας από τους τελεστές αυτούς ξεχωριστά.

- ***“Replace Missing Values”***

  Ο τελεστής αυτός εντοπίζει στα δεδομένα ελλιπής τιμές και τις αντικαθιστά με την τιμή που εμείς επιλέγουμε. Συνήθως επιλέγεται ο μέσος όρων των τιμών της εκάστοτε μεταβλητής.

- ***“Detect Outlier (LOF)”***

  Ο τελεστής αυτός είναι ένας από τους πιο αξιόπιστους τρόπους εντοπισμού ακραίων τιμών, ιδίως όταν τα δεδομένα μας είναι πολυδιάστατα και δεν ακολουθούν την κανονική κατανομή, όπως στην προκειμένη περίπτωση. Συγκεκριμένα, εξετάζει το αν μια παρατήρηση βρίσκεται σε περιοχή με πολύ μικρή πυκνότητα σε σχέση με τους κοντινούς της με αποτέλεσμα αυτή να είναι πιθανό outliner. Ο χρήστης θα πρέπει να ορίσει το πόσους κοντινούς γείτονες θα εξετάσει ο τελεστής με αυτή την τιμή συνήθως να κυμαίνεται από 10 έως 20

γείτονες. Το αποτέλεσμα δίνει ακόμα μία στήλη συνεχών αριθμών οι οποίοι προδίδουν την πιθανότητα μια μεταβλητή να είναι ακραία τιμή. Έτσι στη συνέχεια μπορεί εύκολα να πραγματοποιηθεί σχετικός έλεγχος, να αφαιρεθούν τυχόν λάθος καταχωρήσεις και να αποφασιστεί εάν θα διατηρήσουμε τυχόν παρατηρήσεις που αποτελούν «ειδικές περιπτώσεις» που όντως υπάρχουν.

- ***"Filter Examples"***

  Με τη βοήθεια του τελεστή αυτού μπορούμε να φιλτράρουμε και να κρατήσουμε συγκεκριμένες μόνο παρατηρήσεις από ένα σύνολο δεδομένων που πληρούν μια η περισσότερες συνθήκες. Παραδείγματος χάριν και στην περίπτωση των ακραίων τιμών στις οποίες αναφερθήκαμε και προηγουμένως, μπορούμε με τη βοήθεια του "Filter Examples" και επιλέγοντας το "attribute_value_filter", στο πεδίο "parameter string" να γράψουμε τη συνθήκη που θέλουμε να ικανοποιείται, δηλαδή το LOF outlier score να είναι μεγαλύτερο από μια συγκεκριμένη τιμή.

- ***"Split Data"***

  Ο τελεστής αυτός χωρίζει όλο το σετ δεδομένων σε υποσύνολα όπου το καθένα μπορεί στη συνέχεια να χρησιμοποιηθεί ξεχωριστά. Συγκεκριμένα μπορούμε εμείς μέσω της παραμέτρου partitions που μας προσφέρει ο τελεστής να ορίσουμε το ποσοστό του συνόλου των δεδομένων που θα έχει κάθε υποσύνολο. Στην δική μας εφαρμογή μπορεί να μας βοηθήσει να χωρίσουμε τα δεδομένα σε δύο υποσύνολα, εκπαίδευσης και τεστ, σε ποσοστά 70% και 30% αντίστοιχα, όπως συνηθίζεται και στις περισσότερες εφαρμογές αυτού του είδους.

Στη συνεχεία και αφού ολοκληρωθεί αυτός ο πρώτος «καθαρισμός» και διαχωρισμός του συνόλου των δεδομένων, μπορούμε να προχωρήσουμε στην ανάπτυξη του μοντέλου που θέλουμε να δημιουργήσουμε. Ακολουθεί η διαδικασία αυτή σε θεωρητικό επίπεδο.

1. ***Εισαγωγή Δεδομένων Εκπαίδευσης & Δοκιμής.***

   Αρχικά θα πρέπει να εισαχθούν τα δεδομένα εκπαίδευσης και δοκιμής του προς δημιουργία μοντέλου μηχανικής μάθησης. Τα δεδομένα αυτά περιλαμβάνουν όλες τις ανεξάρτητες αλλά και την εξαρτημένη μεταβλητή και μπορούν να εισαχθούν σε πολλές διαφορετικές μορφές αρχείων, όπως έχουμε ήδη αναφέρει

και σε προηγούμενο υπό-κεφάλαιο. Σε αυτό το σημείο ορίζεται και το είδος κάθε μεταβλητής που εισάγεται στο λογισμικό. Ορίζουμε δηλαδή το αν πρόκειται για ονομαστική μεταβλητή, ακέραια κ.λπ.

2. ***Προεπεξεργασία Δεδομένων.***

   Εδώ, στην περίπτωσή μας, περιλαμβάνονται η κωδικοποίηση των ονομαστικών μεταβλητών σε αριθμητικές (π.χ. η γειτονιά του ακινήτου) μέσω του τελεστή "nominal to numerical" και η κανονικοποίηση των αριθμητικών μεταβλητών μέσω του τελεστή "Normalize" με σκοπό αυτές να έρθουν σε κοινή κλίμακα, χωρίς ταυτόχρονα να αλλοιωθεί η μεταξύ τους διαφορά και έτσι ο αλγόριθμος να καταφέρει να τις «αντιληφθεί» και επεξεργαστεί σωστά δίνοντας μας ένα έγκυρο μοντέλο πρόβλεψης. Η κανονικοποίηση έχει νόημα να εφαρμόζεται κυρίως σε αλγόριθμους οι οποίοι είναι ευαίσθητοι στις απόλυτες τιμές των χαρακτηριστικών, όπως είναι τα Νευρωνικά Δίκτυα και αυτό γιατί εκπαιδεύονται μέσω της σταδιακής προσαρμογής βαρών με backpropagation. Συνεπώς εάν τα χαρακτηριστικά είναι σε διαφορετικές κλίμακες, οι παράγωγοι εκτιμώνται κακώς και το δίκτυο μπορεί να μαθαίνει πολύ αργά ή να αποτυγχάνει. Από την άλλη σε αλγόριθμους όπως τα Τυχαία Δάση, ο οποίος βασίζεται σε δενδρική διάσπαση και όχι σε αποστάσεις ή βαθμίδες, η κανονικοποίηση των δεδομένων δεν έχει νόημα να εφαρμόζεται.

3. ***Ορισμός ρόλων.***

   Στη συνέχεια και με τη βοήθεια του τελεστή "Set Role" ορίζεται η στήλη των δεδομένων που αποτελεί τη μεταβλητή «στόχο» (τιμή ενοικίασης) και την οποία προσπαθούμε να προβλέψουμε με την ανάπτυξη του μοντέλου.

4. ***Επιλογή Αλγορίθμου.***

   Ακολουθεί η επιλογή του κατάλληλου για το εκάστοτε πρόβλημα αλγόριθμου μηχανικής μάθησης και η ρύθμιση των απαιτούμενων παραμέτρων με σκοπό τη βέλτιστη λειτουργία του για τα δεδομένα του προβλήματος.

   Στην περίπτωση επιλογής του αλγόριθμου των **Νευρωνικών Δικτύων**, οι παράμετροι που πρέπει να ορισθούν από το χρήστη είναι οι εξής:

   - <u>Training Cycles (Κύκλοι Εκπαίδευσης)</u>

     Η συγκεκριμένη παράμετρος αφορά τον αριθμό των επαναλήψεων που το δίκτυο θα περάσει από τα δεδομένα εκπαίδευσης κατά τη διαδικασία της εκμάθησης. Σε κάθε μια επανάληψη (κύκλο) πραγματοποιείται η

ενημέρωση των βαρών του δικτύου βάσει των λαθών που εντοπίζονται κάθε φορά. Η τιμή της συγκεκριμένης παραμέτρου μπορεί να κυμαίνεται από 100 έως 1.000 ή και περισσότερους κύκλους και εξαρτάται πάντα από την πολυπλοκότητα των δεδομένων και το είδος του προβλήματος. Πολύ υψηλές τιμές συνεπάγονται καλύτερη απόδοση καθώς υπάρχει καλύτερη προσαρμογή στα δεδομένα εκπαίδευσης, ωστόσο αυτό μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting), στην οποία έχουμε αναφερθεί και σε προηγούμενο κεφάλαιο, όπου το μοντέλο προσαρμόζεται πολύ καλά στα δεδομένα εκπαίδευσης με αποτέλεσμα την αδυναμία καλής γενίκευσης σε άλλα, νέα δεδομένα. Από την άλλη μεριά ένας αρκετά μικρός αριθμός κύκλων μπορεί να μην επαρκεί για την εκμάθηση των μοτίβων των δεδομένων με αποτέλεσμα την υποπροσαρμογή (underfitting) του. Γενικά μια καλή και αποδεκτή τιμή για την παράμετρο αυτή θεωρείτε το 200.

- Learning Rate (Ρυθμός Μάθησης)

  Η παράμετρος αυτή αφορά το ρυθμό της αλλαγής των βαρών του δικτύου σε κάθε αναπροσαρμογή τους, δηλαδή σε κάθε κύκλο εκπαίδευσης. Η τιμή του κυμαίνεται μεταξύ 0,001 και 0,1 με την πιο συνηθισμένη τιμή να είναι 0,01 ενώ και πάλι η βέλτιστη τιμή εξαρτάται από τη φύση του προβλήματος. Χαμηλές τιμές συνεπάγονται βέλτιστο αποτέλεσμα όμως έχουν σαν επίπτωση τον υπερβολικά μεγάλο χρόνο εκπαίδευσης του δικτύου. Υψηλές τιμές συνεπάγονται γρηγορότερη μάθηση αλλά και την πιθανότητα παράλειψης σημαντικών μοτίβων με αποτέλεσμα την αστάθεια ή αναποτελεσματική εκπαίδευση του δικτύου.

- Momentum (Ορμή)

  Αυτή η παράμετρος λειτουργεί όπως η φυσική έννοια της ορμής από τη μηχανική. Με λίγα λόγια αυτή είναι που επιτρέπει στο μοντέλο να προχωρά την εκπαίδευσή του προς τη σωστή κατεύθυνση ακόμα και όταν στα δεδομένα ή στα βάρη παρουσιάζονται αλλαγές που προκαλούν ταλαντεύσεις και έτσι το βοηθά στο να μην παγιδευτεί σε τοπικά ελάχιστα. Ουσιαστικά η παράμετρος αυτή χρησιμοποιώντας τις προηγούμενες αλλαγές στα βάρη, βοηθάει στο δίκτυο να διατηρήσει μια

ομαλή κατεύθυνση και αποτρέπει τις πολύ μεγάλες αλλαγές. Συνήθως η τιμή του κυμαίνεται από 0,5 έως 0,9 με την τελευταία να αποτελεί την πιο βέλτιστη τιμή. Χαμηλή τιμή της παραμέτρου συνεπάγεται αργή μάθηση του μοντέλου με πολλές ταλαντεύσεις ενώ αντιθέτως μια υψηλότερη τιμή σημαίνει γρηγορότερη μάθηση και λιγότερες ταλαντεύσεις. Από την άλλη μια πολύ υψηλή τιμή της «ορμής» (π.χ. από 1 και πάνω) μπορεί να κάνει το μοντέλο πολύ «απότομο» με αποτέλεσμα την μη σταθεροποίησή του στη βέλτιστη λύση.

- Error Epsilon (Σφάλμα Epsilon)

  Η παράμετρος Σφάλμα Έψιλον καθορίζει την τιμή ανοχής του σφάλματος κατά την εκπαίδευση. Είναι ουσιαστικά η τιμή κάτω από την οποία θα πρέπει να πέσει το μέσο σφάλμα έτσι ώστε η εκπαίδευση να τερματιστεί. Συνήθως η τιμή της παραμέτρου κυμαίνεται από 0,00001 έως 0,001 αναλόγως την ακρίβεια που απαιτείται στην εκάστοτε εφαρμογή. Και πάλι μια υψηλή τιμή της παραμέτρους "Error Epsilon" ενδέχεται να οδηγήσει σε πρόωρη διακοπή της εκπαίδευσης με αποτέλεσμα το μοντέλο να μην έχει μάθει καλά τα δεδομένα ενώ μια χαμηλή τιμή αυτού οδηγεί στην υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης με ό,τι αυτό συνεπάγεται όπως αναλύθηκε προηγουμένως.

Όσον αφορά τώρα τις παραμέτρους του αλγόριθμου **Τυχαία Δάση** που εφαρμόστηκε πρέπει να ορισθούν τα εξής:

- Number of trees (Αριθμός Δένδρων)

  Αυτή η παράμετρος καθορίζει τον αριθμό των δένδρων που θα δημιουργήσει το μοντέλο. Συνήθως το εύρος των δένδρων κυμαίνεται από 50 έως 500. Μεγάλος αριθμός δένδρων συνεπάγεται καλύτερη ακρίβεια και γενίκευση του μοντέλου (μείωση του overfitting) όμως αυξάνεται σημαντικά ο χρόνος εκπαίδευσης καθώς και η απαιτούμενη μνήμη. Από την άλλη μικρός αριθμός δένδρων συνεπάγεται γρήγορη εκπαίδευση που μπορεί να οδηγήσει σε χαμηλότερη ακρίβεια. Μια τυπικά καλή τιμή την οποία μπορεί να πάρει η παράμετρος αυτή είναι το 100.

- Criterion (Κριτήριο διαχωρισμού)

  Η συγκεκριμένη παράμετρος καθορίζει το κριτήριο διαχωρισμού που θα χρησιμοποιήσει ο αλγόριθμος για να διαχωρίσει τα δεδομένα σε κάθε κόμβο του δένδρου. Υπάρχει μια ποικιλία επιλογών, καθεμία από τις οποίες είναι καταλληλότερη ανάλογα με το είδος του προβλήματος, τη φύση των μεταβλητών και τον τύπο της μεταβλητής που επιδιώκουμε να προβλέψουμε. Για προβλήματα παλινδρόμησης η παράμετρος αυτή θα πρέπει να είναι η squared error η οποία μετράει το σφάλμα μεταξύ των προβλέψεων και των πραγματικών τιμών.

- Maximal Depth (Μέγιστο βάθος δένδρου)

  Η παράμετρος αυτή καθορίζει το μέγιστο βάθος που μπορεί να έχει το κάθε δένδρο απόφασης μέσα στο Random Forest. Ορίζει, ουσιαστικά, το πόσα επίπεδα κόμβων μπορεί να έχει ένα δένδρο πριν σταματήσει να διακλαδίζεται. Οι τιμές που μπορεί να πάρει η συγκεκριμένη παράμετρος είναι απεριόριστες, ωστόσο συνήθως επιλέγονται τιμές μεταξύ 10 και 50. Μεγάλο βάθος συνεπάγεται ισχυρά μοντέλα που μπορούν να απομνημονεύσουν τα δεδομένα, είναι κατάλληλο για μεγάλα dataset και χρησιμοποιείται συχνά σε πολύπλοκα προβλήματα ταξινόμησης. Ωστόσο χρειάζεται προσοχή αφού μπορεί να μάθει πολύ καλά τα δεδομένα εκπαίδευσης με κίνδυνο την υπερπροσαρμογή του μοντέλου. Ένα μικρό βάθος δένδρων αποκλείει την υπερπροσαρμογή του μοντέλου και το κάνει πιο γενικευμένο. Είναι καλό για απλά προβλήματα με λίγα δεδομένα, ωστόσο περιορίζει την ικανότητα των δένδρων να προσαρμοστούν καλά στα δεδομένα εκπαίδευσης. Ένα μέτριο βάθος (10-30 επίπεδα) εξισορροπεί την ακρίβεια και την γενίκευση και αποτελεί μια καλή επιλογή για τα περισσότερα σετ δεδομένων.

5. ***Αξιολόγηση Απόδοσης Μοντέλου.***

   Ακολουθεί ο υπολογισμός της απόδοσης του παραγόμενου μοντέλου με τη βοήθεια του τελεστή “Performance” και συγκεκριμένα του “Performance (Regression)” μιας και το πρόβλημα της παρούσας αφορά πρόβλημα παλινδρόμησης. Ο συγκεκριμένος τελεστής επιτρέπει την αξιολόγηση του παραγόμενου μοντέλου μέσω διάφορων μέτρων. Τα μέτρα αξιολόγησης όμως

των οποίων τη σημασία θα εξηγήσουμε περαιτέρω καθώς είναι αυτά που θα αναλύσουμε είναι το τετράγωνο συντελεστή συσχέτισης (Squared correlation) – R squared ($R^2$) και η Ρίζα Μέσης Τετραγωνικής Απόδοσης - Root Mean Squared Error (RMSE).

- $R^2$ – R squared

  Ο συντελεστής αυτός αποτελεί ένα από τα βασικότερα στατιστικά μέτρα για την αξιολόγηση ενός μοντέλου πρόβλεψης. Συγκεκριμένα, εκφράζει το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που μπορεί να εξηγήσει το μοντέλο από τις ανεξάρτητες μεταβλητές. Τιμή κοντά στη μονάδα (1) υποδεικνύει ότι το μοντέλο εξηγεί σε μεγάλο βαθμό τις μεταβολές της μεταβλητής - στόχο και είναι ιδιαίτερα αποδοτικό. Αντίθετα χαμηλές τιμές υποδεικνύουν ότι το μοντέλο δεν μπορεί να εξηγήσει επαρκώς τη σχέση μεταξύ των μεταβλητών κάτι που μπορεί να σημαίνει ελλιπή δεδομένα εισόδου είτε ακατάλληλη μορφή μοντελοποίησης. Στο σημείο αυτό θα πρέπει να επισημανθεί ότι ο συγκεκριμένος συντελεστής εκφράζει την ικανότητα εξήγησης της διακύμανσης και όχι την απόλυτη ακρίβεια των προβλέψεων, για την οποία υπολογίζονται άλλα μέτρα όπως η «Ρίζα μέσης τετραγωνικής απόδοσης» (Root Mean Squared Error – RMSE).

- Root Mean Squared Error (RMSE)

  Ο συντελεστής αυτός μετράει τη μέση απόκλιση των προβλέψεων του μοντέλου από τις πραγματικές τιμές, υπολογίζεται ως η τετραγωνική ρίζα της μέσης τιμής των τετραγώνων των σφαλμάτων και εκφράζεται στις ίδιες μονάδες με την εξαρτημένη μεταβλητή. Μικρότερη τιμή του συντελεστή αυτού συνεπάγεται μεγαλύτερη ακρίβεια του παραγόμενου μοντέλου.

Στο επόμενο κεφάλαιο, ακολουθεί η πρακτική εφαρμογή της προαναφερθείσας μεθοδολογίας, με στόχο την εξαγωγή και αξιολόγηση αποτελεσμάτων από τα πραγματικά δεδομένα του δείγματος.

# 4. Εφαρμογή

Στο κεφάλαιο αυτό εφαρμόζονται στην πράξη όλα όσα αναλύθηκαν σε θεωρητικό επίπεδο προηγουμένως και δημιουργούνται τα 2 μοντέλα που θα προβλέπουν την τιμή ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης, το ένα με τον αλγόριθμο Νευρωνικά Δίκτυα και το άλλο με αυτό των Τυχαίων Δασών, στο περιβάλλον του Rapid Miner.

Στην 1η υπό-ενότητα αναλύεται ο τρόπος άντλησης και προ-επεξεργασίας των δεδομένων που χρησιμοποιήθηκαν. Στη συνεχεία (2η υπό-ενότητα) παρουσιάζεται αναλυτικά ο τρόπος ανάπτυξης και δημιουργίας του μοντέλου με τη χρήση του αλγόριθμου Νευρωνικά Δίκτυα στο Rapid Miner· επίσης παρουσιάζονται και αναλύονται τα εξαγόμενα αποτελέσματα. Αντίστοιχα, το ίδιο γίνεται και στην 3η υπό-ενότητα γιατον αλγόριθμο Τυχαία Δάση (Random Forest). Τέλος και στην 4η υπό-ενότητα γίνεται η σύγκριση των αποτελεσμάτων και των αποδόσεων των 2 παραγόμενων μοντέλων.

## 4.1 Δεδομένα

Η συλλογή των δεδομένων είναι ένα κρίσιμο στάδιο που επηρεάζει σημαντικά την επιτυχή εφαρμογή της ανάλυσης, καθώς αυτά αποτελούν τη βάση ανάπτυξης και υλοποίησης των αλγορίθμων Μηχανικής Μάθησης που θα χρησιμοποιηθούν για την εκτίμηση των τιμών ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης.

### *4.1.1 Inside AIRBNB*

Τα δεδομένα που χρησιμοποιήθηκαν αντλήθηκαν από την ιστοσελίδα Inside Airbnb. Το Inside Airbnb είναι μια ιστοσελίδα που δημιουργήθηκε από τον καλλιτέχνη, ακτιβιστή και τεχνολόγο Murray Cox, ο οποίος συνέλεξε, ανέλυσε τα δεδομένα από τις καταχωρήσεις στην πλατφόρμα AIRBNB και δημιούργησε την ιστοσελίδα έχοντας ως στόχο την αμφισβήτηση του ισχυρισμού της τελευταίας, ότι δηλαδή το 87% των οικοδεσποτών νοικιάζει την κατοικία στην οποία ζει. Η ιστοσελίδα παρέχει τη δυνατότητα πρόσβασης και άμεσης λήψης δεδομένων σχετικά με καταχωρήσεις και κριτικές από ένα μεγάλο αριθμό πόλεων παγκοσμίως. Τα δεδομένα για κάθε περιοχή περιλαμβάνουν πληροφορίες όπως ο τύπος του καταλύματος (ολόκληρο σπίτι, ιδιωτικό δωμάτιο κ.λπ.), ο αριθμός των κριτικών, οι βαθμολογίες, οι συντεταγμένες του καταλύματος, η τιμή ανά διανυκτέρευση και πολλές άλλες ενώ συλλέγονται νέα

δεδομένα περιοδικά, αντικαθιστώντας τα υπάρχοντα για κάθε τοποθεσία με τα ενημερωμένα. Τα δεδομένα παρέχονται σε μορφή CSV καθιστώντας εύκολη την επεξεργασία τους για έρευνα, ανάλυση και δημιουργία μοντέλων πρόβλεψης. Το Inside AIRBNB χρησιμοποιείται ευρέως από ερευνητές, ακαδημαϊκούς, δημοσιογράφους ακόμα και πολιτικούς φορείς.

### *4.1.2 Συλλογή & προ-επεξεργασία δεδομένων*

Η περιοχή μελέτης της παρούσας διπλωματικής εργασίας είναι η πόλη της Θεσσαλονίκης, μια από τις μεγαλύτερες και πιο τουριστικές πόλεις της Ελλάδας καθ' όλη τη διάρκεια του χρόνου, με αυξημένη ζήτηση στα καταλύματα βραχυχρόνιας μίσθωσης. Έτσι από την ιστοσελίδα του Inside AIRBNB έγινε λήψη των δεδομένων σε αρχείο μορφής CSV. Το αρχείο περιλαμβάνει δεδομένα για 4.774 καταλύματα βραχυχρόνιας μίσθωσης.

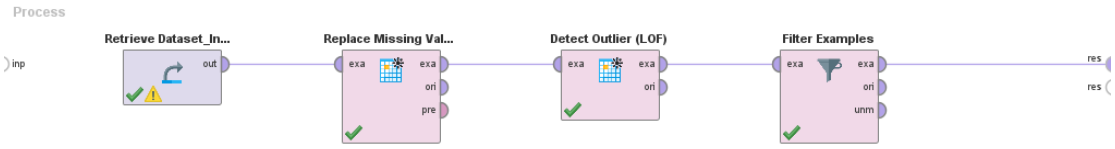
Στο συγκεκριμένο αρχείο περιλαμβάνεται μια πληθώρα πληροφοριών για τα καταχωρημένα στην πλατφόρμα του AIRBNB καταλύματα, από βασικές για το κάθε κατάλυμα πληροφορίες όπως ο μέγιστος αριθμός επισκεπτών, ο αριθμός των υπνοδωματίων των μπάνιων κ.λπ., μέχρι πιο εκειδικευμένες όπως τον μοναδικό αναγνωριστικό αριθμό (ID) του κάθε καταλύματος, τη διαδικτυακή διεύθυνση (URL) που συνδέει το κάθε κατάλυμα με την καταχώρησή του στην πλατφόρμα AIRBNB κ.α.

Αφού το αρχείο CSV με τα παραπάνω δεδομένα μετασχηματίστηκε σε αρχείο EXCEL, με σκοπό την καλύτερη διαχείριση και ανάλυσή του, διατηρήθηκαν οι βασικότερες για κάθε κατάλυμα πληροφορίες, δηλαδή αυτές που επηρεάζουν άμεσα την τιμή ενοικίασής τους. Έτσι επιλέχθηκαν τα εξής δεδομένα: η γειτονιά στην οποία βρίσκεται το κατάλυμα (neighborhood), ο τύπος του καταλύματος (room_type), ο μέγιστος αριθμός επισκεπτών (accommodates), ο αριθμός των συνολικών δωματίων (rooms), ο αριθμός των μπάνιων (bathrooms), ο αριθμός των υπνοδωματίων (bedrooms), ο αριθμός των κλινών (beds), ο αριθμός του ελάχιστου και μέγιστου αριθμού διανυκτερεύσεων (minimum_nights & maximum nights), το πλήθος των κριτικών για το κατάλυμα (number_of_reviews), η συνολική βαθμολογία των αξιολογήσεων (review_scores_rating) και η τιμή ενοικίασης (price).

Ακολουθεί μια πρώτη προετοιμασία των δεδομένων με σκοπό την ορθή χρήση τους στην ανάπτυξη του μοντέλου. Για το σκοπό αυτό πραγματοποιείται η εύρεση και ο

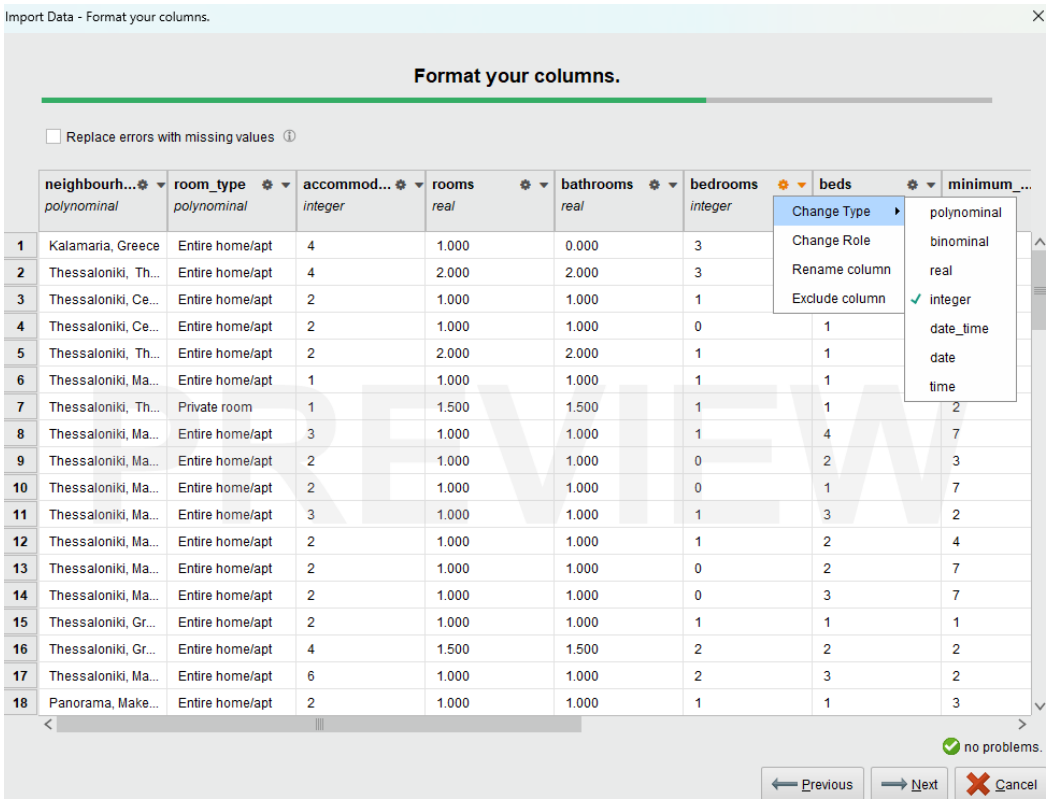
καθαρισμός των δεδομένων, δηλαδή η αντικατάσταση ή/και εξάλειψη ελλιπών και ακραίων τιμών με σκοπό να εξασφαλιστεί η ομαλή λειτουργία των αλγορίθμων. Στη συνέχεια, με σκοπό την εκπαίδευση και αξιολόγηση του μοντέλου, ακολουθεί ο διαχωρισμός του διαθέσιμου συνόλου δεδομένων σε δύο υποσύνολα: ένα σύνολο εκπαίδευσης (training set), που αντιστοιχεί στο 70% των δεδομένων, και ένα σύνολο δοκιμής (test set), που αντιστοιχεί στο υπόλοιπο 30%.

Τα προηγούμενα, όπως αναλύσαμε και σε θεωρητικό επίπεδο (Κεφ. 3, Ενότητα 3.3) μπορούν να πραγματοποιηθούν με τη βοήθεια των τελεστών **"Replace Missing Values", "Detect Outlier (LOF)", "Filter Examples"** και **"Split Data"**. Η πρώτη διεργασία αντικατάστασης ελλιπών τιμών και εύρεσης & ελέγχου των ακραίων φαίνεται στην *εικόνα 4.1*.



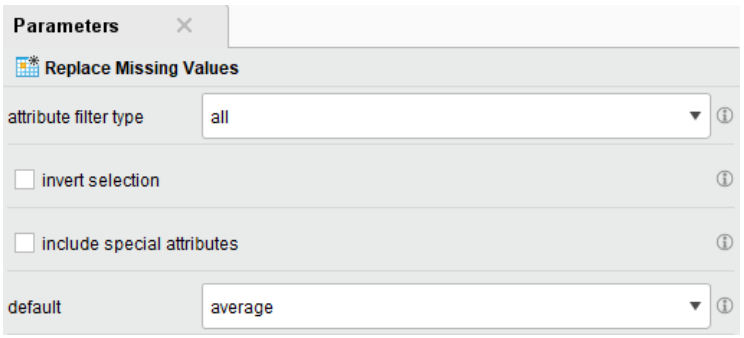

*Εικόνα 4.1 Διάγραμμα ροής προ-επεξεργασίας δεδομένων*

Πιο αναλυτικά, στο περιβάλλον εργασίας εισάγεται το σύνολο των δεδομένων από το αρχείο excel και ορίζεται ο τύπος της κάθε μεταβλητής έτσι ώστε το λογισμικό να μπορεί να τις επεξεργαστεί αναλόγως (*εικόνα 4.2*). Έτσι, όταν έχουμε μια μεταβλητή η οποία είναι κατηγορική με πολλές κατηγορίες, π.χ. η περιοχή του ακινήτου, αυτή ορίζεται ως "polynominal". Από την άλλη όσες μεταβλητές παίρνουν αριθμητικές τιμές, που μπορεί να είναι ακέραιες ή πραγματικές (π.χ. μέγιστος αριθμός επισκεπτών ή τιμή ενοικίασης) αυτές ορίζονται ως "integer" και "real", αντίστοιχα.




*Εικόνα 4.2 Καθορισμός τύπου μεταβλητών δεδομένων κατά την εισαγωγή στο λογισμικό Rapid Miner*

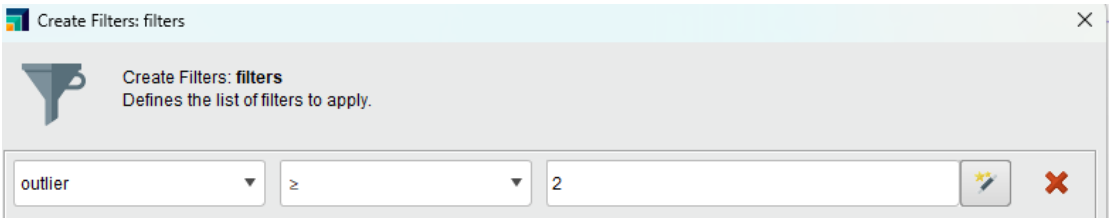
Στη συνέχεια με τον τελεστή “Replace Missing Values” αντικαθίστανται πιθανές κενές τιμές στα δεδομένα μας με το Μ.Ο. της εκάστοτε μεταβλητής (*εικόνα 4.3*).




*Εικόνα 4.3 Ρύθμιση παραμέτρων τελεστή “Replace Missing Values”*

Ακολουθεί ο έλεγχος όλων των παρατηρήσεων, μέσω του τελεστή “Detect Outlier (LOF)”, οι οποίες μπορεί να περιέχουν ακραίες τιμές ώστε να εξακριβωθεί εάν αυτές οφείλονται σε κάποιο λάθος κατά την εισαγωγή των δεδομένων ή εάν απλά πρόκειται για κάποιες «ειδικές» περιπτώσεις ακινήτων που μπορεί να υφίστανται στην πραγματικότητα. Ο τελεστής αυτός προσδίδει ένα outlier score σε κάθε παρατήρηση. Μεγαλύτερη τιμή του 2 στο score αυτό κατά κανόνα συνεπάγεται μεγάλη πιθανότητα ύπαρξης ακραίας τιμής (outlier).

Τέλος και με σκοπό να εξαχθούν μόνο οι παρατηρήσεις με outlier score μεγαλύτερο ή ίσο του 2 (*εικόνα 4.4*), ώστε στη συνέχεια να μπορέσουμε να τις ελέγξουμε χωριστά και να αποφασίσουμε εάν θα προχωρήσουμε στη δημιουργία των μοντέλων έχοντας διατηρήσει ορισμένες ή και όλες τις παρατηρήσεις αυτές, προστίθεται ο τελεστής "Filter Examples".




*Εικόνα 4.4 Δημιουργία συνθήκης στον τελεστή "Filter Examples"*

Τα αποτελέσματα της παραπάνω διεργασίας δίνουν έναν πίνακα αρκετών παρατηρήσεων και συγκεκριμένα 462, τμήμα του οποίου φαίνεται στην *εικόνα 4.5*.

ExampleSet (Filter Examples)

Open in Turbo Prep Auto Model Interactive Analysis Filte

| Row No. | outlier | neighbourho... | room_type | accommoda... | bathrooms_... | bedrooms | beds | minimum_ni... | maximum_n... | price |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.025 | Thessaloniki | Entire home/... | 4 | 2 | 3 | 3 | 28 | 1125 | 50 |
| 2 | 3.789 | Neapolis – Sy... | Entire home/... | 3 | 1 | 2 | 2 | 20 | 60 | 38 |
| 3 | 2.088 | Thessaloniki | Entire home/... | 2 | 1 | 0 | 1 | 7 | 1125 | 45 |
| 4 | 2.650 | Thessaloniki | Entire home/... | 4 | 1 | 1 | 4 | 7 | 1125 | 54 |
| 5 | 2.517 | Thessaloniki | Entire home/... | 2 | 1 | 0 | 1 | 10 | 1125 | 41 |
| 6 | 2.695 | Thessaloniki | Entire home/... | 5 | 1 | 0 | 3 | 10 | 1125 | 80 |
| 7 | 2.197 | Thessaloniki | Entire home/... | 6 | 1 | 2 | 3 | 2 | 1125 | 50 |
| 8 | 2.051 | Thessaloniki | Entire home/... | 4 | 1 | 1 | 1 | 2 | 40 | 56 |
| 9 | 2.580 | Thessaloniki | Entire home/... | 4 | 1 | 3 | 4 | 29 | 1125 | 60 |
| 10 | 2.423 | Neapolis – Sy... | Entire home/... | 4 | 1 | 1 | 2 | 14 | 1125 | 42 |
| 11 | 2.165 | Neapolis – Sy... | Entire home/... | 4 | 1 | 2 | 3 | 3 | 1125 | 53 |
| 12 | 2.203 | Thessaloniki | Private room | 2 | 1 | 1 | 3 | 5 | 1125 | 62 |
| 13 | 2.928 | Thessaloniki | Entire home/... | 5 | 1 | 2 | 3 | 3 | 181 | 84 |
| 14 | 2.584 | Thessaloniki | Entire home/... | 13 | 2 | 4 | 9 | 2 | 1125 | 108 |
| 15 | 2.542 | Ampelokipon ... | Private room | 1 | 2 | 1 | 1 | 14 | 1125 | 15 |
| 16 | 2.435 | Thessaloniki | Entire home/... | 2 | 1 | 1 | 1 | 15 | 1125 | 50 |
| 17 | 2.185 | Thessaloniki | Private room | 2 | 1.500 | 1 | 2 | 1 | 59 | 22 |
| 18 | 2.404 | Thessaloniki | Entire home/... | 4 | 1 | 1 | 2 | 31 | 1124 | 60 |
| 19 | 4.022 | Thessaloniki | Entire home/... | 2 | 1 | 1 | 1 | 15 | 366 | 49 |
| 20 | 6.531 | Thessaloniki | Entire home/... | 2 | 1.500 | 1 | 1 | 90 | 1125 | 16 |
| 21 | 2.550 | Thessaloniki | Entire home/... | 7 | 1 | 3 | 3 | 4 | 55 | 70 |
| 22 | 2.359 | Neapolis – Sy... | Entire home/... | 9 | 1.500 | 2 | 6 | 3 | 252 | 68 |
| 23 | 2.835 | Thessaloniki | Entire home/... | 4 | 1 | 1 | 4 | 10 | 1125 | 100 |
| 24 | 2.883 | Thessaloniki | Entire home/... | 1 | 1 | 1 | 2 | 15 | 1125 | 15 |
| 25 | 2.849 | Thessaloniki | Entire home/... | 4 | 1 | 2 | 2 | 10 | 300 | 500 |

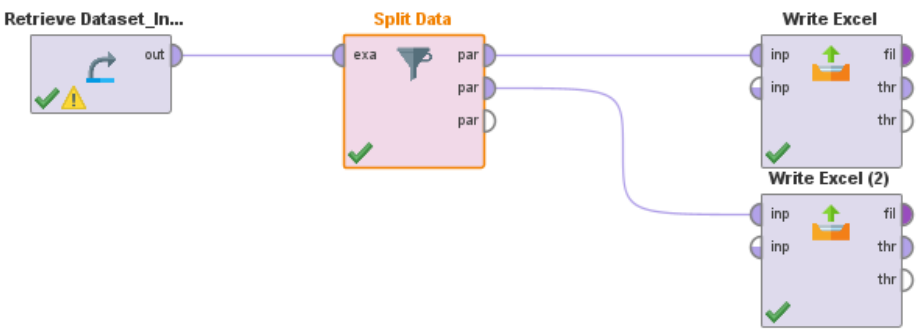
ExampleSet (462 examples,1 special attribute,9 regular attributes)

*Εικόνα 4.5 Παρατηρήσεις με outlier score μεγαλύτερο του 2*

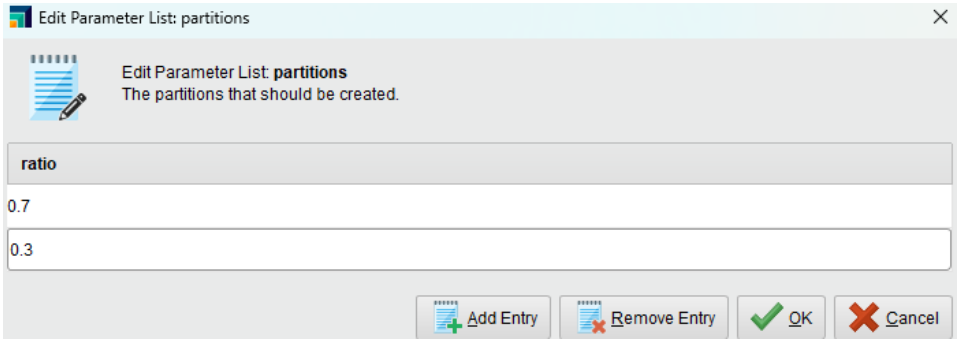
Οι παρατηρήσεις αυτές στη συνέχεια ελέγχθηκαν μεμονωμένα και από την ανάλυση προέκυψε ότι δεν πρόκειται για σφάλματα κατά την εισαγωγή ή καταγραφή των

δεδομένων, αλλά για έγκυρες περιπτώσεις που αιτιολογούνται βάσει των χαρακτηριστικών των καταλυμάτων. Ειδικότερα, παρατηρήθηκαν περιπτώσεις με υψηλό αριθμό επισκεπτών, ο οποίος ωστόσο δικαιολογείται από τον αντίστοιχο αριθμό υπνοδωματίων και κλινών· καταχωρήσεις με πολύ υψηλές τιμές ελάχιστης ή μέγιστης διαμονής· καθώς και καταλύματα με αυξημένο κόστος ενοικίασης, που οφείλεται είτε στον μεγάλο αριθμό επισκεπτών που μπορούν να φιλοξενήσουν τα καταλύματα αυτά είτε στην καλή τους τοποθεσία, σε δημοφιλείς περιοχές της πόλης. Συνεπώς, οι συγκεκριμένες παρατηρήσεις αν και εμφανίζουν υψηλή τιμή στο outlier score, είναι υπαρκτά και συχνά φαινόμενα στην αγορά βραχυχρόνιας μίσθωσης, δεν πρόκειται για σφάλματα ή ανωμαλίες στα δεδομένα, αλλά για περιπτώσεις που μπορούν να εξηγηθούν από τα χαρακτηριστικά των καταλυμάτων. Για τον λόγο αυτό, κρίθηκε απαραίτητο να διατηρηθούν στο σύνολο δεδομένων που θα χρησιμοποιηθεί για την ανάπτυξη του μοντέλου.

Προχωράμε λοιπόν στο διαχωρισμό των «καθαρισμένων» δεδομένων σε 2 υποσύνολα, εκπαίδευσης και επαλήθευσης (training set και test set) με ποσοστά 70% και 30% του συνόλου των δεδομένων αντίστοιχα. Αυτό θα γίνει με τη βοήθεια του τελεστή “Split Data” όπως φαίνεται και στις *εικόνες 4.6 & 4.7.* Ο τελεστής “Write Excel” «γράφει» τα δημιουργούμενα υποσύνολα σε διαφορετικά φύλλα εργασίας Excel δίνοντας μας τη δυνατότητα να τα χρησιμοποιήσουμε στη συνέχεια για την εκπαίδευση και επαλήθευση του μοντέλου που θέλουμε να δημιουργήσουμε. Από την παραπάνω διεργασία προκύπτει ένα αρχείο εκπαίδευσης με δεδομένα 3.342 καταλυμάτων και ένα αρχείο επαλήθευσης με δεδομένα 1.432 καταλυμάτων.




*Εικόνα 4.6 Διάγραμμα ροής στο RapidMiner για τον διαχωρισμό των δεδομένων σε υποσύνολα εκπαίδευσης (training set) και ελέγχου (test set).*




*Εικόνα 4.7 Καθορισμός ποσοστών διαχωρισμού του συνόλου δεδομένων σε training και test set.*

Αφού ολοκληρώθηκε η πρώτη, βασική επεξεργασία του συνόλου των δεδομένων μας και μετά τον διαχωρισμό τους στα υποσύνολα training & test set, είμαστε έτοιμοι να προχωρήσουμε στην δημιουργία των μοντέλων εκτίμησης της τιμής ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης με τους αλγόριθμους Νευρωνικά Δίκτυα & Τυχαία Δάση.

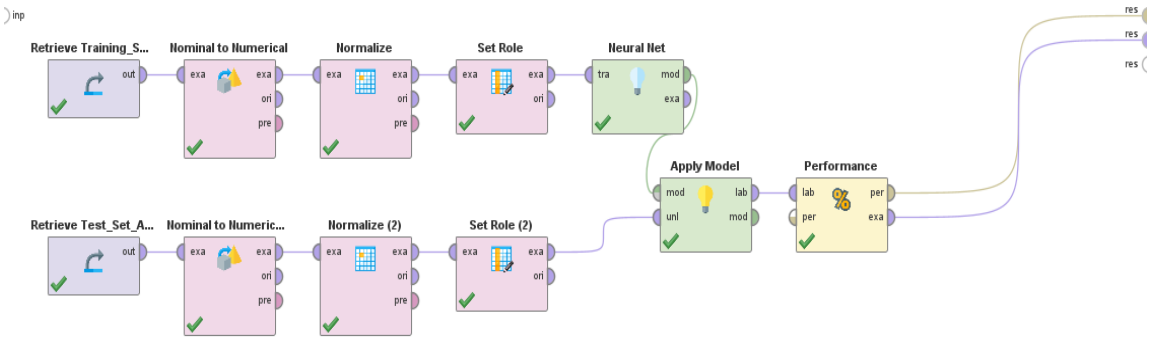
## 4.2 Αλγόριθμος Νευρωνικά Δίκτυα – Neural Network

Εδώ, θα εφαρμόσουμε όσα σε θεωρητικό επίπεδο αναλύθηκαν στο προηγούμενο Κεφάλαιο με σκοπό τη δημιουργία του μοντέλου πρόβλεψης των τιμών ενοικίασης και την αξιολόγηση της απόδοσής του με τη χρήση του αλγορίθμου Νευρωνικά Δίκτυα (Neural Network).

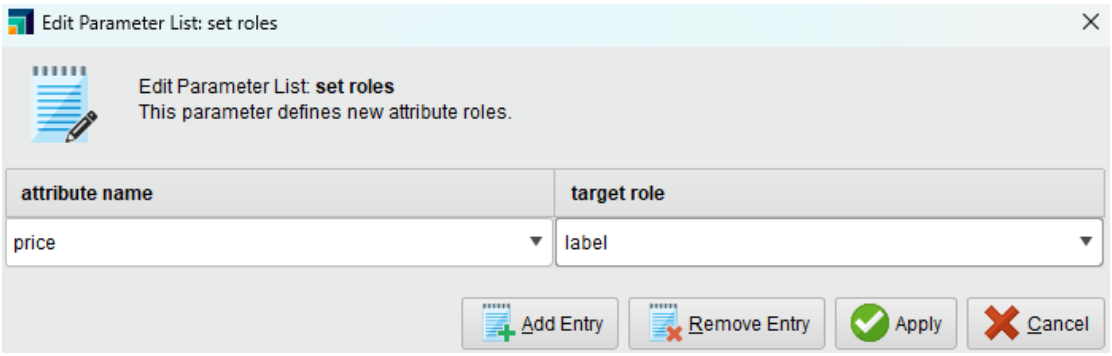
### 4.2.1 Ανάπτυξη μοντέλου

Στην *εικόνα 4.8* φαίνεται το διάγραμμα ροής του δημιουργούμενου μοντέλου στο λογισμικό Rapid Miner.

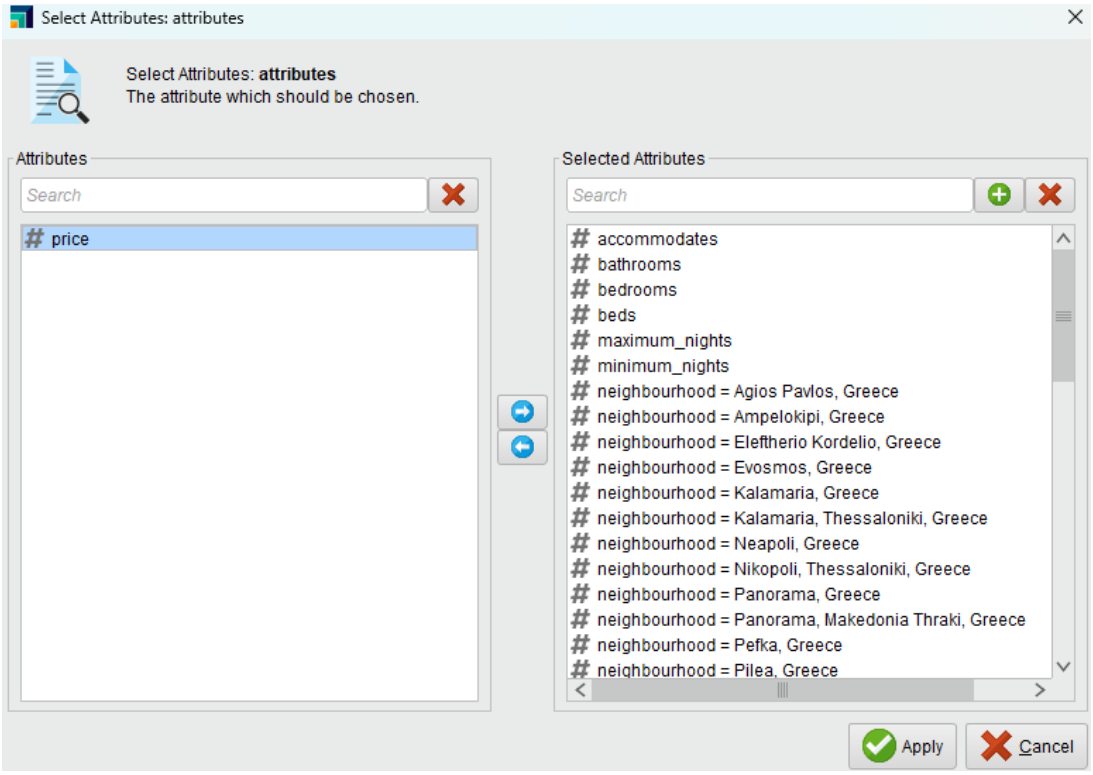
Αρχικά, εισάγονται τα δεδομένα εκπαίδευσης (Training set) και στη συνέχεια ο τελεστής “Nominal to Numerical” κωδικοποιεί όλες τις ονομαστικές μεταβλητές των δεδομένων με σκοπό τη μετατροπή τους σε αριθμητικές. Ακολουθεί ο τελεστής “Normalize” ο οποίος κανονικοποιεί όλα τα αριθμητικά δεδομένα, εκτός από την μεταβλητή «στόχο», ώστε να μπορέσει να λειτουργήσει σωστά το μοντέλο (*εικόνα 4.10*) και στη συνέχεια εφαρμόζεται ο τελεστής “Set Role” στον οποίο ορίζεται η παράμετρος «στόχος» του προβλήματός μας, δηλαδή η τιμή ενοικίασης των καταλυμάτων (price) (*εικόνα 4.9).*




*Εικόνα 4.8 Διάγραμμα ροής στο Rapid Miner για τη δημιουργία μοντέλου πρόβλεψης των τιμών ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης με τον αλγόριθμο Νευρωνικά Δίκτυα.*



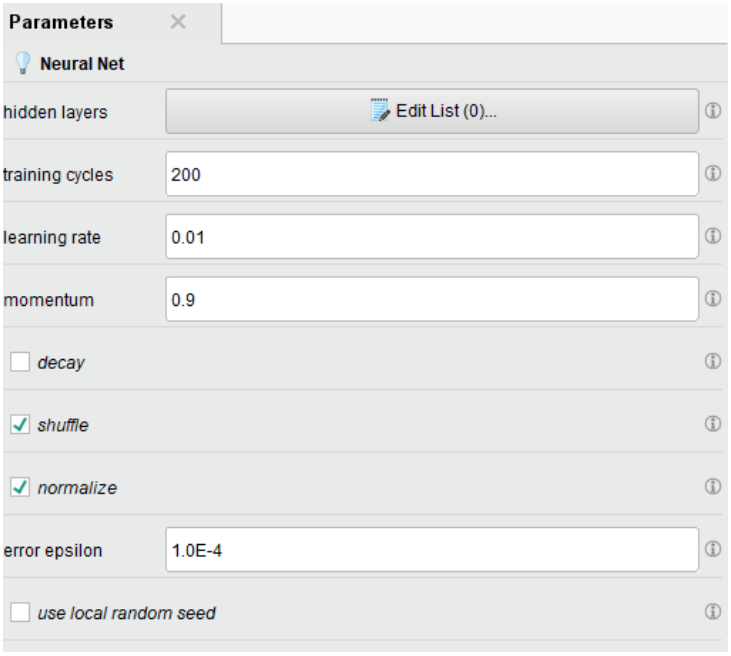

*Εικόνα 4.9 Καθορισμός παραμέτρου «στόχο» του προς παραγωγή μοντέλου.*




*Εικόνα 4.10 Επιλογή αριθμητικών μεταβλητών προς κανονικοποίηση μέσω του τελεστή "Normalize".*

Στη συνέχεια πραγματοποιήθηκαν δύο δοκιμές εκπαίδευσης του μοντέλου· μια χωρίς τη χρήση του τελεστή Cross Validation και μια με τη χρήση αυτού και με την δοκιμή

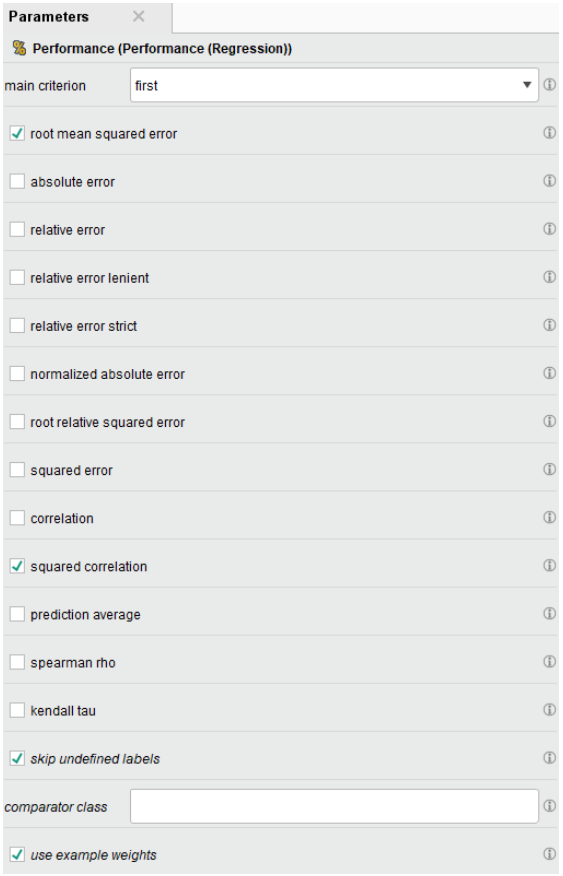
διαφόρων τιμών k υποσυνόλων (k-folds). Αυτό που παρατηρήθηκε είναι ότι παρά το γεγονός της εφαρμογής διασταυρωμένης επικύρωσης στην 2η περίπτωση, η τιμή της απόδοσης του παραγόμενου μοντέλου δεν παρουσίαζε διαφορά συγκριτικά με την περίπτωση κατά την οποία αυτή δεν εφαρμόστηκε. Το γεγονός αυτό μπορεί να αποδοθεί στη σχετική ομοιογένεια και καθαρότητα του συνόλου δεδομένων, καθώς και στη σταθερότητα του αλγόριθμου Νευρωνικά Δίκτυα. Έτσι και μιας και η χρήση πολλαπλών επαναλήψεων με διασταυρούμενη επικύρωση αύξανε σημαντικά τον υπολογιστικό χρόνο, χωρίς αντίστοιχη βελτίωση στην αξιοπιστία των αποτελεσμάτων, επιλέχθηκε η χρήση απλής κατανομής εκπαίδευσης/ελέγχου (train/test split), προκειμένου να εξοικονομηθεί υπολογιστικός χρόνος χωρίς να θιγεί η εγκυρότητα των συμπερασμάτων. Εισάγεται λοιπόν ο αλγόριθμος Νευρωνικά Δίκτυα και ορίζονται όλοι οι παράμετροι αυτού (*εικόνα 4.11*). Με βάση τα όσα αναλύθηκαν σε θεωρητικό επίπεδο στο προηγούμενο Κεφάλαιο το ποσοστό μάθησης ορίζεται ίσο με 0.01, η ορμή 0.9, το σφάλμα-Ε ίσο με 0.0001 ενώ για τον αριθμό των κύκλων εκπαίδευσης πραγματοποιήθηκαν κάποιες δοκιμές με διάφορες τιμές, προκειμένου να εντοπιστεί το βέλτιστο σημείο σύγκλισης του μοντέλου. Η διαδικασία αυτή είχε στόχο την επίτευξη της καλύτερης δυνατής απόδοσης χωρίς υπερπροσαρμογή (overfitting) του μοντέλου. Ακολουθεί ο τελεστής “Apply model” που συνδέεται με το μοντέλο που έχουμε πριν εκπαιδεύσει με τον αλγόριθμο και «περιμένει» τα δεδομένα επικύρωσης για να μπορέσει να υπολογίσει την απόδοση αυτού.




*Εικόνα 4.11 Καθορισμός παραμέτρων αλγόριθμου Neural Network*

Στη συνέχεια λοιπόν, εισάγονται και τα δεδομένα επικύρωσης (test_set) και αφού υποβληθούν στην ίδια αρχική επεξεργασία με αυτή των δεδομένων εκπαίδευσης, μέσω των τελεστών "Nominal to numerical", "Normalize" και "Set Role" ομοίως, συνδέονται και αυτά στον ίδιο τελεστή "Apply Model", ο οποίος θα εφαρμόσει το εκπαιδευόμενο μοντέλο στα νέα δεδομένα.

Τελευταίος εισάγεται ο τελεστής "Performance (Regression)" ο οποίος θα υπολογίσει τους δείκτες απόδοσης που θα εξετάσουμε στο συγκεκριμένο πρόβλημα, δηλαδή το τετράγωνο συντελεστή συσχέτισης - Squared correlation ($R^2$) και τη ρίζα μέσης τετραγωνικής απόκλισης – Root Mean Squared Error (RMSE) (*εικόνα 4.12*). Τέλος, όπως φαίνεται και στο διάγραμμα ροής της *εικόνας 4.8* οι δύο έξοδοι "performance" και "example set" του τελεστή αυτού συνδέονται με τις εξόδους του συνολικού διαγράμματος ροής επιτρέποντας την εξαγωγή και παρουσίαση των αποτελεσμάτων απόδοσης για περαιτέρω ανάλυση, αξιολόγηση και σύγκριση του μοντέλου.




*Εικόνα 4.12 Επιλογή δεικτών απόδοσης προς υπολογισμό από τον τελεστή "Performance"*

### 4.2.2 Αποτελέσματα εφαρμογής μοντέλου

Αφού εφαρμόσουμε το παραπάνω μοντέλο μπορούμε να προχωρήσουμε στην αξιολόγηση των αποτελεσμάτων. Τα εξαγόμενα αποτελέσματα είναι ένα νέο example set που περιλαμβάνει τις αρχικές εγγραφές του test set μαζί με τις νέες προβλέψεις του μοντέλου για τις τιμές ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης καθώς και οι συντελεστές απόδοσης που έχουμε επιλέξει.

Στην *εικόνα 4.13* παρουσιάζεται τμήμα του νέου example set όπου έχουν επισημανθεί οι στήλες με την πραγματική τιμή και την πρόβλεψη που έδωσε το μοντέλο που παρήγαμε (στήλη $1^{η}$ και $2^{η}$ αντίστοιχα). Με τον τρόπο αυτό μας επιτρέπεται η άμεση σύγκριση μεταξύ πραγματικών και προβλεπόμενων τιμών και μπορούμε οπτικά να βγάλουμε κάποια αρχικά συμπεράσματα σχετικά με την ακρίβεια του παραγόμενου μοντέλου.

Αυτό που είναι αρχικά εμφανές είναι ότι υπάρχουν παρατηρήσεις με μεγαλύτερες και άλλες με μικρότερες αποκλίσεις, περίπου σε ίδιο ποσοστό. Συνεπώς δεν μπορούμε οπτικά και μόνο να βγάλουμε κάποιο ασφαλές συμπέρασμα για την απόδοση του μοντέλου. Σε αυτό θα μας βοηθήσουν οι συντελεστές απόδοσης $R^2$ και RMSE που προέκυψαν.

| price | prediction(p... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourh |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 186 | 140.854 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 117 | 97.063 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 104 | 104.046 | 0 | 0 | -0.544 | 0 | -1.350 | 0 | 0 | 6.405 | 0 | 0 | 0 |
| 114 | 124.053 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 129 | 132.336 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 123 | 115.547 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 170 | 127.070 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 150 | 128.390 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 100 | 115.064 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 111 | 113.092 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 132 | 126.379 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 176 | 177.937 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 100 | 100.640 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 116 | 118.729 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 135 | 118.555 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 109 | 104.561 | 0 | 0 | -0.544 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 104 | 104.247 | 0 | 0 | -0.544 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 143 | 116.031 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 125 | 95.999 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 107 | 111.272 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 104 | 105.105 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 124 | 113.926 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 105 | 122.228 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 189 | 176.628 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |

*Εικόνα 4.13 Πίνακας των προβλεπόμενων και πραγματικών τιμών ενοικίασης κάθε παρατήρησης.*

Στον ***πίνακα 4.1*** παρουσιάζονται τα αποτελέσματα των δεικτών απόδοσης $R^2$ και RMSE για τις διαφορετικές τιμές αριθμού κύκλων εκπαίδευσης, όπως προέκυψαν από τις δοκιμές. Η συγκριτική αξιολόγησή τους επιτρέπει την επιλογή του βέλτιστου αριθμού κύκλων, ο οποίος οδηγεί στη μέγιστη ακρίβεια πρόβλεψης του μοντέλου χωρίς να προκαλείται υπερπροσαρμογή.

| Training Cycles | 100 | 150 | 200 |
|---|---|---|---|
| $R^2$ | 0,733 | **0,733** | 0,692 |
| RMSE | 18,446 | **18,435** | 18,627 |

*Πίνακας 4.1: Δείκτες απόδοσης $R^2$ και RMSE για διαφορετικούς αριθμούς κύκλων εκπαίδευσης.*

Παρατηρούμε λοιπόν ότι για 100 κύκλους εκπαίδευσης επιτεύχθηκε τιμή $R^2 = 0,733$ και RMSE = 18,446, ενώ για 150 κύκλους, το **$R^2$** παρέμεινε στο **0,733**, με ελαφρώς βελτιωμένο **RMSE = 18,435**. Αντίθετα, στους 200 κύκλους παρατηρήθηκε μείωση της απόδοσης, με $R^2 = 0,692$ και RMSE = 18,627.

Από τα παραπάνω καθίσταται σαφές ότι η αύξηση των κύκλων εκπαίδευσης πάνω από τους 150 δεν οδηγεί σε βελτίωση της απόδοσης, ενώ ενδέχεται να προκαλεί υπερεκπαίδευση (overfitting). Συνεπώς, ως καταλληλότερη ρύθμιση επιλέγονται οι **150 κύκλοι εκπαίδευσης**, καθώς επιτυγχάνεται η χαμηλότερη τιμή σφάλματος RMSE με σταθερά υψηλή τιμή $R^2$.

Ο συντελεστής $R^2$ λοιπόν προέκυψε ίσος με 0,733 που σημαίνει ότι το μοντέλο είναι ικανό να εξηγήσει περίπου το 73,3% της συνολικής διακύμανσης της εξαρτημένης μας μεταβλητής, δηλαδή της τιμής εκμίσθωσης των καταλυμάτων βραχυχρόνιας μίσθωσης. Η απόδοση αυτή μπορεί να θεωρηθεί αρκετά ικανοποιητική λόγω της φύσης του προβλήματος και της ύπαρξης εξωτερικών παραγόντων που πιθανώς επηρεάζουν την τιμή και δεν έχουν ληφθεί υπόψιν (π.χ. εποχικότητα, εσωτερική διακόσμηση καταλυμάτων, παροχές κ.λπ.).

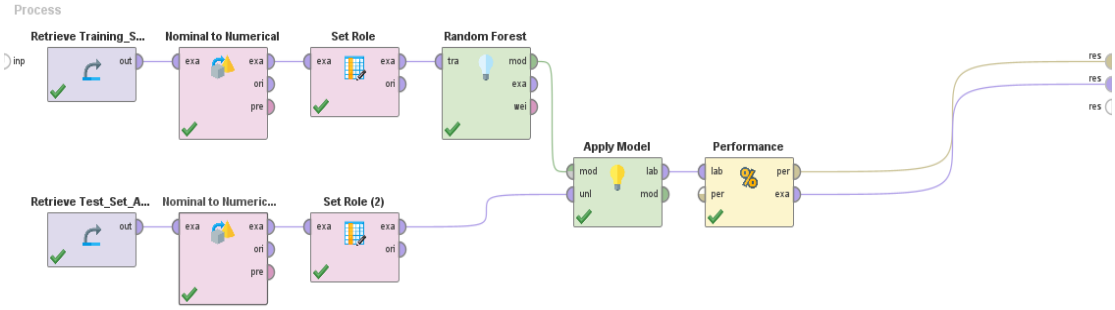
Ο συντελεστής RMSE υπολογίστηκε ίσος με 18,435, που πρακτικά σημαίνει ότι η μέση απόκλιση των προβλέψεων του μοντέλου από τις πραγματικές τιμές ενοικίασης είναι περίπου 18,44 ευρώ. Δεδομένου λοιπόν ότι οι τιμές των καταλυμάτων στο σύνολο δεδομένων κυμαίνονται από 20€ έως 240€, η απόκλιση αυτή θεωρείται σχετικά χαμηλή και υποδηλώνει ικανοποιητική ακρίβεια πρόβλεψης του μοντέλου μας.

## 4.3 Αλγόριθμος Τυχαία Δάση – Random Forest

Εδώ, θα εφαρμόσουμε όσα σε θεωρητικό επίπεδο αναλύθηκαν στο προηγούμενο Κεφάλαιο με σκοπό τη δημιουργία του μοντέλου πρόβλεψης των τιμών ενοικίασης και την αξιολόγηση της απόδοσής του με τη χρήση του αλγορίθμου Τυχαία Δάση (Random Forest).

### 4.3.1 Ανάπτυξη μοντέλου

Στην *εικόνα 4.14* φαίνεται το διάγραμμα ροής του δημιουργούμενου μοντέλου στο λογισμικό Rapid Miner.



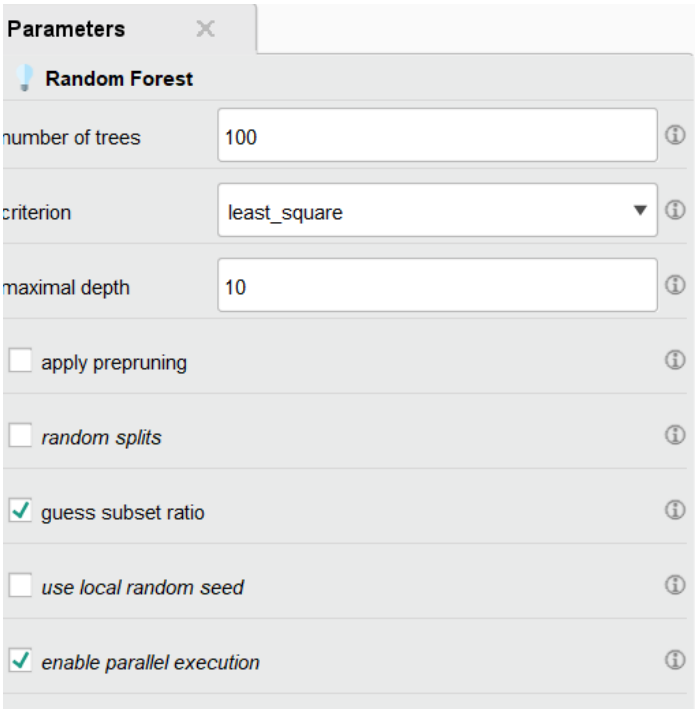

***Εικόνα 4.14 Διάγραμμα ροής στο Rapid Miner για τη δημιουργία μοντέλου πρόβλεψης των τιμών ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης με τον αλγόριθμο Random Forest.***

Όπως φαίνεται και στην παραπάνω εικόνα, η συνολική ροή στο διάγραμμα παραμένει αμετάβλητη, ίδια δηλαδή με την προηγούμενη προσέγγιση, με την μόνη διαφορά την απουσία του τελεστή που πραγματοποιεί την κανονικοποίηση των δεδομένων “Normalization”, ο οποίο στην περίπτωση του αλγορίθμου «Τυχαία Δάση» δεν χρησιμοποιήθηκε καθώς δεν προσφέρει κάτι στην απόδοση και καλύτερη προετοιμασία των εκπαιδευόμενων δεδομένων, όπως εξηγήσαμε και σε θεωρητικό επίπεδο προηγουμένως. Έτσι ακολουθούνται τα εξής βήματα επεξεργασίας των δεδομένων: “Nominal to Numerical” και “Set Role”, “Normalize", στη συνέχεια επιλέγεται ο αλγόριθμος Random Forest, η εφαρμογή του μοντέλου (“Apply Model”) και η αξιολόγησή του [“Performance (Regression)”], όπως ακριβώς έγινε και με τον αλγόριθμο των Νευρωνικών Δικτύων.

Εδώ θα πρέπει να επισημάνουμε ότι και πάλι πραγματοποιήθηκαν δύο δοκιμές εκπαίδευσης του μοντέλου· μια χωρίς τη χρήση του τελεστή Cross Validation και μια με τη χρήση αυτού και με την δοκιμή διαφόρων τιμών k υποσυνόλων (k-folds). Αυτό που παρατηρήθηκε είναι ότι και σε αυτή την περίπτωση παρά το γεγονός της εφαρμογής διασταυρωμένης επικύρωσης, η τιμή της απόδοσης του παραγόμενου

μοντέλου δεν παρουσίαζε διαφορά συγκριτικά με την περίπτωση κατά την οποία αυτή δεν εφαρμόστηκε. Το γεγονός αυτό και πάλι μπορεί να αποδοθεί στη σχετική ομοιογένεια και καθαρότητα του συνόλου δεδομένων, καθώς και στη σταθερότητα του αλγόριθμου Τυχαία Δάση. Έτσι και μιας και η χρήση πολλαπλών επαναλήψεων με διασταυρούμενη επικύρωση αύξανε σημαντικά τον υπολογιστικό χρόνο, χωρίς αντίστοιχη βελτίωση στην αξιοπιστία των αποτελεσμάτων, επιλέχθηκε η χρήση απλής κατανομής εκπαίδευσης/ελέγχου (train/test split), προκειμένου να εξοικονομηθεί υπολογιστικός χρόνος χωρίς να θιγεί η εγκυρότητα των συμπερασμάτων.

Στο πλαίσιο τώρα της βελτιστοποίησης του αλγορίθμου, ακολουθήθηκε μια διαδοχική διαδικασία ρύθμισης υπερπαραμέτρων. Αρχικά, πραγματοποιήθηκαν δοκιμές με διαφορετικό αριθμό δέντρων (number of trees), προκειμένου να εντοπιστεί η τιμή που προσέφερε τη βέλτιστη ακρίβεια πρόβλεψης και αφού καθορίστηκε ο βέλτιστος αριθμός δέντρων, πραγματοποιήθηκαν επιπλέον δοκιμές μεταβάλλοντας το μέγιστο βάθος των δέντρων (maximum depth), με στόχο την περαιτέρω βελτίωση της απόδοσης και τη μείωση του κινδύνου υπερπροσαρμογής (overfitting). Με τη διαδικασία αυτή αναπτύχθηκε τελικά το βέλτιστο αποδοτικά μοντέλο πρόβλεψης με τον αλγόριθμο τω Τυχαίων Δασών. Τα αποτελέσματα των δοκιμών παρουσιάζονται αναλυτικά στους αντίστοιχους πίνακες της επόμενης υποενότητας.




*Εικόνα 4.15 Καθορισμός παραμέτρων αλγόριθμου Random Forest*

### 4.3.2 Αποτελέσματα εφαρμογής μοντέλου

Αφού εφαρμόσουμε το παραπάνω μοντέλο στο Rapid Miner μπορούμε να προχωρήσουμε στην αξιολόγηση των αποτελεσμάτων. Τα εξαγόμενα αποτελέσματα είναι και πάλι το νέο example set που περιλαμβάνει τις αρχικές εγγραφές του test set μαζί με τις νέες προβλέψεις του μοντέλου για τις τιμές ενοικίασης των καταλυμάτων βραχυχρόνιας μίσθωσης, τμήμα του οποίου υπάρχει στην *εικόνα 4.16*. Αυτό που αρχικά παρατηρούμε είναι πως, όπως και στην προηγούμενη εφαρμογή, υπάρχουν παρατηρήσεις με μεγαλύτερες και άλλες με μικρότερες αποκλίσεις, περίπου σε ίδιο ποσοστό. Συνεπώς και πάλι είναι αδύνατον οπτικά και μόνο να βγάλουμε κάποιο ασφαλές συμπέρασμα για την απόδοση του μοντέλου. Σε αυτό θα μας βοηθήσουν οι συντελεστές απόδοσης $R^2$ και RMSE που προέκυψαν.

| price | prediction(p... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourho... | neighbourh |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 186 | 136.180 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 117 | 118.589 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 104 | 117.525 | 0 | 0 | -0.544 | 0 | -1.350 | 0 | 0 | 6.405 | 0 | 0 | 0 |
| 114 | 121.319 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 129 | 127.958 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 123 | 121.685 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 170 | 132.697 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 150 | 132.550 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 100 | 122.789 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 111 | 119.482 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 132 | 123.834 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 176 | 135.783 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 100 | 117.822 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 116 | 121.728 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 135 | 127.918 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 109 | 117.334 | 0 | 0 | -0.544 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 104 | 117.883 | 0 | 0 | -0.544 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 143 | 129.166 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 125 | 120.270 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 107 | 121.091 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 104 | 117.901 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 124 | 121.558 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 105 | 119.813 | 0 | 0 | -0.544 | 0 | 0.723 | 0 | 0 | -0.152 | 0 | 0 | 0 |
| 189 | 142.334 | 0 | 0 | 1.795 | 0 | -1.350 | 0 | 0 | -0.152 | 0 | 0 | 0 |

***Εικόνα 4.16 Πίνακας των προβλεπόμενων και πραγματικών τιμών ενοικίασης κάθε παρατήρησης.***

Στον ***πίνακα 4.2*** παρουσιάζονται τα αποτελέσματα των δεικτών απόδοσης $R^2$ και RMSE για τις διαφορετικές τιμές αριθμού δένδρων απόφασης, όπως προέκυψαν από τις δοκιμές. Η συγκριτική αξιολόγησή τους επιτρέπει την επιλογή του βέλτιστου αριθμού δένδρων.

| Number of trees | 100 | 150 | 200 | 250 | 300 |
|---|---|---|---|---|---|
| $R^2$ | 0,883 | 0,889 | 0,895 | **0,898** | 0,895 |
| RMSE | 13,477 | 13,463 | 13,457 | **13,442** | 13,464 |

*Πίνακας 4.2: Δείκτες απόδοσης $R^2$ και RMSE για διαφορετικούς αριθμούς δένδρων.*

Παρατηρείται λοιπόν ότι, καθώς αυξάνεται ο αριθμός των δέντρων, η απόδοση του μοντέλου βελτιώνεται οριακά, με τον υψηλότερο συντελεστή $R^2$ (0,898) και το χαμηλότερο σφάλμα RMSE (13,442) να καταγράφονται όταν χρησιμοποιούνται 250 δέντρα ενώ οι επιδόσεις για 200 και 300 δέντρα είναι παρόμοιες αλλά ελαφρώς υποδεέστερες, γεγονός που μας οδηγεί στην επιλογή των 250 δέντρων ως βέλτιστη ρύθμιση. Συνεπώς, στο επόμενο στάδιο βελτιστοποίησης, ο αριθμός δέντρων διατηρείται σταθερός (250), ώστε να εξεταστεί η επίδραση του μέγιστου βάθους των δέντρων (maximum depth) στην τελική απόδοση του μοντέλου.

Στον ***πίνακα 4.3*** λοιπόν, παρουσιάζονται τα αποτελέσματα των δεικτών απόδοσης $R^2$ και RMSE για τις διαφορετικές τιμές του μέγιστου βάθους των δένδρων, όπως προέκυψαν από τις δοκιμές με σταθερό αριθμό δένδρων ίσο με 250.

| Maximum Depth | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|
| $R^2$ | **0,898** | 0,887 | 0,887 | 0,887 | 0,887 |
| RMSE | **13,442** | 13,412 | 13,397 | 13,341 | 13,314 |

*Πίνακας 4.3: Δείκτες απόδοσης $R^2$ και RMSE για διαφορετικό μέγιστο βάθος.*

Παρατηρείται λοιπόν ότι η βέλτιστη τιμή $R^2$ (0,898) επιτυγχάνεται όταν το μέγιστο βάθος οριστεί σε 10, ενώ για μεγαλύτερα βάθη (15 έως 30) δεν παρατηρείται κάποια βελτίωση, καθώς οι τιμές του $R^2$ παραμένουν σταθερές (0,887) και στο σφάλμα RMSE παρατηρείτε οριακή μείωση. Το γεγονός αυτό υποδηλώνει ότι η αύξηση του βάθους δεν οδηγεί σε ουσιαστική βελτίωση της απόδοσης και ενδέχεται να εισάγει περιττή πολυπλοκότητα στο μοντέλο και να μας «κοστίσει» σε υπολογιστικό χρόνο. Συνεπώς, η τιμή Maximum Depth ίση με 10 επιλέγεται ως βέλτιστη.

Τελικά, για τις παραμέτρους που επιλέχθηκαν, ο συντελεστής $R^2$ προέκυψε ίσος με 0,898 κάτι που πρακτικά σημαίνει ότι το μοντέλο εξηγεί το 89,8% της συνολικής διακύμανσης της εξαρτημένης μεταβλητής (τιμή ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης) και συνεπώς έχουμε μια πολύ καλή προσαρμογή στα

δεδομένα. Έχουμε δηλαδή ένα αρκετά ακριβές μοντέλο με υψηλή ικανότητα αποτελεσματικής μοντελοποίησης των σχέσεων μεταξύ των μεταβλητών, ιδιαιτέρως αν λάβουμε υπόψιν τη φύση του προβλήματος και την ποικιλότητα των διαφορετικών χαρακτηριστικών από τα οποία μπορεί να εξαρτάται η μεταβολή της μεταβλητής στόχου.

Ο συντελεστής RMSE υπολογίστηκε ίσος με 13,442, που πρακτικά σημαίνει ότι η μέση απόκλιση των προβλέψεων του μοντέλου από τις πραγματικές τιμές ενοικίασης είναι περίπου 13,44 ευρώ. Η τιμή αυτή, αν και δεν είναι αμελητέα, κυμαίνεται σε ικανοποιητικά επίπεδα, δεδομένης της συνολικής διακύμανσης των τιμών ενοικίασης των καταλυμάτων στο σύνολο δεδομένων (20€ έως 240€). Να επισημανθεί επίσης ότι ο συγκεκριμένος δείκτης απόδοσης είναι αρκετά ευαίσθητος σε outliers. Επομένως η διατήρησή του σε αυτό το επίπεδο υποδηλώνει ότι το μοντέλο δεν παρουσιάζει συστηματικά μεγάλα σφάλματα πρόβλεψης.

### 4.4 Σύγκριση αλγορίθμων

Για την πρόβλεψη των τιμών ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης, εφαρμόστηκαν δύο διαφορετικοί αλγόριθμοι μηχανικής μάθησης: οι Νευρωνικά Δίκτυα και Τυχαία Δάση. Και τα δύο μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν βάσει των ίδιων δεδομένων και των ίδιων μετρικών. Συγκεκριμένα για την αξιολόγηση χρησιμοποιήθηκαν οι συντελεστές: προσδιορισμού $R^2$ και σφάλματος RMSE. Στην υπό-ενότητα αυτή λοιπόν πραγματοποιείται η συγκριτική ανάλυση των 2 μοντέλων και παρέχεται ουσιαστική εικόνα για την σχετική τους αποτελεσματικότητα.

Στον παρακάτω πίνακα 4.4 παρουσιάζονται οι δείκτες απόδοσης $R^2$ και RMSE που προέκυψαν τελικά για τον κάθε αλγόριθμο.

| Αλγόριθμος | Νευρωνικά Δίκτυα | Τυχαία Δάση |
|---|---|---|
| $R^2$ | 0,733 | **0,898** |
| RMSE | 18,435 | **13,442** |

Παρατηρούμε τελικά ότι το μοντέλο με τον αλγόριθμο Νευρωνικά Δίκτυα πέτυχε τιμή $R^2$=0,733 που σημαίνει ότι εξηγεί περίπου το 73,3% της διακύμανσης στις τιμές

ενοικίασης. Το αντίστοιχο RMSE ήταν 18,435, υποδηλώνοντας μια μέση απόκλιση 18 ευρώ μεταξύ των πραγματικών και των προβλεπόμενων τιμών. Η απόδοση αυτή χαρακτηρίζεται ως ικανοποιητική, ωστόσο φανερώνει ότι υπάρχει ένα μη αμελητέο ποσοστό σφάλματος που δεν μπορεί να εξηγηθεί από το μοντέλο.

Αντίθετα, το μοντέλο που βασίστηκε στον αλγόριθμο Τυχαία Δάση κατέγραψε σημαντικά καλύτερη επίδοση στον δείκτη $R^2$, με τιμή 0,898, δηλαδή κατάφερε να εξηγήσει 89.8% της συνολικής διακύμανσης. Πρόκειται για σαφώς υψηλότερη προσαρμογή, γεγονός που δείχνει ότι το μοντέλο ενσωματώνει και αξιοποιεί καλύτερα τις σχέσεις μεταξύ των ανεξάρτητων και της εξαρτημένης μεταβλητής. Από πλευράς RMSE, η τιμή διαμορφώθηκε στα 13,442 ευρώ, δηλαδή και πάλι σημαντικά βελτιωμένο μέσο σφάλμα σε σχέση με τον αλγόριθμο των Νευρωνικών Δικτύων.

Επίσης, θα πρέπει να λάβουμε υπόψιν ότι ο αλγόριθμος των Νευρωνικών Δικτύων παρουσίασε εκτός από περιορισμένη γενικευσιμότητα παρουσιάζει και μικρότερη σταθερότητα, μιας και παρατηρήθηκαν μεγαλύτερες διακυμάνσεις στις δοκιμές. Από την άλλη το μοντέλο των Τυχαίων Δασών είχε σχεδόν σταθερή απόδοση σε διαφορετικές τιμές παραμέτρων γεγονός που το καθιστά πιο ανθεκτικό στην υπερπροσαρμογή και του επιτρέπει να γενικεύει καλύτερα. Τέλος, τα Νευρωνικά Δίκτυα απαιτούν κανονικοποίηση των μεταβλητών με σκοπό την ορθή λειτουργία και χρειάζεται μεγαλύτερο χρόνο εκπαίδευσης, ιδίως όταν αυξάνεται ο αριθμός των κύκλων. Αντίθετα, το Random Forest δεν προϋποθέτει κανονικοποίηση των δεδομένων και είναι πιο αποδοτικό χρονικά.

Τελικά, και όπως προκύπτει από όλα τα παραπάνω, το μοντέλο Τυχαίων Δασών φαίνεται να υπερτερεί συνολικά στο συγκεκριμένο πρόβλημα παλινδρόμησης, καθώς παρέχει πιο συνεπή και ισχυρή ερμηνεία της μεταβλητότητας των τιμών ενοικίασης καθώς και χαμηλότερο μέσο σφάλμα στις μεμονωμένες προβλέψεις. Το μοντέλο των Νευρωνικών Δικτύων, από την άλλη, υστερεί σε βάθος πρόβλεψης και ικανότητα γενίκευσης.

# 5. Συμπεράσματα - Προτάσεις για μελλοντική έρευνα

Μετά την ολοκλήρωση της ανάλυσης και ανάπτυξης των μοντέλων πρόβλεψης και την αξιολόγηση των επιδόσεων των επιλεγμένων αλγορίθμων μηχανικής μάθησης, ακολουθεί η παρουσίαση των βασικών συμπερασμάτων που προκύπτουν από την παρούσα μελέτη. Παράλληλα, παρατίθεται μια συνοπτική αποτίμηση της μεθοδολογικής προσέγγισης που ακολουθήθηκε, ορισμένοι περιορισμοί καθώς και προτάσεις για τη διεύρυνση της έρευνας σε μελλοντικό επίπεδο.

## 5.1 Συμπεράσματα

Η παρούσα εργασία μελέτησε την εφαρμογή αλγορίθμων μηχανικής μάθησης για την πρόβλεψη των τιμών ενοικίασης καταλυμάτων βραχυχρόνιας μίσθωσης στην πόλη της Θεσσαλονίκης. Στόχος ήταν να διερευνηθεί σε ποιον βαθμό μοντέλα όπως τα Νευρωνικά Δίκτυα και τα Τυχαία Δάση μπορούν να εξηγήσουν και να προβλέψουν την τιμή ενός καταλύματος βάσει χαρακτηριστικών του όπως η τοποθεσία, ο τύπος ακινήτου κ.ά. και να διερευνηθεί πιο από αυτά λειτουργεί πιο αποτελεσματικά και μπορεί να δώσει αξιόπιστα αποτελέσματα.

Ειδικότερα, ο αλγόριθμος Τυχαία Δάση (Random Forest) παρουσίασε τη μεγαλύτερη προσαρμογή στα δεδομένα, επιτυγχάνοντας υψηλή τιμή $R^2$ (0,898) και μικρότερο μέσο σφάλμα πρόβλεψης RMSE (13,442), αναδεικνύοντας το ως την καλύτερη επιλογή για το υπό μελέτη πρόβλημα.

Τα αποτελέσματα επιβεβαιώνουν ότι η έξυπνη αξιοποίηση των δεδομένων μέσω της μηχανικής μάθησης μπορεί να αποτελέσει ένα ισχυρό εργαλείο υποστήριξης αποφάσεων για ιδιοκτήτες, επενδυτές και πλατφόρμες βραχυχρόνιας μίσθωσης. Ωστόσο, υπάρχουν ακόμα σημαντικά περιθώρια βελτίωσης και εμβάθυνσης.

## 5.2 Περιορισμοί της έρευνας

Παρά το γεγονός ότι η παρούσα ερευνητική εργασία σχεδιάστηκε και υλοποιήθηκε με προσοχή και συνέπεια, ορισμένοι περιορισμοί που μπορεί να σχετίζονται είτε με την ίδια τη φύση του προβλήματος, είτε με πρακτικά ζητήματα, είναι αναπόφευκτοι και η αναγνώρισή τους είναι αναγκαία τόσο για την έγκυρη ερμηνεία των αποτελεσμάτων όσο και για τη διαμόρφωση προτάσεων για μελλοντική έρευνα.+

- ***Περιορισμένος αριθμός δείγματος.***

  Η μελέτη βασίστηκε σε δεδομένα περίπου 5.000 παρατηρήσεων, που κατά κανόνα μπορεί να θεωρηθούν ένα ικανοποιητικό μέγεθος δείγματος για προβλήματα παλινδρόμησης. Ωστόσο, μεγαλύτερα σύνολα δεδομένων θα μπορούσαν να ενισχύσουν περαιτέρω τη γενικευσιμότητα των αποτελεσμάτων.

- ***Απλουστευμένη χωρική πληροφορία.***

  Αν και χρησιμοποιήθηκε η βασική διαθέσιμη γεωγραφική πληροφορία, η περιοχή δηλαδή του διαμερίσματος, δεν ενσωματώθηκαν πιο σύνθετα χωρικά χαρακτηριστικά (όπως αποστάσεις από σημεία ενδιαφέροντος, γεωγραφικές συντεταγμένες κ.α.), τα οποία θα μπορούσαν να ενισχύσουν την ακρίβεια των προβλέψεων.

- ***Έλλειψη χρονικών/εποχικών μεταβλητών.***

  Η μελέτη εστίασε σε αριθμητικά και κατηγορικά χαρακτηριστικά που ήταν διαθέσιμα στο dataset. Η ενσωμάτωση επιπλέον εξωτερικών δεδομένων, όπως η εποχικότητα της ζήτησης, ξεπερνούσε το πεδίο της παρούσας εργασίας, αλλά θα μπορούσε να αποτελέσει επέκταση μελλοντικά καθώς επηρεάζουν σημαντικά τις τιμές ενοικίασης βραχυχρόνιας μίσθωσης.

- ***Χρήση του RapidMiner ως κύριου εργαλείου.***

  Το RapidMiner επιλέχθηκε λόγω της δυνατότητας δοκιμών χωρίς κώδικα. Αν και ιδιαίτερα αποδοτικό, περιορίζει τη δυνατότητα εκτεταμένης προσαρμογής σε σύγκριση με προγραμματιστικά εργαλεία (π.χ. Python ή R).

Οι παραπάνω περιορισμοί καθιστούν σαφές ότι υπάρχει περιθώριο περαιτέρω βελτίωσης της μεθοδολογίας και εμπλουτισμού των δεδομένων, γεγονός που ανοίγει τον δρόμο για μελλοντική ερευνητική δραστηριότητα.

## 5.3 Προτάσεις για μελλοντική έρευνα

Για μελλοντική έρευνα, προτείνεται να ενταχθούν εμπλουτισμένα και εξωτερικά δεδομένα (π.χ. τουριστική ζήτηση, εποχικότητα, κοινωνικοοικονομικά χαρακτηριστικά περιοχών, πιο συγκεκριμένα χωρικά δεδομένα) ώστε να ενισχυθεί το πληροφοριακό περιεχόμενο των παραμέτρων με σκοπό την αύξηση της ακρίβειας των προβλέψεων και την ανάδειξη επιπλέον συσχετίσεων. Επιπλέον, η διεύρυνση του γεωγραφικού πεδίου μελέτης και των χρονικών σειρών, θα επέτρεπε τη μελέτη της δυναμικής μεταβολής των τιμών τόσο σε διαφορετικά τουριστικά προφίλ όσο και στο χρόνο.

Τέλος, η αξιολόγηση και άλλων, πιο σύγχρονων και εξελιγμένων αλγορίθμων (όπως XGBoost, LightGBM ή ακόμα και deep learning προσεγγίσεων) θα μπορούσε να προσφέρει βελτιωμένα αποτελέσματα και βαθύτερη κατανόηση της πολυπλοκότητας του προβλήματος.

# Βιβλιογραφία

**Ξενόγλωσση**


Baharun N., Faezah N., Masrom S. & Mohamad Yusri N. A. (2022). *Auto Modelling for Machine Learning: A Comparison Implementation between RapidMiner and Python.* International Journal of Emerging Technology and Advanced Engineering 12(5):15-27

Barbierato E. & Gatti A. (2024). *The Challenges of Machine Learning: A Critical Review*. *Electronics, 13*(2), 416.

Fernando J. (2024). *R-Squared: Definition, Calculation, and Interpretation.* Investopedia.

Fitria F. & Pebriadi M. S. (2025). *House Price Prediction Using the Random Forest Algorithm on the Rapidminer Application.* Formosa Journal of Science and Technology 4(2):727-738

Hu C., Huang R. & Li H. (2022). *Prediction and Analysis of Rental Price using Random Forest Machine Learning Technique Take Shanghai and Wuhan for example.* International Conference on Mathematical Statistics and Economic Analysis (MSEA 2022) (pp.587-593)

Islam, M. D., Li, B., Islam, K. S., Ahasan, R., Mia, M. R., & Haque, M. E. (2022). Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model. *Machine Learning with Applications*

Jinwen Tang, Jinlin Cheng & Min Zhang. (2023). *Forecasting Airbnb prices through machine learning.* Managerial and Decision Economics

Jovanovic M. Z., Vukicevic M., Delibašić B. & Suknovic M. (2014*). Using RapidMiner for Research: Experimental Evaluation of Learners.* Research Gate.

Karthikraj V., Patil V., Thanneermalai S. & Murugesh V. (2021). International Conference on Advances in Computing, Communication, and Control (ICAC3). Mumbai, India.

Ketkar Y. & Gawade S. (2022). *A decision support system for selecting the most suitable machine learning in healthcare using user parameters and requirements.*

Masrom S., Baharun N., Razi N. F. M. & Rahman R. A. (2022). *Particle Swarm Optimization in Machine Learning Prediction of Airbnb Hospitality Price Prediction.* International Journal of Emerging Technology and Advanced Engineering

Samarasinghe, R., & Mollah, M. B. (2022). *The determinants of Airbnb prices in an emerging market: Evidence from Colombo, Sri Lanka*. Managerial and Decision Economics, 43

Shekh R., Neyaz A., Ahmed A. & Singh A. (2023). *Machine Learning for Rental Price Prediction: Regression Techniques and Random Forest Model.* SSRN Electronic Journal

Tarando-Vera G., Galindo-Villardón G., Merchán-Sánchez-Jara J., Salazar-Pozo J., Moreno-Salazar A. & Salazar-Villalva V. (2021). *Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature.* The Journal of Supercomputing.

Yang S. (2021). *Learning-based Airbnb Price Prediction Model.* 2nd International Conference on E-Commerce and Internet Technology (ECIT). Hangzhou, China.

Zhou, Y., Wang, H., & Liu, H. (2021). *Research on Airbnb house price prediction based on machine learning*. In Y. Wang & J. Liu (Eds.), Algorithms and architectures for parallel processing. ICA3PP 2021. Lecture Notes in Computer Science (Vol. 13077, pp. 133–143). Springer.

Zhu A., Li R. & Xie Z. (2020). *Machine Learning Prediction of New York Airbnb Prices.* Third International Conference on Artificial Intelligence for Industries (AI4I)

**Ελληνική**

Βιαννιτάκη Β. (2014). *Τεχνητή νοημοσύνη, ευφυής πράκτορες και εφαρμογές στην πληροφορική υγείας* (Πτυχιακή εργασία, τμήμα Διοίκησης επιχειρήσεων και οργανισμών, ΤΕΙ Πελοποννήσου).

Γεωργακόπουλος Α. (2022). *Πρόβλεψη μέσης ωριαίας κατανάλωσης ηλεκτρικής ισχύος με χρήση μεθόδων μηχανικής μάθησης* (Διπλωματική εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο).

Γκόγκος Χ. (2022). *Εξέλιξη της μηχανικής μάθησης και της επεξεργασίας φυσικής γλώσσας και η επίδρασή τους στην οικονομίας* [Διαφάνειες Παρουσίασης]. Θερινό σχολείο ιστορίας και οικονομίας Πρέβεζας.

Ζησόπουλος Φ. (2019). *Αναλυτική δεδομένων στις επιχειρήσεις: Μελέτη περίπτωσης AIRBNB* (Διπλωματική εργασία, Διατμηματικό πρόγραμμα μεταπτυχιακών σπουδών στη διοίκηση επιχειρήσεων, Πανεπιστήμιο Μακεδονίας).

Ιακωβίδης Σ. (2008). *Ηλεκτρομαγνητική ανίχνευση υπόγειων στόχων με νευρωνικά δίκτυα* (Διπλωματική εργασία, τμήμα Ηλεκτρονικής φυσικής (Ραδιοηλεκτρολογίας), Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης).

Καλούτσα Β. (2018). *Μοντελοποίηση ανεμογενήτριας με νευρωνικά δίκτυα* (Διπλωματική εργασία, τμήμα Μηαχνολόγων μηχανικών, Πανεπιστήμιο Θεσσαλίας).

Καραμπλιάς Β. (2023). *Τεχνικές ανάπτυξης αλγόριθμων μηχανικής μάθησης: επισκόπηση μεθόδων και εφαρμογή σε ναυτιλιακά δεδο*μένα (Διπλωματική εργασία, τμήμα Ηλεκτρολόγων μηχανικών και μηχανικών υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο).

Κονταξάκης Α. (2017). *Ανάλυση δεδομένων σε ένα πέρασμα στο Rapid Miner* (Διπλωματική εργασία, Τμήμα Ηλεκτρολόγων και μηχανικών υπολογιστών, Πολυτεχνείο Κρήτης).

Λαμπρινή Χ. (2024). *Χρήση αλγόριθμων μηχανικής μάθησης για την εκτίμηση τιμής πώλησης ακινήτου* (Διπλωματική εργασία, τμήματα πληροφορικής και οικονομικών επιστημών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης).

Λυκοθανάσης Σ. (χ.χ.). *Εισαγωγή στα Τεχνητά Νευρωνικά Δίκτυα* [Διαφάνειες Παρουσίασης]. Εργαστήριο Αναγνώρισης Προτύπων, Τμήμα Μηχανικών Η/Υ και Πληροφορικής.

Ματθαίου Χ. (2020). *Μηχανική Μάθηση*. Nowmag.gr

Μηναϊδη Μ. (2020). *Τεχνικές μηχανικής μάθησης σε προβλήματα πρόβλεψης τιμών ακινήτων* (Διπλωματική εργασία, τμήμα Ηλεκτρολόγων μηχανικών και μηχανικών υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο).

Μπάστας Κ. (2022). *Τεχνικές μηχανικής μάθησης για την ανάλυση συναισθημάτων σε κείμενο* (Πτυχιακή εργασία, τμήμα Διοικητικής επιστήμης και τεχνολογίας, Πανεπιστήμιο Πατρών).

Παλαιολόγος Χ. (2009). *Ταξινόμηση με χρήση αλγορίθμων Data Mining και Ασαφούς λογικής: Μια εφαρμογή στο μετρό του Παρισιού* (Διπλωματική εργασία, τμήμα Ηλεκτρολόγων μηχανικών και μηχανικών υπολογιστών, Πολυτεχνείο Κρήτης).

Πασόη Γ. (2022). *Εφαρμογή προβλέψεων στη χρηματοοικονομική αγορά με τεχνικές μηχανικής μάθησης και ανάλυσης συναισθήματος* (Διπλωματική εργασία, τμήμα Εφαρμοσμένης πληροφορικής, Πανεπιστήμιο Μακεδονίας).

Περίκος Ι. (χ.χ.). *Κεφάλαιο Μηχανική Μάθηση: Ευφυή συστήματα λήψης απόφασης στις επιστήμες υγείας* [Διαφάνειες Παρουσίασης]. Τεχνολογικό Εκπαιδευτικό Ίδρυμα Δυτικής Ελλάδας.

Πολύζος Μ. (2019). *Μεθοδολογία πρόβλεψης της υπολειμματικής αξίας δομικών μηχανημάτων με τη χρήση λογισμικού μηχανικής μάθησης Rapid Miner* (Διπλωματική εργασία, τμήμα Πολιτικών μηχανικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης).

Σταυρουλάκης Ι. (2020). *Πρόβλεψη τιμών και εξόρυξη γνώμης από κριτικές σε διαδικτυακές πλατφόρμες βραχυχρόνιας μίσθωσης ακινήτων* (Διπλωματική εργασία, τμήμα Μηχανικών παραγωγής και διοίκησης, Πολυτεχνείο Κρήτης).

Τζεδάκης Χ. (2014). *Ανασκόπηση της εφαρμογής των μεθόδων μηχανικής μάθησης στη βιοπληροφορική* (Διπλωματική εργασία, τμήμα Ηλεκτρολόγων μηχανικών και μηχανικών υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο).

Τζέκας Α. (2018). *Το Rapid Miner ως εργαλείο εφαρμογών Big Data Analytics* (Διπλωματική εργασία, τμήμα Ψηφιακών συστημάτων, Πανεπιστήμιο Πειραιώς).

Τσίπουρας Μ. (2016). *Τεχνητή νοημοσύνη* [Διαφάνειες Παρουσίασης]. Τμήμα Διοίκησης Επιχειρήσεων. ΤΕΙ Δυτικής Μακεδονίας.

**Links**

*20 Best Machine Learning Software Tools in 2024*

*KDD in Data Mining- Scaler Topics*

*Root Mean Square Error (RMSE)*

*Κατανόηση της επιβλεπόμενης και μη επιβλεπόμενης μάθησης*

*Η ιστορία της «Airbnb» που σήμερα αποτιμάται στα 25,5 δισ. δολάρια*

*Τι είναι η τεχνητή νοημοσύνη και πώς χρησιμοποιείται; | Θέματα | Ευρωπαϊκό Κοινοβούλιο*