ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ : ΕΠΙΣΤΗΜΗ & ΤΕΧΝΟΛΟΓΙΑ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΡΟΒΛΕΨΗ ΤΙΜΩΝ ΣΤΗΝ ΑΓΟΡΑ ΑΚΙΝΗΤΩΝ ΤΗΣ ΣΙΓΚΑΠΟΥΡΗΣ

ΒΑΣΙΛΕΙΟΣ ΠΑΠΑΖΗΣΗΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
ΑΓΓΕΛΟΣ ΣΙΦΑΛΕΡΑΣ

ΘΕΣΣΑΛΟΝΙΚΗ
2023-2024

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία με τίτλο «Ανάλυση Δεδομένων και Πρόβλεψη Τιμών στην Αγορά Ακινήτων της Σιγκαπούρης» αναφέρεται στην εφαρμογή της επιστήμης δεδομένων στον τομέα της αγοράς ακινήτων. Αναλυτικότερα, πραγματοποιήθηκε ανάλυση δεδομένων της βάσης ‘’Resale flat prices based on registration date from Jan-2017 onwards ’’, η οποία αντλήθηκε από την ιστοσελίδα της Κυβερνητικής Υπηρεσίας Σιγκαπούρης (https://beta.data.gov.sg). Έπειτα, δημιουργήθηκαν τέσσερα μοντέλα επιβλεπόμενης μηχανικής μάθησης που προβλέπουν την τιμή μεταπώλησης, τα οποία αξιολογήθηκαν με την χρήση τεσσάρων μετρικών και στην συνέχεια συγκρίθηκε η απόδοσή τους για να ανακηρυχθεί το καλύτερο.

Η πτυχιακή εργασία έχει συγγραφεί με βάση τα επιστημονικά περιοδικά και την διεθνή βιβλιογραφία. Η ανάλυση δεδομένων πραγματοποιήθηκε με γνώμονα την εξαγωγή πληροφοριών σχετικά με διάφορες πτυχές της αγοράς ακινήτων, βασιζόμενη στην οπτικοποίηση των δεδομένων μέσα από διαγράμματα όπως, boxplot, scatterplot , ραβδόγραμμα και διάγραμμα πίτα. Παράλληλα, για την δημιουργία των μοντέλων επιλέχθηκαν αλγόριθμοι όπως η παλινδρόμηση Ridge , η παλινδρόμηση Lasso, τα τυχαία δάση και τα δέντρα απόφασης. Στη συνέχεια αξιολογήθηκαν με την χρήση τεσσάρων μετρικών αξιολόγησης όπως, το μέσο τετραγωνικό σφάλμα, η ρίζα μέσου τετραγωνικού σφάλματος, το μέσο απόλυτο σφάλμα και η μετρική R-squared.

Η βάση δεδομένων που αξιοποιήθηκε και ο κώδικας εξ΄ ολοκλήρου γραμμένος σε Python, βρίσκεται αναρτημένος στο GitHub (https://github.com/PapBill) και στο Kaggle (https://www.kaggle.com/vasilispapazisis)

Λέξεις κλειδιά: μηχανική μάθηση, ανάλυση δεδομένων, επιστήμη δεδομένων, python.

ABSTRACT

This thesis titled "Data Analysis and Price Forecasting in the Singapore Property Market" is about the application of data science in the property market. Specifically, data analysis of the database ''Resale flat prices based on registration date from Jan-2017 onwards'', which was obtained from the Singapore Government Office website (https://beta.data.gov.sg), was conducted. Then, four supervised machine learning models were created to predict the resale price, which were evaluated using four metrics and then, their performance was compared to declare the best one.

The thesis has been written based on the journals and international literature. Data analysis was carried out with a view to extract information on various aspects of the real estate market based on data visualization through charts such as, boxplot, scatterplot , bar chart and pie chart. At the same time, algorithms such as Ridge regression, Lasso regression, random forests and decision trees were chosen to create the models. They were evaluated using four evaluation metrics such as, mean square error, root mean square error, root mean square error, mean absolute error and R-squared metric.

The utilized database and the code written entirely in Python can be found posted on GitHub (https://github.com/PapBill) and Kaggle (https://www.kaggle.com/vasilispapazisis)

Keywords: machine learning, supervised learning, python, data analysis, data science.

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΕΙΚΟΝΩΝ

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

1.Εισαγωγή

1.1 Πρόβλημα

Η επιστήμη δεδομένων είναι ένας πολυδιάστατος κλάδος που συνδυάζει στοιχεία πληροφορικής, στατιστικής και μαθηματικών για την εξόρυξη πολύτιμων πληροφοριών που συντελούν στην λήψη αποφάσεων. Ως επί το πλείστων, η επιστήμη δεδομένων ασχολείται με την ανάλυση συνόλων δεδομένων , με σκοπό την αναγνώριση μοτίβων , την εξαγωγή πληροφοριών και την δημιουργία προβλέψεων. Πρόκειται για έναν τομέα άρρηκτα συνδεδεμένο με την μηχανική μάθηση, συνδυασμός, που φέρει καινοτόμους τρόπους αξιοποίησης και αντιμετώπισης των δεδομένων. Η μηχανική μάθηση είναι ένα υποσύνολο της τεχνητής νοημοσύνης, βασιζόμενη σε αλγορίθμους που εκπαιδεύονται από τα δεδομένα εισόδου , βελτιώνονται χωρίς να απαιτούνται ρητές προγραμματιστικές οδηγίες και εξάγουν τιμές που ανήκουν σε ένα συγκεκριμένο εύρος. Χάρις στους αλγορίθμους μηχανικής μάθησης, οι διαδικασίες λήψης αποφάσεων και προβλέψεων αυτοματοποιούνται με βάση τα σύνολα δεδομένων που εισάγονται κάθε φορά.

Η μηχανική μάθηση συνδράμει σημαντικά σε πολλούς τομείς, από την εκπαίδευση και την υγειονομική περίθαλψη, έως τις μεταφορές και τις χρηματοοικονομικές υπηρεσίες. Η αγορά ακινήτων, αποτελεί έναν από αυτούς τους τομείς, ο οποίος λαμβάνει μεγάλη δυναμική τα τελευταία χρόνια. Η εφαρμογή μηχανικής μάθησης, προσφέρει πολλές δυνατότητες στην αγορά ακινήτων που βασίζονται στην ανάλυση δεδομένων και στον εντοπισμό μοτίβων που συντελούν στην αναγνώριση τάσεων της αγοράς.

Για την βαθύτερη κατανόηση του συνόλου δεδομένων της βάσης Resale flat prices based on registration date, μελετήθηκαν αρκετά notebooks από χρήστες της ιστοσελίδας του Kaggle που επιλύσαν το ίδιο ή και παρόμοια προβλήματα. Η μελέτη περιλάμβανε την παρατήρηση των αλγορίθμων παλινδρόμησης και την σύγκριση των αποτελεσμάτων. Στην συνέχεια, ξεκίνησε η συγγραφή του κώδικα της παρούσας πτυχιακής εργασίας.

Αρχικά, πραγματοποιήθηκε στατιστική ανάλυση δεδομένων της βάσης, η οποία οπτικοποιήθηκε με την χρήση διαγραμμάτων, με γνώμονα την αναγνώριση προτύπων και τάσεων της αγοράς. Έπειτα , έγινε επεξεργασία των μεταβλητών προκειμένου να μπορέσουν να εφαρμοστούν στα μοντέλα παλινδρόμησης που επιλέχθηκαν. Τέλος, συγκρίνονται τα αποτελέσματα που προέκυψαν από τα επιβλεπόμενα μοντέλα μηχανικής μάθησης.

1.2 Στόχος

Το μεγαλύτερο μέρος της αγοράς ακινήτων της Σιγκαπούρης (82%) αποτελείται από δημόσιες κατοικίες ρυθμιζόμενες από το Συμβούλιο Στέγασης και Ανάπτυξης που θεσπίζει αυστηρά μέτρα στοχεύοντας στην σταθεροποίηση της αγοράς. Η υψηλή ποιότητα ζωής , η ασφάλεια και η ευνοϊκή φορολογία προσελκύουν πολλούς ξένους και μη, αγοραστές που αναζητούν ένα διαμέρισμα είτε για να μείνουν οι ίδιοι με τις οικογένειές τους, είτε να το εκμεταλλευτούν ως εισοδηματική πηγή.

Η παρούσα πτυχιακή εργασία έχει δύο σκοπούς. Πρώτος στόχος είναι η παροχή πολύτιμων πληροφοριών που απορρέουν από την ανάλυση των δεδομένων της βάσης, ενώ δεύτερος στόχος είναι η κατασκευή τεσσάρων μοντέλων μηχανικής μάθησης.

Η στατιστική ανάλυση των δεδομένων της βάσης αποσκοπεί στην βελτιστοποίηση της πολιτικής στέγασης καθώς κατανοούνται οι ανάγκες και οι προτιμήσεις των κατοίκων , βοηθώντας παράλληλα στον σχεδιασμό νέων κατοικιών που βασίζονται στα χαρακτηριστικά που επιθυμούν οι πολίτες. Επίσης, εντοπίζονται μοτίβα στην τάση ζήτησης με συνέπεια την πρόβλεψη μελλοντικών αναγκών.

Ο δεύτερος στόχος, αφορά την δημιουργία επιβλεπόμενων μοντέλων μηχανικής μάθησης που προβλέπουν την τιμή πώλησης των διαμερισμάτων. Στα μοντέλα αυτά εφαρμόζονται μετρικές αξιολόγησης για την κατανόηση της απόδοσης και της αποτελεσματικότητάς τους και πραγματοποιείται βελτιστοποίηση των παραμέτρων τους.

1.3 Περιεχόμενα

Το κεφάλαιο 2 , αφορά το θεωρητικό υπόβαθρο στο οποίο γίνεται μια εισαγωγή στην Μηχανική Μάθηση, στους αλγορίθμους και τις μετρικές αξιολόγησης που θα χρησιμοποιηθούν. Η αναφορά βασίζεται σε βιβλιογραφία.

Στο κεφάλαιο 3 , γίνεται βιβλιογραφική ανασκόπηση σχετικά με την Σιγκαπούρη. Παρέχονται πληροφορίες για το Συμβούλιο Στέγασης και Ανάπτυξης και τον τρόπο με τον οποίο διαχειρίζεται τα διαμερίσματα.

Το κεφάλαιο 4, αφορά την μεθοδολογία της εργασίας. Αναφέρονται πληροφορίες σχετικά με την βάση δεδομένων, τις μετρικές και τα εργαλεία που χρησιμοποιήθηκαν.

Στο κεφάλαιο 5, γίνεται η στατιστική ανάλυση της βάσης , όπου συλλέγονται, επεξεργάζονται και ερμηνεύονται τα δεδομένα. Ακολουθεί η μοντελοποίηση και η αξιολόγησή τους με βάση τους αλγορίθμους και τις μετρικές που επιλέχθηκαν. Οι αλγόριθμοι είναι η παλινδρόμηση Ridge, παλινδρόμηση Lasso , τυχαία δάση και δένδρα απόφασης. Τέλος, έγινε η σύγκριση των αποτελεσμάτων.

Στο κεφάλαιο 6 εξάγονται συμπεράσματα της ανάλυσης δεδομένων και της απόδοσης των τεσσάρων μοντέλων.

2. Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

Μηχανική μάθηση (Machine Learning), ορίζεται η τεχνική που δίνει την δυνατότητα στους υπολογιστές να μαθαίνουν από τα δεδομένα χωρίς να είναι άμεσα προγραμματισμένοι να το κάνουν. Διερευνά την δημιουργία αλγορίθμων , οι οποίοι μπορούν να μαθαίνουν από τα δεδομένα που εισάγονται και να κάνουν προβλέψεις σχετικά με αυτά. Ο Tom M. Mitchell το 1997, ορίζει τη μηχανική μάθηση ως «ένα πρόγραμμα υπολογιστών λέγεται ότι μαθαίνει από την εμπειρία *E*, σε σχέση με κάποια τάξη εργασιών *T* και μέτρηση απόδοσης *P* (Performance Measure) εάν η απόδοσή του σε εργασίες στο *T*, όπως μετριέται από το *P*, βελτιώνεται με την εμπειρία *E*.»

Επεξηγηματικά, η μηχανική μάθηση μπορεί να θεωρηθεί ως ένα σύνολο μεθόδων που βασίζονται σε διάφορους αλγορίθμους, με σκοπό την αυτοματοποιημένη αναγνώριση μοτίβων που βρίσκονται στα δεδομένα και έπειτα με βάση τα μοτίβα αυτά να πάρουν αποφάσεις που υπόκεινται σε περιορισμούς ή να προβλέψουν μελλοντικά αποτελέσματα.

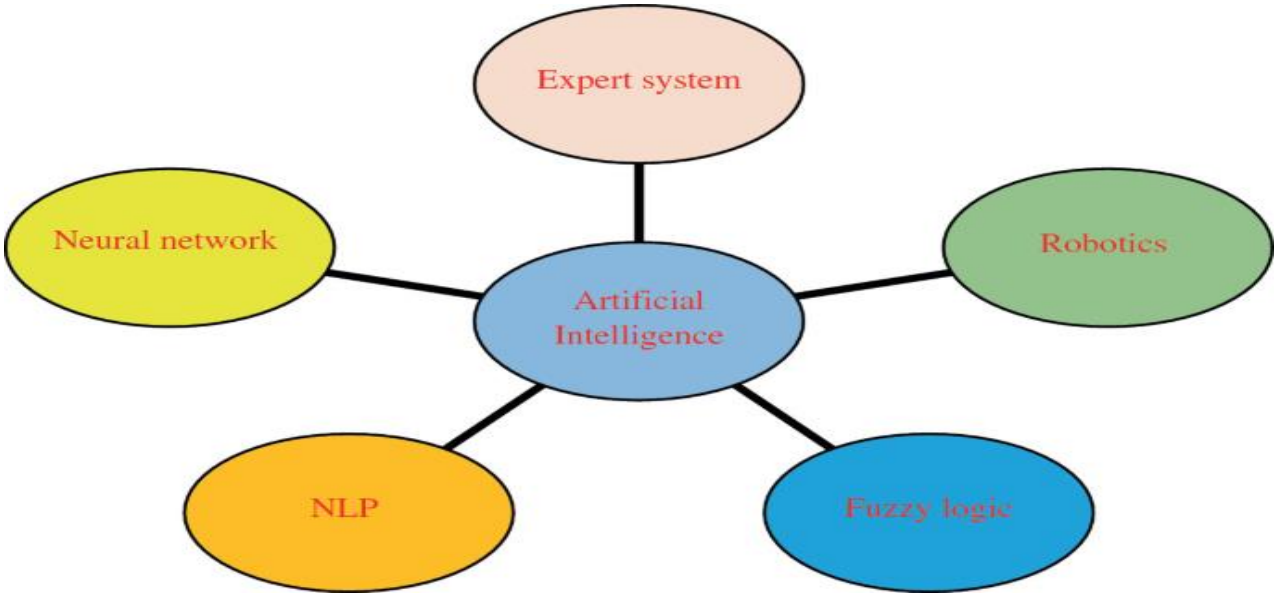
Η μηχανική μάθηση αποτελεί υποσύνολο της τεχνητής νοημοσύνης. Με τον όρο Τεχνητή Νοημοσύνη (Artificial Intelligence), ορίζεται η ικανότητα μιας μηχανής να αναπαράγει τις γνωστικές λειτουργίες ενός ανθρώπου , όπως είναι η μάθηση, ο σχεδιασμός και η δημιουργικότητα. Το πεδίο της τεχνητής νοημοσύνης αποτελείται συνολικά από πέντε κλάδους:

- Νευρωνικά Δίκτυα (Neural Networks), είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που διδάσκει στους υπολογιστές να επεξεργάζονται δεδομένα με τρόπο. εμπνευσμένο από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου.

- Ρομποτική (Robotics), έχει ως αντικείμενο τη μελέτη των μηχανών και ασχολείται με το σχεδιασμό, την κατασκευή, τη λειτουργία και τη χρήση ρομπότ.

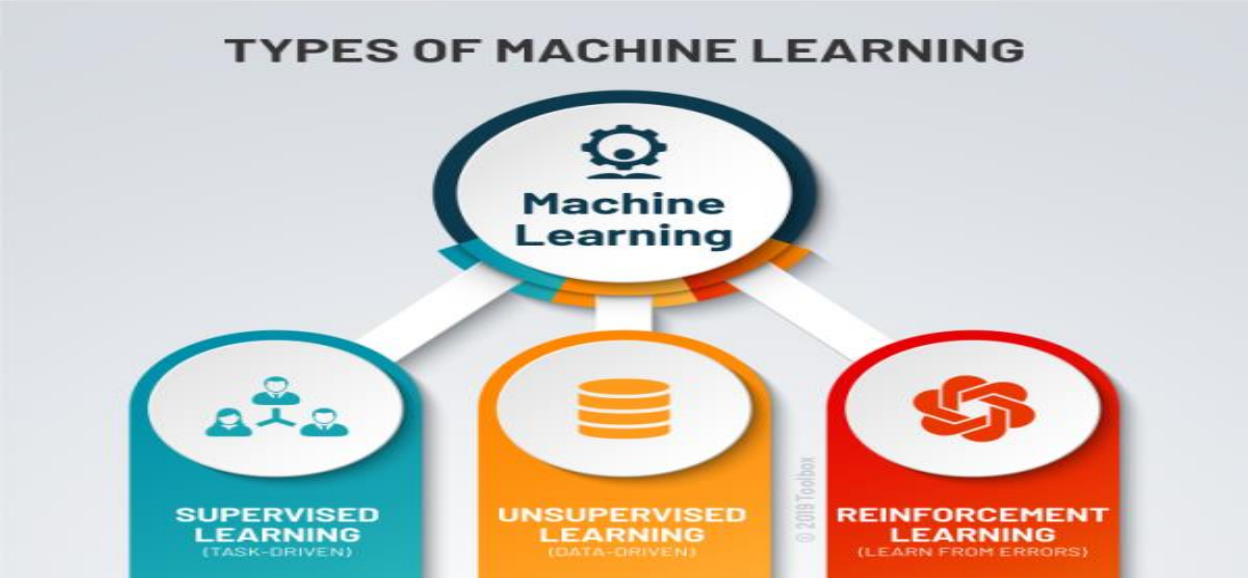
- Έμπειρα Συστήματα (Expert Systems) είναι υπολογιστικά συστήματα σχεδιασμένα να επιλύουν πολύπλοκα προβλήματα, προσομοιώνοντας την κρίση και την συμπεριφορά ενός ανθρώπου που διαθέτει τεχνογνωσία και εμπειρία σε έναν συγκεκριμένο τομέα.

- Ασαφής Λογική (Fuzzy Logic), είναι η τεχνική κατά την οποία επεξεργάζονται αβέβαιες πληροφορίες , βασιζόμενη σε "βαθμούς αλήθειας" και όχι στη συνήθη λογική "αληθές ή ψευδές" (1 ή 0) στην οποία βασίζεται ο σύγχρονος υπολογιστής.



Εικόνα 1 : Κλάδοι της Τεχνητής Νοημοσύνης(What Is Artificial Intelligence: Definition Of AI, 2023)

Η μηχανική μάθηση ανάλογα με το πρόβλημα που πρέπει να επιλύσει, χρησιμοποιεί διάφορες αλγοριθμικές τεχνικές. Χωρίζεται σε τρείς κατηγορίες ανάλογα με την φύση των δεδομένων και το επιδιωκόμενο αποτέλεσμα. Οι κατηγορίες αυτές είναι : Η Επιβλεπόμενη Μάθηση (supervised machine learning), η Μη-Επιβλεπόμενη Μάθηση (unsupervised machine learning) και η Ενισχυτική Μάθηση (reinforcement learning).[1], [2]

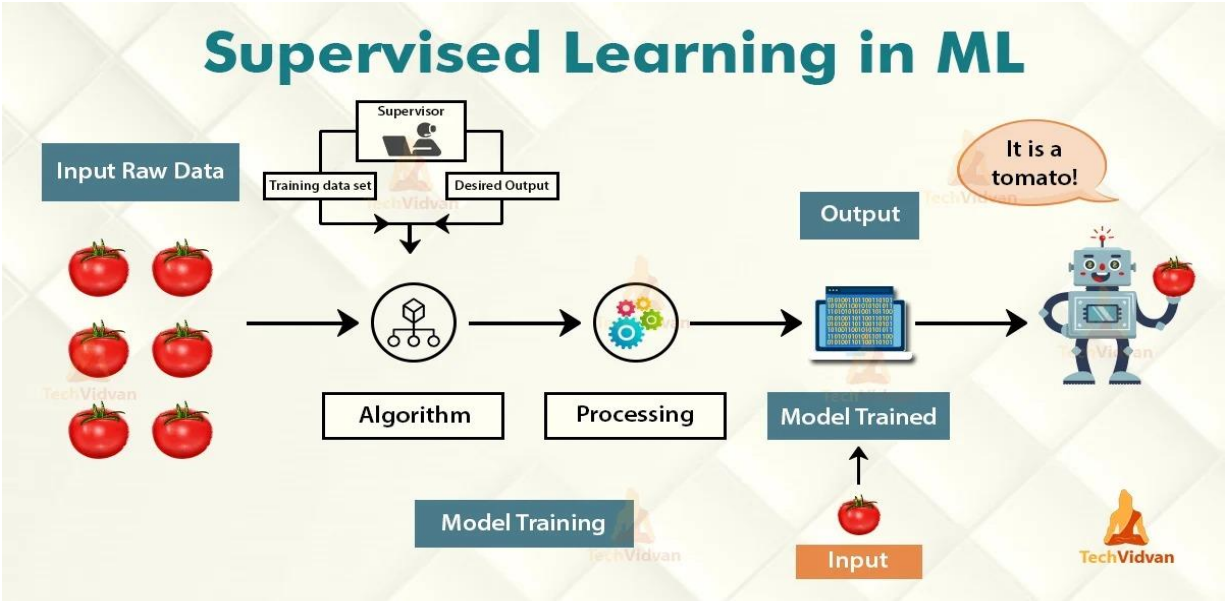


Εικόνα 2 : Τύποι Μηχανικής Μάθησης (What is Machine Learning - Potentia Analytics)

2.1.1 Επιβλεπόμενη Μάθηση

Η επιβλεπόμενη μάθηση είναι ένα από τα συνηθέστερα και πιο επιτυχημένα μοντέλα μηχανικής μάθησης, το οποίο αξιοποιείται για την πρόβλεψη συγκεκριμένων αποτελεσμάτων. Τα επιτηρούμενα μοντέλα μάθησης απαρτίζονται από ζεύγη δεδομένων "εισόδου" και "εξόδου", όπου στην έξοδο το μοντέλο πραγματοποιεί την πρόβλεψη της επιθυμητής τιμής. Σε αυτό το είδος μάθησης ο αλγόριθμος, βασισμένος στα δεδομένα εισόδου, εκπαιδεύει τον εαυτό του και δημιουργεί ένα νέο σύστημα το οποίο θα εφαρμοστεί στο νέο σύνολο δεδομένων . Η διαδικασία αυτή συνεχίζεται αναδρομικά μέχρι να πραγματοποιηθεί η πρόβλεψη της επιθυμητής τιμής. Με την πάροδο του χρόνου έχει αναπτυχθεί ένα μεγάλο πλήθος αλγορίθμων επιβλεπόμενης μάθησης ,από το οποίο πιο συχνά χρησιμοποιούνται :

- Οι Μηχανές Διανυσμάτων στήριξης (Support Vector Machine, SVM)
- Η Γραμμική Παλινδρόμηση (Linear Regression)
- Λογιστική Παλινδρόμηση(Logistic Regression)
- Τα τυχαία δάση (Random Forests)
- Τα δένδρα απόφασης(Decisision Trees)
- Ο Αλγόριθμος Naive Bayes

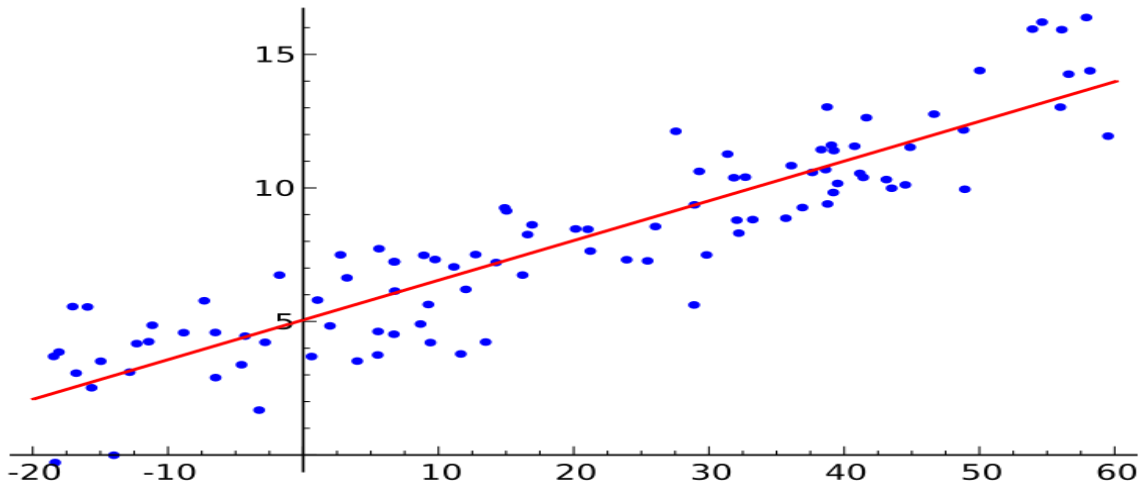


Εικόνα 3 : Λειτουργία επιβλεπόμενης μάθησης (TechVidvan,2023).

Η παραπάνω εικόνα αποτελεί παράδειγμα ενός μοντέλου μάθησης με επίβλεψη. Αναλυτικότερα, στην φάση της εκπαίδευσης (Model Training), στον αλγόριθμο (Algorithm) εισάγεται το σύνολο δεδομένων εκπαίδευσης (Training Data Set) έχοντας κάποια ετικέτα που το προσδιορίζει (Desired Output). Έπειτα, αφού ολοκληρωθεί η εκπαίδευση, εισάγονται νέα δεδομένα στο μοντέλο πρόβλεψης (Model Trained) ,το οποίο με την σειρά του παράγει το αποτέλεσμα πρόβλεψης (Output).

Τα επιτηρούμενα μοντέλα μάθησης εστιάζουν σε δύο τύπους προβλημάτων , το πρόβλημα της Παλινδρόμησης (Regression) και το πρόβλημα της Ταξινόμησης (Classification)[3] .

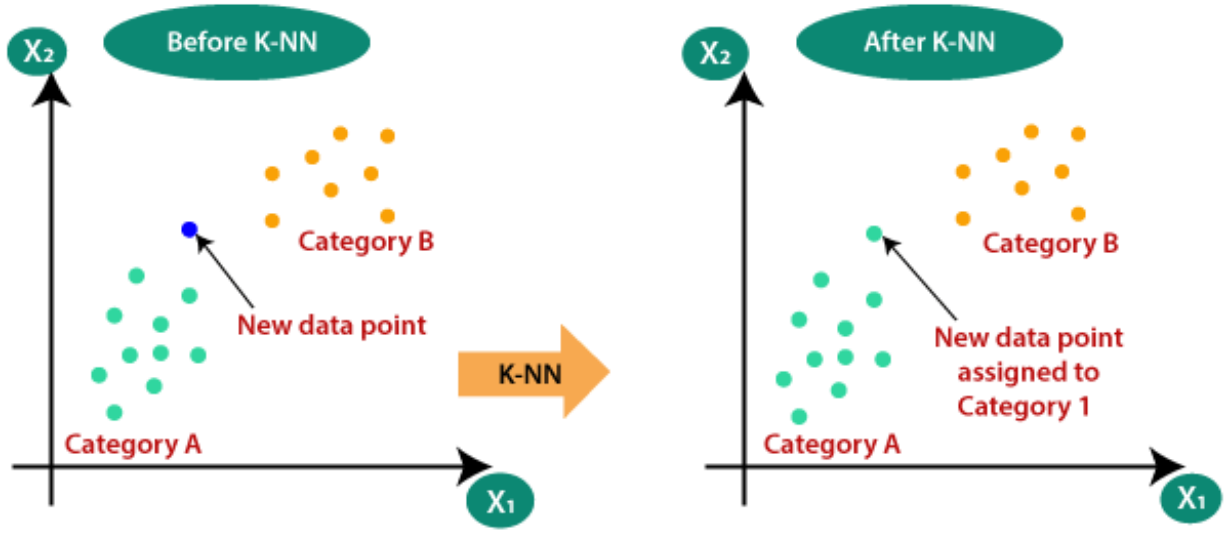
- Παλινδρόμηση :Πρώτος τύπος επιβλεπόμενης μάθησης είναι η μέθοδος της παλινδρόμησης. Η παλινδρόμηση μας βοηθά να εντοπίσουμε μοτίβα μεταξύ των δεδομένων. Με την βοήθεια γραφικών παραστάσεων μπορούμε να κατανοήσουμε την σχέση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών. Η παλινδρόμηση έχει ένα ευρύ φάσμα εφαρμογών στην καθημερινότητα, σε προβλήματα που περιλαμβάνουν συνεχείς αριθμούς όπως : Πρόβλεψη τιμών κατοικιών, μετοχών ή πρόβλεψη ετήσιου εισοδήματος ενός ατόμου με βάση τα χρόνια προϋπηρεσίας και τον τόπο διαμονής. Υπάρχουν πολλοί τύποι παλινδρόμησης, ωστόσο οι δύο κυριότεροι είναι η γραμμική παλινδρόμηση και η λογιστική παλινδρόμηση.



Εικόνα 4 : Διάγραμμα διασποράς Γραμμικής Παλινδρόμησης(Wikipedia).

Η παραπάνω εικόνα αποτελεί παράδειγμα γραμμικής παλινδρόμησης. Έχουμε ένα σύνολο δειγμάτων $\{\chi_i , y_i\}$, με τις μεταβλητές x να μην θεωρούνται τυχαίες, σε αντίθεση όμως οι y θεωρούνται τυχαίες . Η σχέση μεταξύ των δύο μεταβλητών περιγράφεται από το ευθύγραμμο τμήμα, με τύπο $y = ax + \beta$.

- Ταξινόμηση :Η ταξινόμηση αποτελεί ένα, βαθιά μελετημένο μοντέλο το οποίο χρησιμοποιείται όταν η επιθυμητή τιμή είναι διακριτή. Αποτελείται από δύο στάδια, την εκπαίδευση και την ταξινόμηση . Αρχικά, το σύνολο δεδομένων χωρίζεται σε δεδομένα εκπαίδευσης τα οποία αποτελούνται από εγγραφές με γνωστή ετικέτα κατηγορίας και δεδομένα δοκιμής τα οποία δεν φέρουν ετικέτα. Μέσα από ένα σύνολο κανόνων και από τα δεδομένα εκπαίδευσης, δημιουργείται ένα νέο μοντέλο ταξινόμησης , το οποίο στη συνέχεια αξιολογείται στο σύνολο δεδομένων δοκιμής προτού χρησιμοποιηθεί για την εκτέλεση πρόβλεψης σε νέα δεδομένα. Πρόκειται δηλαδή για μία μέθοδο που ταξινομεί δεδομένα σε διάφορες κλάσεις , καθώς την επόμενη φορά που θα εισαχθούν δεδομένα ,το μοντέλο θα τα συγκρίνει και θα τα ταξινομήσει στην αντίστοιχη κατηγορία[4]. Για παράδειγμα , ένα σύστημα μπορεί να εκπαιδευτεί στο να ανιχνεύει αποδοτικά αν κάποιο άτομο έχει COVID-19. Μερικοί από τους αλγόριθμους ταξινόμησης είναι : Τα τυχαία δάση, ο αλγόριθμος Naive Bayes , οι μηχανές διανυσμάτων υποστήριξης , τα δένδρα απόφασης, ο K-πλησιέστεροι γείτονες.



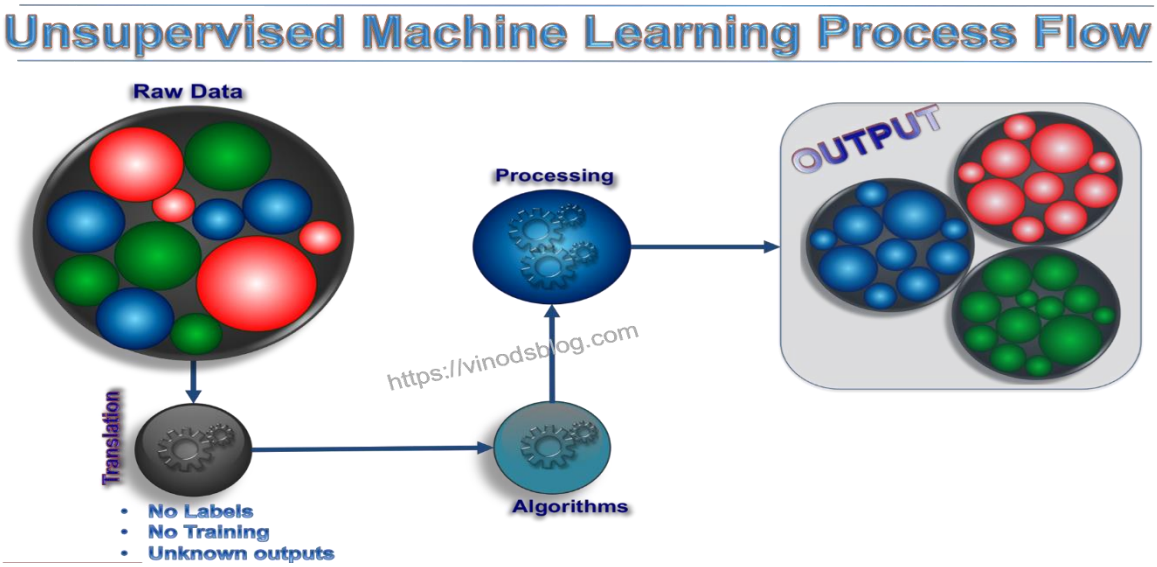
Εικόνα 5 : Αλγόριθμος Κ-πλησιέστεροι γείτονες (JavaPoint.com)

Στο παράδειγμα της εικόνας ο αλγόριθμος ταξινομεί τα νέα δεδομένα σε δύο κατηγορίες (Α και Β), με βάση τους *Κ*-πλησιέστερους γείτονες. Οι *Κ*-γείτονες επιλέγονται σύμφωνα με την μικρότερη ευκλείδεια απόσταση από το νέο δεδομένο. Το νέο αντικείμενο θα ανατεθεί σε εκείνη την κατηγορία για την οποία ο αριθμός των *Κ*-πλησιέστερων γειτόνων είναι μέγιστος.

2.1.2 Μη-Επιβλεπόμενη Μάθηση

Δεύτερος τύπος μηχανικής μάθησης είναι η μη επιβλεπόμενη μάθηση. Σε αντίθεση με την επιβλεπόμενη μάθηση , στην οποία προσφέρονται στο σύστημα δεδομένα με ετικέτες, στην μη επιβλεπόμενη μάθηση εισάγονται δεδομένα και εναπόκειται στο σύστημα να εξαγάγει γνώση από αυτά. Αναλυτικότερα, στα μη επιβλεπόμενα μοντέλα μάθησης , οι αλγόριθμοι εκπαιδεύονται στο να δέχονται μη επισημασμένα ή ταξινομημένα σύνολα δεδομένων με σκοπό τον εντοπισμό μοτίβων και συσχετίσεων προκειμένου να αποκτήσουν την ικανότητα να βαθμολογήσουν τα δεδομένα σύμφωνα με τους κανόνες αυτούς . Η “εμπειρία” του μοντέλου ορίζεται από την ποσότητα των δεδομένων που εισάγονται και καθίστανται διαθέσιμα ,με συνέπεια να απαιτούνται μεγάλα σύνολα για την δημιουργία ενός αποδοτικού συστήματος[1]. Ένα δείγμα των πιο διαδεδομένων αλγόριθμων μη επιβλεπόμενης μάθησης είναι :

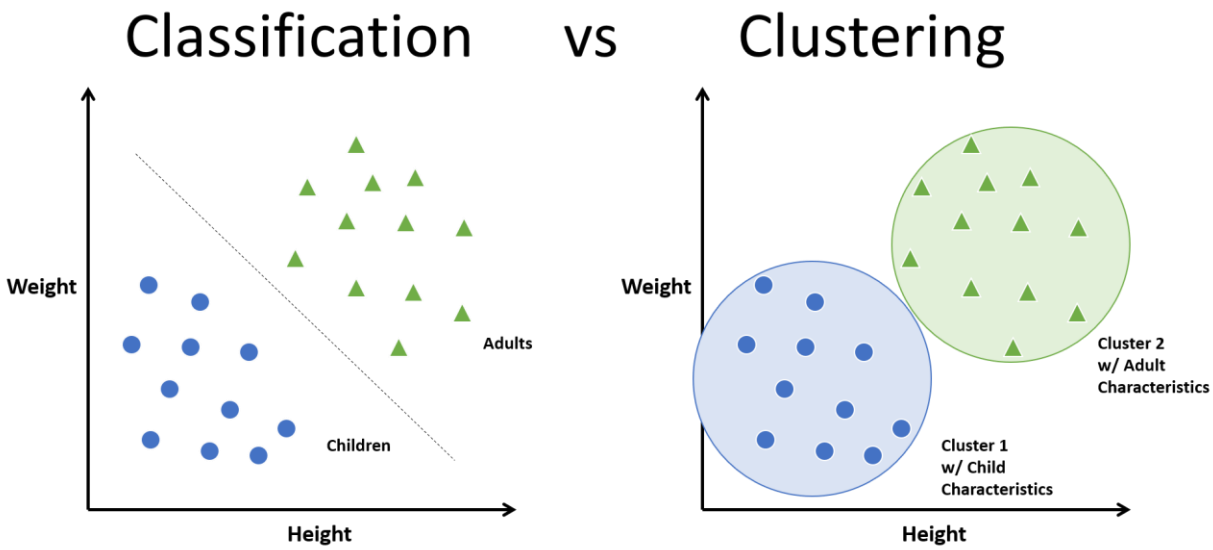
- Η Τυποποίηση (Standardization)
- Η Κανονικοποίηση (Normalization)
- Η Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)
- Ο Ομαδοποιητής K-Means
- Αλγόριθμος Apriori



Εικόνα 6 : Διάγραμμα ροής μη επιβλεπόμενης μάθησης (Data Science Central , 2016)

Στην παραπάνω εικόνα περιγράφεται η λειτουργία της επιβλεπόμενης μάθησης. Αναλυτικότερα, υπάρχουν τα δεδομένα που αρχικά εισάγουμε και τα τρία στάδια του μη επιβλεπόμενου μοντέλου. Στο πρώτο στάδιο (Translation) το σύστημα μας αναλύει χωρίς βοήθεια τα δεδομένα, βρίσκοντας συσχετίσεις και μοτίβα , δημιουργώντας έτσι το μοντέλο κατά το οποίο θα κατηγοριοποιηθούν τα δεδομένα(Algorithm). Στο τέλος προκύπτει η έξοδος, η οποία προήλθε από την επεξεργασία των δεδομένων.
Διάφορες τεχνικές για την μάθηση χωρίς επίβλεψη έχουν παρουσιαστεί τα τελευταία χρόνια. Δημοφιλέστερες αποτελούν η συσταδοποίηση (Clustering) , ανίχνευση ανωμαλιών (Anomaly Detection), κανόνες συσχετίσεων (Association Rules), μείωση διάστασης (Dimencionality Reduction).[1], [3]

- Συσταδοποίηση : Πρόκειται για μια τεχνική που οργανώνει τα δεδομένα σε ομάδες έτσι ώστε κάθε πληροφορία να ανήκει σε μόνα μία συστάδα[5]. Χωρίς να δίνονται ετικέτες στα δεδομένα , ο αλγόριθμος τα ομαδοποιεί βάσει των ομοιοτήτων και των διαφορών που απορρέουν από το χαρακτηριστικά τους. Αυτό έχει ως αποτέλεσμα την εύρεση συστάδων αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε συστάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων συστάδων. Η αποδοτικότητα μιας μεθόδου συσταδοποίησης εξαρτάται από το πόσο μεγάλη είναι η ομοιότητα εντός της συστάδας και πόσο μικρή είναι η ομοιότητα ανάμεσα στις συστάδες.[6]



Εικόνα 7 : Διαφορά ταξινόμησης και Συσταδοποίησης (Analytics Vidhya ,2022)

Παραπάνω απεικονίζεται η διαφορά μεταξύ ταξινόμησης και συσταδοποίησης. Στην ταξινόμηση το πλήθος των τάξεων είναι γνωστό και απαιτείται η εισαγωγή δεδομένων προκειμένου το σύστημα να τα αξιοποιήσει για την ταξινόμηση μελλοντικών αντικειμένων στις ήδη καθορισμένες τάξεις. Αντίθετα , στην συσταδοποίηση ο αριθμός των κλάσεων δεν είναι γνωστός και δεν απαιτούνται δεδομένα εκπαίδευσης καθώς, σκοπός της είναι η οργάνωση μιας συλλογής από στοιχεία σε συστάδες με βάση κάποιο μέτρο ομοιότητας. [7]

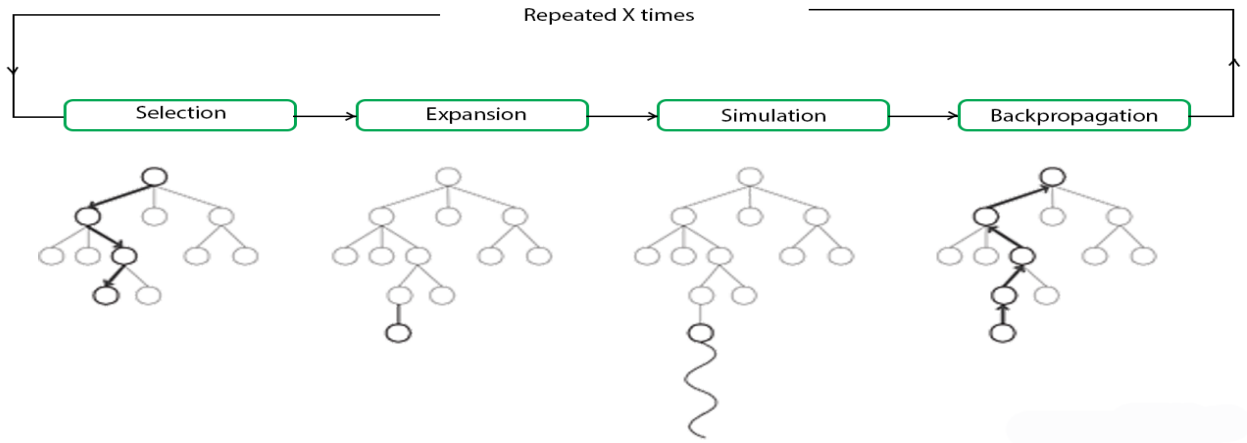
- Ανίχνευση ανωμαλιών : Η ανίχνευση σφαλμάτων είναι μία σημαντική τεχνική για την εύρεση και τον εντοπισμό ακραίων τιμών συμβάλλοντας στην πρόληψη απάτης ή εισβολής στο δίκτυο.

- Μείωση διάστασης : Σκοπός αυτής της μεθόδου είναι η αφαίρεση περιορισμών που εφαρμόζονται σε ένα σύστημα καθώς το υπολογιστικό κόστος είναι πολύ υψηλό.

- Κανόνες συσχετίσεων : Αυτή η τεχνική μη επιβλεπόμενης μάθησης χρησιμοποιείται για να εντοπίσει συνδέσεις μεταξύ διαφόρων συνόλων δεδομένων. Οι κανόνες συσχέτισης ως τεχνική μάθησης είναι εξαιρετικά αποδοτικοί ,διότι μπορούν να ανταπεξέλθουν σε μεγάλες βάσεις δεδομένων ή σε βάσεις δεδομένων με “θόρυβο“ ,προβλήματα που καθιστούν παραδοσιακές τεχνικές μάθησης λιγότερο αποδοτικές καθώς η εκπαίδευσή τους πρέπει να γίνεται σε “καθαρά” και όχι τόσο σύνθετα δεδομένα . [8]

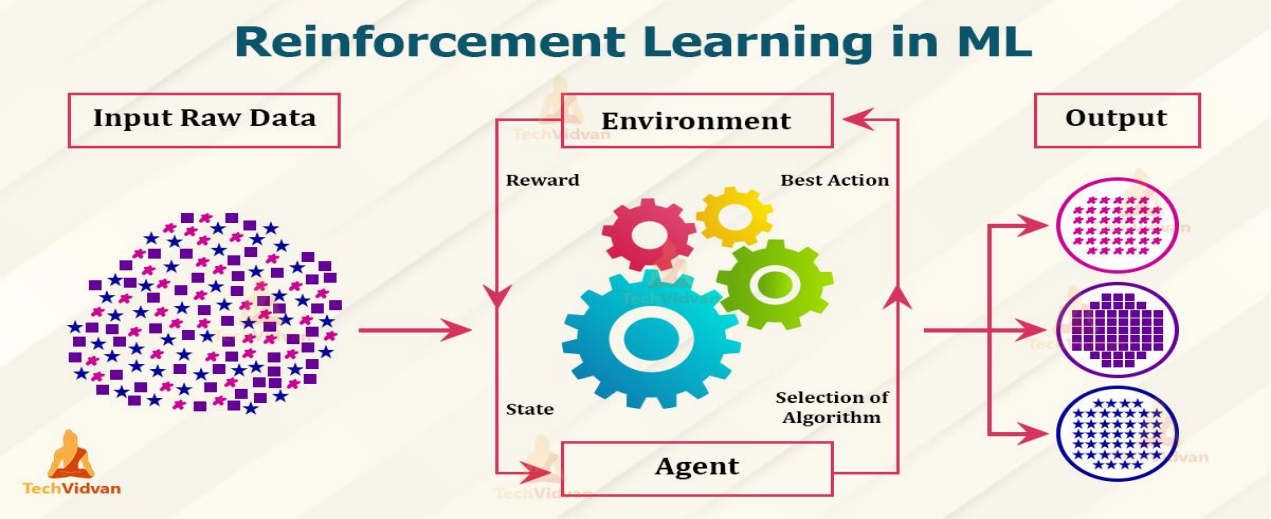
2.1.3 Ενισχυτική Μάθηση

Το τρίτο μοντέλο μηχανικής μάθησης είναι η ενισχυμένη μάθηση. Σε αυτή την περίπτωση το μοντέλο εκπαιδεύεται να αλληλοεπιδρά σε ένα περιβάλλον αβέβαιο. Δεν περιλαμβάνεται κάποιο κλειδί απάντησης ,παρά μόνο ένα σύνολο από επιτρεπόμενες ενέργειες με αποτέλεσμα κάθε φορά που ο αλγόριθμος πραγματοποιεί μια σωστή επιλογή , δέχεται μία ανταμοιβή από το περιβάλλον, της οποίας η τιμή είναι ανάλογη με το πόσο αυτή η ενέργεια συνέβαλε στην επίτευξη του στόχου[9]. Έτσι λοιπόν , μέσα από μία διαδικασία δοκιμής και σφάλματος ο αλγόριθμος μπορεί να μάθει να επιλέγει βέλτιστες ενέργειες χωρίς να έχει πλήρη επίγνωση του περιβάλλοντος. Παράδειγμα αποτελούν τα επιτραπέζια παιχνίδια όπως το σκάκι όπου ο παίχτης τοποθετεί το πιόνια σε σημεία που θα του αυξήσουν τις πιθανότητες νίκης και αναπροσαρμόζεται στις κινήσεις του ανάλογα με τις στρατηγική που ακολουθεί ο αντίπαλος παίχτης. Μία πολύ διαδεδομένη τεχνική ενισχυμένης μάθησης είναι ο αλγόριθμος Monte Carlo Tree Search (MCTS).



Εικόνα 8 : Αλγόριθμος Monte Carlo Search Tree (geeksforgeeks, 2023)

Η παραπάνω εικόνα δείχνει ένα στιγμιότυπο του αλγορίθμου MCTS. Ο ευρετικός αλγόριθμος MCTS χρησιμοποιείται σε μεγάλο βαθμό στον τομέα της λήψης αποφάσεων καθώς είναι ιδιαίτερα αποδοτικός σε μεγάλες περιοχές αναζήτησης ,στις οποίες παραδοσιακοί αλγόριθμοι δυσκολεύονται επειδή βασίζονται στην εξαντλητική εξερεύνησή της. Αντίθετα ο MCTS εξερευνά μόνο υποσχόμενους κλάδους, γεγονός που περιορίζει σε μεγάλο βαθμό την περιοχή αναζήτησης. [9]



Εικόνα 9 : Λειτουργία Επιβλεπόμενης μάθησης (pinterest, 2020)

Στην παραπάνω εικόνα περιγράφεται η διαδικασία της ενισχυμένης μάθησης. Σύμφωνα με το σχήμα ο πράκτορας (agent) εκτελεί επαναληπτικά ενέργειες μέσα σε ένα περιβάλλον το οποίο περιέχει ένα σύνολο από καταστάσεις και ανταμείβεται(Reward) ανάλογα με την αποτελεσματικότητα της ενέργειάς του.

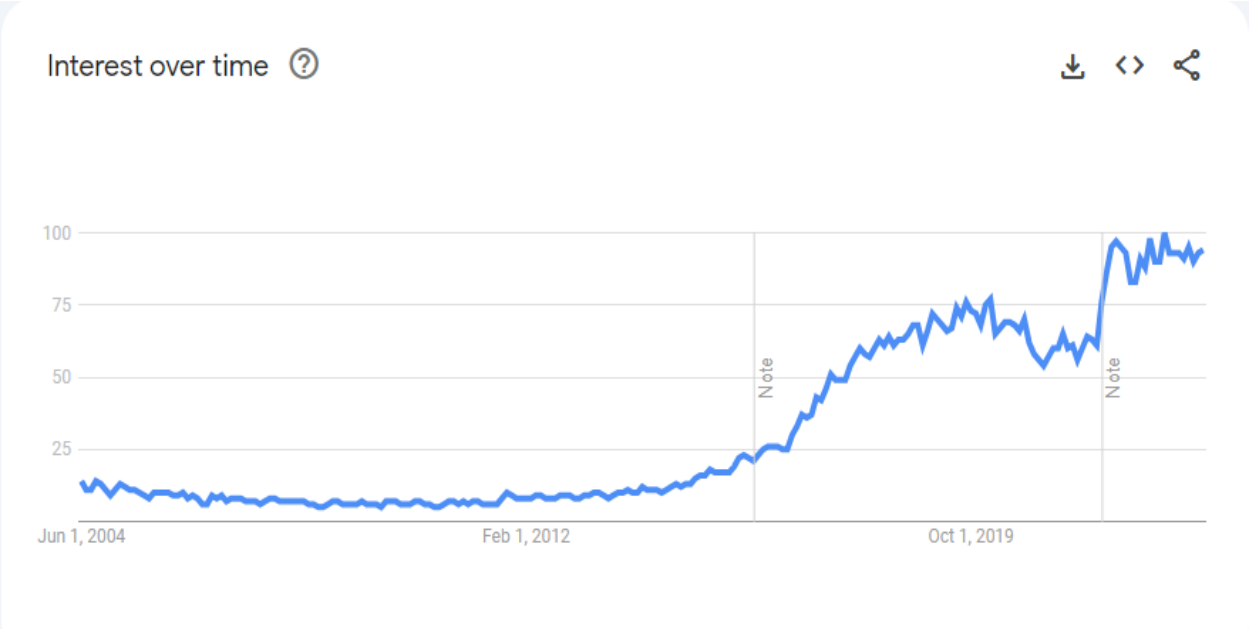
Παρακάτω ακολουθεί ένας συνοπτικός πίνακας με τις διαφορές μεταξύ των τριών μοντέλων μηχανικής μάθησης.

ΕΠΙΒΛΕΠΟΜΕΝΗ –	Η ΕΠΙΒΛΕΠΟΜΕΝΗ –	ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ
• Δέχεται δεδομένα με ετικέτες. • Κύριος στόχος είναι η πρόβλεψη. • Διαχωρισμός δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμής. • Τεχνικές : Παλινδρόμηση και κατηγοριοποίηση.	• Μη επισημασμένα δεδομένα εισόδου. • Κύριος στόχος είναι η κατανόηση/ανάλυση της βάσης δεδομένων • Το σύνολο δεδομένων δεν διαχωρίζεται • Τεχνικές : Συσταδοποίηση, Εντοπισμός Ανω/λιών, Μείωση Διάστασης ,Κανόνες Συσχετίσεων	• Απόφαση με βάση την εμπειρία που αποκτά από τις προηγούμενες ενέργειες. • Σύστημα ανταμοιβής • Τεχνικές : MCTS

Πίνακας 1: Συγκριτικός πίνακας

2.1.4 Προσεγγίσεις

Αναφορές προς την μηχανική μάθηση ξεκίνησαν να γίνονται στην δεκαετία του 1960, ωστόσο η χρήση των τεχνικών της εντείνεται έντονα την δεκαετία του 1990 ως απόρροια της ανάπτυξης κλάδων της επιστήμης υπολογιστών όπως είναι η εξόρυξη δεδομένων. Μάλιστα, τα τελευταία χρόνια η ραγδαία ανάπτυξη της τεχνολογίας αύξησε κατακόρυφα τον μέγεθος των δεδομένων που παράγονται, χρησιμοποιούνται και αποθηκεύονται κάθε χρόνο. Ως αποτέλεσμα , καθίσταται επιτακτική η ανάγκη δημιουργίας νέων τεχνικών ανάλυσης ώστε ο μεγάλος όγκος δεδομένων να διαχειρίζεται αποτελεσματικά. Αυτό οδήγησε στην κατακόρυφη αύξηση του ενδιαφέροντος για τις προσεγγίσεις που μπορεί να προσφέρει η μηχανική μάθηση.



Εικόνα 10 : Απήχηση της μηχανικής μάθησης ανά χρόνο (Google Trends).

Από την παραπάνω εικόνα φαίνεται ξεκάθαρα η κορύφωση του ενδιαφέροντος για την μηχανική μάθηση ειδικά τα τελευταία τρία χρόνια, αγγίζοντας βαθμό κοντά στο 100 παγκοσμίως. Οι αριθμοί αντιπροσωπεύουν το ενδιαφέρον αναζήτησης σε σχέση με το υψηλότερο σημείο του διαγράμματος για τη συγκεκριμένη περιοχή και χρονική στιγμή. Μια τιμή 100 είναι η μέγιστη δημοτικότητα για τον όρο. Μια τιμή 50 σημαίνει ότι ο όρος είναι κατά το ήμισυ δημοφιλής. Μια τιμή 0 σημαίνει ότι δεν υπήρχαν αρκετά δεδομένα για τον όρο αυτό.

Ένας από τους πολλούς τομείς που εφαρμόζεται η μηχανική μάθηση είναι η ιατρική . Η αξιοποίηση βάσεων δεδομένων που αφορούν ασθενείς αποδείχθηκε καρποφόρα καθώς προβλέπονται καταστάσεις με εγκεφαλικό επεισόδιο ,καταστάσεις της καρδιακής λειτουργίας του εμβρύου και έγκαιρος εντοπισμός καρκίνου του μαστού [10]. Παράλληλα σημαντική προσφορά σημειώνεται στον τομέα της εφοδιαστικής αλυσίδας από την βελτιστοποίηση της διαχείρισης αποθεμάτων, την ορατότητα σε πραγματικό χρόνο για τη βελτίωση της εμπειρίας του πελάτη, την προβλεπτική ανάλυση έως και την αναβάθμιση των διαδικασιών αποθήκευσης προϊόντων[2]. Στον οικονομικό κλάδο η μηχανική μάθηση αξιοποιείται για την βελτίωση της διαχείρισης κινδύνων στις χρηματοοικονομικές

υπηρεσίες , την βαθμολόγηση πιστωτικών μονάδων και τον εντοπισμό ύποπτων συναλλαγών[11].

2.2 Αλγόριθμοι Μηχανικής Μάθησης

Αλγόριθμος ονομάζεται μια σειρά συγκεκριμένων οδηγιών (βημάτων) που στοχεύουν στην λύση ενός προβλήματος. Στον τομέα της μηχανικής μάθησης, οι αλγόριθμοι επεξεργάζονται τα δεδομένα , εντοπίζουν μοτίβα με γνώμονα την λήψη αποφάσεων και βελτιώνουν την απόδοσή.

2.2.1 Παλινδρόμηση Ridge

Η παλινδρόμηση Ridge, αποτελεί μια μορφή γραμμικής παλινδρόμησης η οποία χρησιμοποιεί την τεχνική στατιστικής κανονικοποίησης για την μείωση των σφαλμάτων, με σκοπό την μεγαλύτερη ακρίβεια του μοντέλου. Ουσιαστικά προστίθεται μια παράμετρος ποινής, με μέγεθος ίσο με το τετράγωνο των συντελεστών, συμβάλλοντας στην αποφυγή υπερπροσαρμογής του μοντέλου.
Η συνάρτηση της παλινδρόμησης Ridge εκφράζεται ως εξής :

Minimize

$$\sum (y_i - \beta_0 - \sum \beta_j x_{ij})^2 + \lambda \sum \beta_j^2$$

Όπου:

- y_i είναι η τιμή της εξαρτημένης μεταβλητής που θέλουμε να προβλέψουμε.
- x_{ij} είναι η τιμή της ανεξάρτητης μεταβλητής j.
- λ είναι η υπερπαράμετρος της κανονικοποίησης που καθορίζει το μέγεθος της ποινής των συντελεστών.
- β_j είναι ο συντελεστής για την j-οστή μεταβλητή πρόβλεψης x_{ij}.

2.2.2 Παλινδρόμηση Lasso

Η παλινδρόμηση Lasso είναι ένας τύπος γραμμικής παλινδρόμησης που διαθέτει την τεχνική της κανονικοποίησης για να αυξήσει την αποδοτικότητα του μοντέλου. Παρόμοια με την παλινδρόμηση Lasso, υπολογίζει το τετραγωνισμένο άθροισμα των υπολειπόμενων τιμών. Ωστόσο, προσθέτει έναν όρο ποινής ,μεγέθους ίσο με την απόλυτη τιμή των συντελεστών, το οποίο συνοδεύεται από μια παράμετρο λ. Αυτό έχει ως αποτέλεσμα τον περιορισμό της τιμής των συντελεστών των ανεξάρτητων μεταβλητών.
Η συνάρτηση της παλινδρόμησης Ridge εκφράζεται ως εξής :

Minimize

$$\sum (y_i - \beta_0 - \sum \beta_j x_{ij})^2 + \lambda \sum |\beta_j|$$

Όπου:

- y_i είναι η τιμή της εξαρτημένης μεταβλητής που θέλουμε να προβλέψουμε.
- x_{ij} είναι η τιμή της ανεξάρτητης μεταβλητής j.
- λ είναι η υπερπαράμετρος της κανονικοποίησης που καθορίζει το μέγεθος της ποινής των συντελεστών.
- βj είναι ο συντελεστής για την j-οστή μεταβλητή πρόβλεψης x_{ij}.

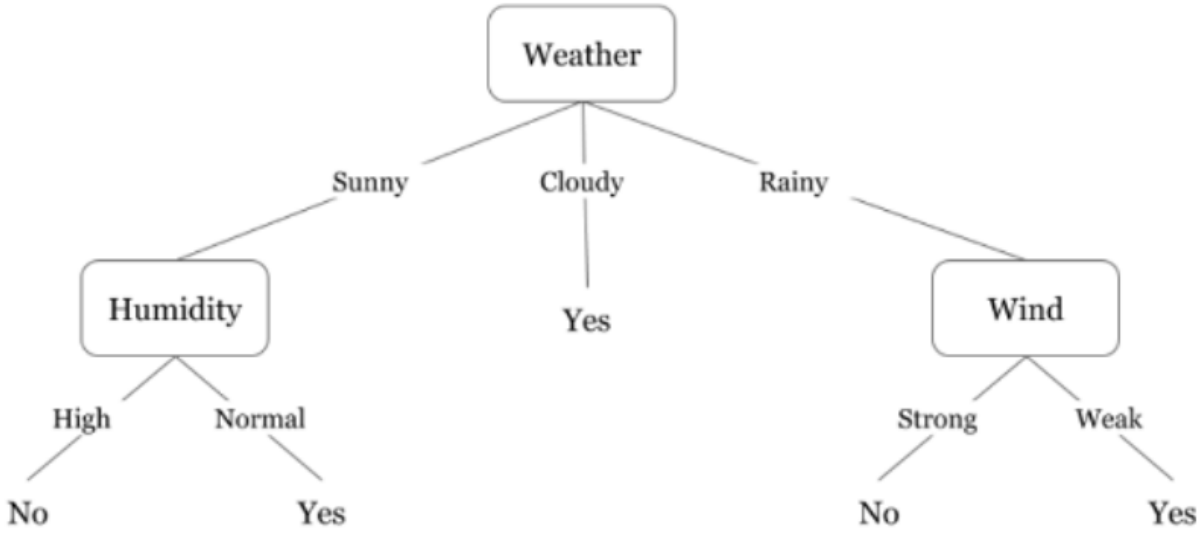
Η ειδοποιός διαφορά μεταξύ παλινδρομήσεων Lasso και Ridge εντοπίζεται στον τρόπο με τον οποίο μειώνουν την πολυπλοκότητα του μοντέλου. Η παλινδρόμηση Lasso μηδενίζει συντελεστές, με στόχο την εξάλειψη των ανεξαρτήτων μεταβλητών που συμμετέχουν σε μεγάλο βαθμό στην εξαγωγή αποτελέσματος. Αντίθετα, η παλινδρόμηση Ridge , μειώνει τους συντελεστές κοντά στο μηδέν ,χωρίς όμως αυτοί ποτέ να πάρουν την τιμή μηδέν. Έτσι ,αγγίζουν αμελητέες τιμές με συνέπεια την μικρότερη επιρροή του προγνωστικού παράγοντα (που συνοδεύουν) στο μοντέλο. Εν κατακλείδι, η παλινδρόμηση Lasso μειώνει τον αριθμό των ανεξάρτητων μεταβλητών που επηρεάζουν την έξοδο, ενώ η παλινδρόμηση Ridge μειώνει τη βαρύτητα που έχει κάθε ανεξάρτητη μεταβλητή στην έξοδο.

2.2.3 Δένδρα Απόφασης

Τα δέντρα απόφασης είναι ένας αλγόριθμος επιβλεπόμενης μάθησης που δεν διέπεται από παραμέτρους, χρησιμοποιείται τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Η δενδρική του δομή είναι αυστηρά ιεραρχημένη , παρόμοια με ένα διάγραμμα ροής στην οποία όλοι οι εσωτερικοί κόμβοι φέρουν μία ιδιότητα, κάθε κλάδος διέπεται από έναν κανόνα απόφασης της μορφής ‘’ εάν/αλλιώς (if/else)” και κάθε κόμβος φύλλου αντιπροσωπεύει το αποτέλεσμα. Η δομή ξεκινάει πάντα από την ρίζα, η οποία δεν έχει εισερχόμενους κλάδους. Οι εξερχόμενοι κλάδοι ενώνουν την ρίζα με τους κόμβους απόφασης που με την σειρά τους φέρουν διαφορετικούς εξερχόμενους κλάδους καταλήγοντας σε άλλους κόμβους. Η διαδικασία αυτή εκτελείται επαναληπτικά με αποτέλεσμα να διεξάγονται αξιολογήσεις που σχηματίζουν ομοιογενή υποσύνολα ,καταλήγοντας έτσι σε τερματικούς κόμβους που αντιπροσωπεύουν το σύνολο των πιθανών αποτελεσμάτων. Όσο μεγαλύτερο το βάθος του δένδρου, τόσο πολυπλοκότεροι είναι οι κανόνες, χωρίς αυτό να συνεπάγεται με μεγαλύτερη ακρίβεια του μοντέλου που δημιουργείται με βάση τον αλγόριθμο αυτό.

Για να καθιστεί κατανοητή η επεξήγηση του μοντέλου , η παρακάτω εικόνα αντικατοπτρίζει ένα παράδειγμα κατηγοριοποίησης μιας απόφασης που πρέπει να λάβει ένα παιδάκι, για να παίξει ποδόσφαιρο, με βάση τις καιρικές συνθήκες. Το παιδάκι έχει στην διάθεσή του δύο επιλογές : ΝΑΙ ή ΟΧΙ , οι οποίες εξαρτώνται από τις καιρικές συνθήκες που επικρατούν. Αρχικά, η ρίζα της δομής αντιπροσωπεύει τον καιρό (Weather), οποίος μπορεί να είναι βροχερός, συννεφιασμένος ή ηλιόλουστος (Rainy, Cloudy, Sunny). Έστω ότι επικρατεί ήλιος, και οδηγεί στον κόμβο του πρώτου επιπέδου του δένδρου που αντιπροσωπεύει την υγρασία (Humidity), η οποία με την σειρά της θα είναι είτε υψηλή, είτε

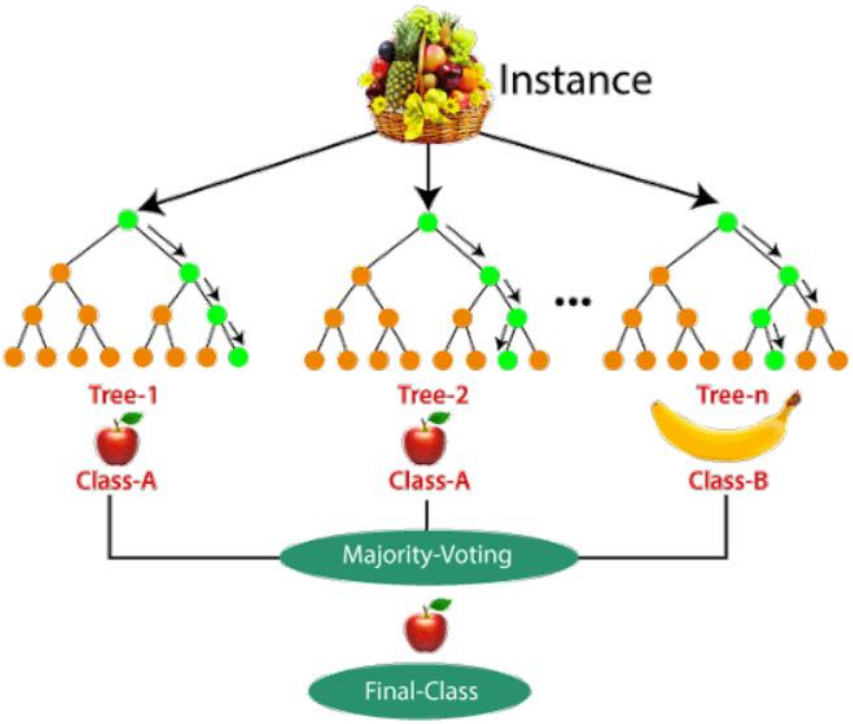
χαμηλή. Εάν η υγρασία είναι υψηλή, τότε οδηγούμαστε στο δεύτερο επίπεδο του δέντρου , στο κόμβο φύλλο με τιμή ‘’όχι (No)’’, με συνέπεια το παιδί να μην μπορεί να παίξει ποδόσφαιρο, ενώ αν η υγρασία βρίσκεται σε κανονικά επίπεδα τότε οδηγούμαστε στο φύλλο με τιμή ‘’ναι (Yes)’’ , επιτρέποντας στο παιδάκι να παίξει ποδόσφαιρο.[12]



Εικόνα 11 : Παράδειγμα δένδρου απόφασης(Analytics Vidhya,2024).

2.2.4 Τυχαία Δάση

Τα τυχαία δάση είναι ένας αλγόριθμος αρκετά όμοιος με τα δένδρα απόφασης, καθώς στην φάση της εκπαίδευσης κατασκευάζει ένα πλήθος από δένδρα απόφασης. Ένα από τα κρισιμότερα χαρακτηριστικά των τυχαίων δασών είναι η ικανότητά τους να αποδίδουν εξίσου καλά τόσο σε προβλήματα ταξινόμησης ,όσο και σε προβλήματα παλινδρόμησης διότι είναι ικανοί να χειριστούν κατηγορικές μεταβλητές (ταξινόμηση), αλλά και συνεχείς μεταβλητές (παλινδρόμηση). Αρχικά, αυτή η τεχνική, δημιουργεί έναν αριθμό από δένδρα απόφασης βασισμένα σε διαφορετικά υποσύνολα του συνόλου δεδομένων με συνέπεια καθένα από αυτά τα δέντρα να δημιουργεί ένα διαφορετικό υποσύνολο χαρακτηριστικών. Έτσι, δημιουργείται μεγάλη ποικιλομορφία δέντρων με μικρή συσχέτιση μεταξύ τους καθώς υπάρχει τυχαιότητα στην επιλογή του υποσυνόλου δεδομένων για την δημιουργία τους, μειώνοντας τον κίνδυνο υπερπροσαρμογής και βελτιώνοντας την ακρίβεια της πρόβλεψης. Έπειτα , στην φάση της πρόβλεψης, ο αλγόριθμος συναθροίζει τα αποτελέσματα των δένδρων, είτε με την πρόβλεψη της πλειοψηφίας τους (ταξινόμηση), είτε με την τον υπολογισμό του μέσου όρου των αποτελεσμάτων που εξάγονται από αυτά (παλινδρόμηση). Η ικανότητα, λοιπόν, του αλγορίθμου να χειρίζεται πολύπλοκα σύνολα δεδομένων τον καθιστά εξαιρετικά αξιόπιστο σε διάφορα προβλήματα πρόβλεψης.[12]



Εικόνα 12 : Παράδειγμα Τυχαίου δάσους (Analytics Vidhya,2024).

Παραπάνω οπτικοποιείται ένα παράδειγμα ταξινόμησης ,με το σύνολο δεδομένων να αντικατοπτρίζεται από το καλάθι των φρούτων. Όπως φαίνεται στην φάση της εκπαίδευσης, για κάθε τυχαίο δείγμα του καλαθιού , δημιουργείται ένα μεμονωμένο δέντρο απόφασης που αποδίδει μία έξοδο. Στη συνέχεια, το τελικό αποτέλεσμα εξάγεται από την πλειοψηφία της κλάσης (φρούτο) που προκύπτει ως έξοδος κάθε δένδρου, με το μήλο (2) στην προκειμένη περίπτωση να είναι αυτό που υπερτερεί , έναντι της μπανάνας (1) .

Καθίσταται εμφανές λοιπόν, ότι ο αλγόριθμος των τυχαίων δασών μοιάζει αρκετά με τον αλγόριθμο των δέντρων απόφασης. Ωστόσο μπορούν να εντοπιστούν αρκετές διαφορές στην λειτουργία τους [12] :

- Τα δένδρα απόφασης με μεγάλο βάθος, δεν εγγυόνται καλύτερα αποτελέσματα διότι μπορεί να πάσχουν από το πρόβλημα της υπερπροσαρμογής. Αντιθέτως τα τυχαία δάση , αποτελούνται από τυχαία δείγματα του συνόλου δεδομένων δημιουργώντας έτσι δένδρα απόφασης με μεγάλη μεταβλητότητα και ποικιλομορφία , γεγονός που αντιμετωπίζει τον κίνδυνο υπερπροσαρμογής.

- Το τυχαίο δάσος επιλέγει με τυχαιότητα τα δείγματα και δημιουργεί δένδρα χωρίς να διέπεται από σύνολο κανόνων. Σε αντίθεση, ο αλγόριθμος δένδρων απόφασης διέπεται αυστηρά από σύνολο κανόνων που καθορίζουν την λήψη αποτελέσματος.

2.3 Μετρικές Αξιολόγησης Μοντέλου

Η κατασκευή ενός μοντέλου μηχανικής μάθησης , επιτάσσει την αξιολόγησή του πάνω σε διάφορες μετρικές, οι οποίες προσφέρουν ένα αξιόπιστο αποτέλεσμα σε αθέατα δεδομένα ώστε να βελτιστοποιηθεί η απόδοσή του. Υπάρχουν πολλές μετρικές αξιολόγησης ενός μοντέλου παλινδρόμησης. Παρακάτω παρατίθενται οι πιο συνηθισμένες.[13]

2.3.1 Μέσο Απόλυτο Σφάλμα

Το μέσο απόλυτο σφάλμα (Mean Absolute Error, MAE), αποτελεί μία μετρική που προτιμάται σε μεγάλο βαθμό στο πεδίο της μηχανικής μάθησης. Ουσιαστικά, υπολογίζει το άθροισμα όλων των απόλυτων διαφορών μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών των δειγμάτων που ανήκουν σε ένα σύνολο δεδομένων και διαιρεί προς το πλήθος των δειγμάτων βέβαια. Ο μαθηματικός τύπος εκφράζεται από την σχέση :

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

Όπου

- n είναι το σύνολο των σημείων/δεδομένων.
- x_i Αναπαριστά την πραγματική τιμή για το i-οστό σημείο.
- y_i Αναπαριστά την τιμή που προβλέφθηκε από το μοντέλο για το i-οστό σημείο.

2.3.2 Μέσο Τετραγωνικό Σφάλμα

Το μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE) αντιπροσωπεύει το άθροισμα της τετραγωνικής ρίζας της απόστασης μεταξύ των προβλεπόμενων και των πραγματικών τιμών των δειγμάτων που ανήκουν στο σύνολο δεδομένων, προς το πλήθος των δειγμάτων. Ο τύπος της μετρικής εκφράζεται από την παρακάτω σχέση :

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

Όπου:

- n είναι το σύνολο των σημείων/δεδομένων.
- x_i αναπαριστά την πραγματική τιμή για το i-οστό σημείο.
- y_i αναπαριστά την τιμή που προβλέφθηκε από το μοντέλο για το i-οστό σημείο.

2.3.3 Τετραγωνική Ρίζα Του MSE

Η τετραγωνική ρίζα του MSE στην ουσία ποσοτικοποιεί τον βαθμό που συντρέχουν οι πραγματικές τιμές των δειγμάτων με τις προβλεπόμενες τιμές του μοντέλου. Ο τύπος φαίνεται παρακάτω:

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

- n είναι το σύνολο των σημείων/δεδομένων.
- x_i αναπαριστά την πραγματική τιμή για το i-οστό σημείο.
- y_i αναπαριστά την τιμή που προβλέφθηκε από το μοντέλο για το i-οστό σημείο.

2.3.4 R^2

Η μετρική τετράγωνο R , αποτελεί ίσως την δημοφιλέστερη στατιστική μέτρηση. Δείχνει πόσο καλά προσαρμόζονται τα δεδομένα στο μοντέλο παλινδρόμησης καθώς εξηγεί την αναλογία της διακύμανσης της εξαρτημένης μεταβλητής, η οποία καθορίζεται από τις ανεξάρτητες μεταβλητές. Η τιμές που μπορεί να πάρει ανήκουν από 0 έως και 1. Ο τύπος παρατίθεται παρακάτω :

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

Όπου:

- n είναι το σύνολο των σημείων/δεδομένων.
- y_i αναπαριστά την πραγματική τιμή για το i-οστό σημείο.
- $\hat{y}_i$ αναπαριστά την τιμή που προβλέφθηκε από το μοντέλο για το i-οστό σημείο.
- $\bar{y}_i$ είναι η μέση τιμή των πραγματικών τιμών.

2.4 Στατιστικές Τεχνικές

Παρακάτω αναφέρονται οι στατιστικές τεχνικές που χρησιμοποιήθηκαν για την ανάλυση δεδομένων και την εξαγωγή συμπερασμάτων.

2.4.1 Μέση Τιμή

Η μέση τιμή αποτελεί τον μέσο όρο των τιμών σε ένα σύνολο δεδομένων και ορίζεται ως :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{n}(t_1 + \cdots + t_n)$$

Όπου n είναι το πλήθος των δειγμάτων και t_i, η i-οστή παρατήρηση.

2.4.2 Μεσαία Τιμή

Η μεσαία τιμή ή διάμεσος, αντιπροσωπεύει την τιμή που βρίσκεται στο κέντρο ενός ταξινομημένου συνόλου n-πλήθους δεδομένων, χωρίζοντάς το σε δύο ίσα μέρη.
Ορίζεται ως :

$$\mathrm{median}(x) = x_{(n+1)/2}$$

2.4.3 Μέθοδος IQR

Στην στατιστική , το ενδοτεταρτημοριακό εύρος (IQR), είναι μία τεχνική που βοηθάει στον εντοπισμό των ακραίων τιμών και χρησιμοποιείται για να κατανοηθεί ο βαθμός της διασποράς των δεδομένων. Ουσιαστικά, περιγράφει την κατανομή των δεδομένων, δείχνοντας την απόσταση μεταξύ του τρίτου και του πρώτου τεταρτημορίου σε ένα σύνολο δεδομένων. Ο τύπος IQR είναι ο εξής :

$$\text{IQR}=Q3-Q1$$

Όπου :

- Q_3 είναι το τρίτο τεταρτημόριο
- Q_1 είναι το πρώτο τεταρτημόριο

Αρχικά, ταξινομούμε τα στοιχεία του συνόλου, κατά αύξουσα σειρά και εντοπίζουμε την μεσαία τιμή τους, η οποία αντιπροσωπεύει το δεύτερο τεταρτημόριο. Στη συνέχεια, η μεσαία τιμή των δεδομένων που βρίσκονται κάτω από το δεύτερο τεταρτημόριο, αποτελεί το πρώτο τεταρτημόριο, ενώ η μεσαία τιμή των δεδομένων που βρίσκονται πάνω από το δεύτερο τεταρτημόριο, αποτελεί το τρίτο τεταρτημόριο. Η τιμή του ενδοτεταρτημοριακού εύρους προκύπτει από την αφαίρεση του Q_1 από το Q_3.

Οι ακραίες τιμές εντοπίζονται, χρησιμοποιώντας την τιμή που βρήκαμε παραπάνω για την δημιουργία κάτω και άνω ορίου :

- Άνω όριο: $Q3+1.5\times\text{IQR}$
- Κάτω όριο: $Q1-1.5\times\text{IQR}$

Συνεπώς οποιαδήποτε τιμή βρίσκεται εκτός των δύο ορίων θεωρείται ακραία τιμή.

2.4.4 Τυποποίηση

Τυποποίηση (Standardization) , είναι μια τεχνική της στατιστικής που εφαρμόζεται σε σύνολα δεδομένων που έχουν μεγάλη απόκλιση τιμών στα δείγματά τους. Μετατρέπει τα δεδομένα έτσι ώστε η μέση τιμή τους να είναι ίση με 0 , και η διακύμανση ίση με 1.
Ορίζεται:

$$z = \frac{x_i - \mu}{\sigma}$$

3.Σιγκαπούρη

Η Σιγκαπούρη είναι μια μικρού μεγέθους πυκνοκατοικημένη νησιωτική πόλη-κράτος που βρίσκεται στην Νοτιοανατολική Ασία, συνορεύοντας με το νότιο τμήμα της Μαλαισίας. Η έκτασή της είναι μόλις 734,4 τετραγωνικά χιλιόμετρα καθιστώντας την, μικρότερη χώρα στην Νοτιοανατολική Ασία. Αποτελείται από πολυεθνή πληθυσμό με κυριότερες χώρες προέλευσης τις: Κίνα , Ταμίλ, Μαλαισία. Το κέντρο της πόλης εντοπίζεται στο νότιο τμήμα της, κοντά στις εκβολές του ποταμού της Σιγκαπούρης. Αποτελούσε βρετανική αποικία μέχρι το 1950, ενώ αργότερα εντάχθηκε στην Μαλαισία από το 1963 μέχρι το 1965, όπου και ανεξαρτητοποιήθηκε.[14]
Θεωρείται ένα ραγδαία εξελισσόμενο κράτος με υψηλό κατά κεφαλήν εισόδημα (56.248 δολάρια) που ξεπερνά μέχρι και την Ιαπωνία το 2017, καθιστώντας την ένα από τα πλουσιότερα κράτη της Ασίας. Σύμφωνα με την απογραφή του 2023 που πραγματοποίησε το Τμήμα Στατιστικών Σιγκαπούρης, ο πληθυσμός της είναι 5.917.600 κάτοικοι εκ των οποίων οι 4.149.000 είναι μόνιμοι κάτοικοι. Το 82% των κατοίκων ζει σε μισθωμένα (99 έτη) διαμερίσματα τα οποία επιδοτούνται από το κράτος και διαχειρίζονται από το Συμβούλιο Στέγασης και Ανάπτυξης . Πρόκειται για έναν κρατικό οργανισμό που ιδρύθηκε το 1960, αποσκοπώντας στην καταπολέμηση της στεγαστικής κρίσης στην Σιγκαπούρη. Ο υπερβολικά αυξανόμενος πληθυσμός οδήγησε στην ανάγκη κατασκευής απλών και οικονομικών διαμερισμάτων προκειμένου να καλυφθεί άμεσα η μεγάλη ζήτηση, ωστόσο με την πάροδο των ετών το συμβούλιο στέγασης εξέλιξε τα διαμερίσματα επενδύοντας σε σπίτια με περιβαλλοντική βιωσιμότητα, δημιουργώντας έτσι εξελιγμένες κοινότητες χάρις τις καινοτόμες και πολυτελείς ανέσεις που παρέχει για διαφορετικές εισοδηματικές ομάδες. Τα νέα διαμερίσματα που διατίθενται προς πώληση μέσω επιδοτούμενων τιμών, έχουν μεγάλο εύρος μεγεθών και πωλούνται αποκλειστικά μόνο σε πολίτες της Σιγκαπούρης που πληρούν συγκεκριμένα κριτήρια. [15]

Είναι σαφές λοιπόν , πως το Συμβούλιο Ανάπτυξης και Στέγασης συνδράμει αποφασιστικά στην εξέλιξη του βιοτικού επιπέδου των πολιτών της Σιγκαπούρης προωθώντας την αειφορία καθώς ο σχεδιασμός και η ανάπτυξη της δημόσιας στέγασης που παρέχει, βασίζεται σε πράσινες τεχνολογίες. Παράλληλα, διαθέτει μια μεγάλη ποικιλομορφία διαμερισμάτων ικανή να καλύψει κάθε είδους ανάγκη , εξυπηρετώντας αποτελεσματικά πολίτες που ανήκουν σε διαφορετικές κοινωνικές τάξεις.

4.Μεθοδολογία

Στο παρόν κεφάλαιο παρατίθενται τα βήματα που πραγματοποιήθηκαν για τον σκοπό της εργασίας και αναφέρονται τα εργαλεία που χρησιμοποιήθηκαν τόσο για την ανάπτυξη του κώδικα, όσο και για την αξιολόγηση των αποτελεσμάτων.

4.1 Πηγή Δεδομένων

Για την υλοποίηση της πειραματικής αξιολόγησης χρησιμοποιήθηκε η βάση δεδομένων '' Resale flat prices based on registration date from Jan-2017 onwards ''[16] από την ιστοσελίδα της Κυβερνητικής Υπηρεσίας Σιγκαπούρης (https://beta.data.gov.sg). Η συγκεκριμένη βάση δεδομένων ανανεώνεται καθημερινά από την υπηρεσία καθώς συνδέεται άμεσα με το Συμβούλιο Στέγασης και Ανάπτυξης της Σιγκαπούρης. Κάθε εγγραφή της βάσης αφορά και ένα διαμέρισμα που τίθεται προς πώληση. Η λήψη του αρχείου μορφής '.csv' πραγματοποιήθηκε στις 12 Απριλίου 2024, συνεπώς η τελευταία εγγραφή της βάσης χρονολογείται μέχρι και τις 12 Απριλίου.

Παρακάτω, αναφέρονται πληροφορίες σχετικά με τις ανεξάρτητες μεταβλητές σύμφωνα με την Κυβερνητική Υπηρεσία Σιγκαπούρης :

- Month : Μήνας πώλησης διαμερίσματος
- Town : Καθορισμένη κατοικημένη περιοχή
- Flat_type : Ταξινόμηση δείγματος ανάλογα με το μέγεθος των δωματίων
- Block : Οικοδομικό τετράγωνο της κατοικημένης περιοχής
- Street_name : Οδός διαμερίσματος
- Storey_range : Εκτιμώμενο εύρος ορόφων διαμερίσματος
- Floor_area_sqm : Συνολικό εμβαδό διαμερίσματοσε τετραγωνικά εκατοστά
- Flat_model : Ταξινόμηση του δείγματος ανάλογα με μοντέλο του διαμερίσματος
- Lease_commence_date : Ημερομηνία εκκίνησης της μίσθωσης
- Remaining_lease : Εναπομείναντα χρόνια μέχρι το τέλος της μίσθωσης
- Resale_price : Τιμή πώλησης του διαμερίσματος

4.2 Βήματα Υλοποίησης

Για την επίτευξη της γραφής του κώδικα, ερευνήθηκε υλικό από notebooks στην πλατφόρμα Kaggle , GitHub και χρειάστηκε να ολοκληρωθεί η παρακολούθηση μιας σειράς από ασύγχρονα μαθήματα του Udemy αλλά και τις πλατφόρμας YouTube.
Αρχικά, προκειμένου να καταστεί περισσότερο κατανοητός ο τομέας της επιστήμης των δεδομένων πραγματοποιήθηκε παρακολούθηση βίντεο ασύγχρονης εκπαίδευσης που σχετίζονται με την ανάλυση, την οπτικοποίηση δεδομένων αλλά και την δημιουργία μοντέλων μηχανικής μάθησης, στις πλατφόρμες Udemy [17][18] και Youtube [19][20]. Επίσης, ανακτήθηκε πολύτιμη γνώση μέσα από την αναζήτηση επιστημονικών άρθρων σε επίσημες ιστοσελίδες.

Στη συνέχεια, μελετήθηκαν διάφορα notebooks από χρήστες του Kaggle , που σχετίζονται με την δημιουργία μοντέλων πρόβλεψης τιμής. Λόγου χάριν, ένας χρήστης επιλέγοντας τον αλγόριθμο των δέντρων απόφασης με προεπιλεγμένες τιμές παραμέτρων, εμφάνισε r^2 -score ίσο με 79%, επιλέγοντας τον αλγόριθμο των τυχαίων δασών εμφάνισε r^2 -score ίσο με 85%, χρησιμοποιώντας τον αλγόριθμο της παλινδρόμησης Ridge εμφάνισε r^2 -score 72% και τέλος επιλέγοντας τον αλγόριθμο της παλινδρόμησης Lasso πέτυχε r^2 -score ίσο με 72%.[21]

Έπειτα, ανακτήθηκε η βάση δεδομένων από την ιστοσελίδα της Κυβερνητικής Υπηρεσίας Σιγκαπούρης, ακολούθησε η εξερεύνηση και η προετοιμασία για ανάλυση δεδομένων. Στο κομμάτι της ανάλυσης δεδομένων εντοπίστηκαν οι σχέσεις μεταξύ των μεταβλητών και η επιρροή τους στην μεταβλητή στόχο(μεταβλητή που πρόκειται να προβλεφθεί αργότερα από το μοντέλο). Συνάμα, εξήχθηκαν χρήσιμες πληροφορίες που συνδράμουν στην κατασκευή των μοντέλων μηχανικής μάθησης με τις απαραίτητες ανεξάρτητες μεταβλητές, οι οποίες έγιναν περισσότερο κατανοητές μέσα από την μελέτη ιστοσελίδων που αφορούν την ακίνητη περιουσία της Σιγκαπούρης [22]. Τέλος, στο κομμάτι της εκπαίδευσης των μοντέλων με τους επιλεγμένους αλγόριθμους ,σε πρώτη φάση οι αλγόριθμοι τρέχουν με τις προεπιλεγμένες τιμές παραμέτρων και γίνεται αξιολόγηση με τις μετρικές που έχουν επιλεχθεί. Σε δεύτερη φάση, όλοι οι αλγόριθμοι υπερ-παραμετροποιήθηκαν ξεχωριστά, έτρεξαν για άλλη μια φορά και αξιολογήθηκαν με τις ίδιες μετρικές. Κλείνοντας, δημιουργούνται γραφήματα στα οποία συγκρίνονται τα αποτελέσματα των εκάστοτε αλγορίθμων και εξάγονται τα αντίστοιχα συμπεράσματα.

4.3 Εργαλεία

Η εργασία υλοποιήθηκε εξ ολοκλήρου σε γλώσσα Python , έκδοση 3.11.4. Η εξερεύνηση και προετοιμασία της βάσης έλαβε χώρα στο περιβάλλον του Spyder έκδοση 5.4.3. Έπειτα , η στατιστική ανάλυση των δεδομένων και η εκπαίδευση των μοντέλων έγινε μέσω Jupyter Notebook έκδοση 6.5.4 και Google Colaboratory.

Το Spyder είναι ένα επιστημονικό προγραμματιστικό περιβάλλον της Python. Διανέμεται δωρεάν μέσω της πλατφόρμας Anaconda και περιλαμβάνει λειτουργίες προηγμένης επεξεργασίας, διαδραστικές δοκιμές και ικανότητα αποσφαλμάτωσης κώδικα .

Το Jupyter Notebook είναι ένα διαδρασικό περιβάλλον που διανέμεται δωρεάν από την πλατφόρμα Anaconda, ικανό να εκτελέσει, να επεξεργαστεί και να αναλύσει δεδομένα.

Το Google Colaboratory αποτελεί πανομοιότυπο εργαλείο με το Jupyter Notebook, το οποίο δεν χρειάζεται εγκατάσταση και παρέχει δωρεάν υπολογιστικούς πόρους.

5.Στατιστική Ανάλυση

Η στατιστική ανάλυση μιας βάσης δεδομένων αφορά την εξερεύνηση, την ερμηνεία και συλλογή συμπερασμάτων τα οποία απορρέουν από τα δεδομένα της βάσης. Με την βοήθεια στατιστικών εργαλείων επιτυγχάνεται η ανάλυση των δεδομένων, που έχει ως στόχο την αναγνώριση προτύπων μέσω οπτικοποιήσεων και την ανίχνευση ανωμαλιών ή αποκλίσεων στα δεδομένα.

5.1 Εξερεύνηση Βάσης Δεδομένων

Στο αρχικό στάδιο , κρίνεται απαραίτητο να εξερευνηθεί η βάση δεδομένων με σκοπό την καλύτερη κατανόησή της ,ενώ παράλληλα θα προετοιμαστεί κατάλληλα για το επόμενο στάδιο της στατιστικής ανάλυσης.
Αρχικά, παρουσιάζονται οι πρώτες πέντε εγγραφές για κάθε μια ανεξάρτητη μεταβλητή που αντιπροσωπεύει το όνομα της στήλης.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
0	2017-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44.0	Improved	1979	61 years 04 months	232000.0
1	2017-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67.0	New Generation	1978	60 years 07 months	250000.0
2	2017-01	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	262000.0
3	2017-01	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68.0	New Generation	1980	62 years 01 month	265000.0
4	2017-01	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67.0	New Generation	1980	62 years 05 months	265000.0

Εικόνα 13 : Οι πρώτες εγγραφές της βάσης

```
Exploring columns....
Index(['month', 'town', 'flat_type', 'block', 'street_name', 'storey_range',
       'floor_area_sqm', 'flat_model', 'lease_commence_date',
       'remaining_lease', 'resale_price'],
      dtype='object')
```

Εικόνα 14 : Ανεξάρτητες μεταβλητές

Εκ πρώτης όψεως, φαίνεται πως οι τιμές κάποιων ανεξάρτητων μεταβλητών είναι ακανόνιστες, καθώς αποτυπώνονται είτε με εξ΄ ολοκλήρου κεφαλαίους χαρακτήρες, είτε με πεζούς χαρακτήρες. Παρόλα αυτά, είναι αρκετά εύκολο να κατανοήσει κάποιος τα δεδομένα καθώς η ερμηνεία των κλάσεων είναι ξεκάθαρη.

Σύμφωνα με τις εικόνες, η βάση απαρτίζεται από 176976 εγγραφές με 11 στήλες έκαστην. Κάθε εγγραφή αποτελεί και ένα διαμέρισμα του οποίου πληροφορίες σχετικά με την τιμή , το μέγεθος, την τοποθεσία, το είδος και την μίσθωσή του παρέχονται από τις στήλες. Είναι σημαντικό να σημειωθεί πως κατά την σάρωση της βάσης δεν βρέθηκαν κενές τιμές. Ωστόσο εντοπίστηκαν 566 διπλότυπες εγγραφές.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 176976 entries, 0 to 176975
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   month                176976 non-null  object
 1   town                 176976 non-null  object
 2   flat_type            176976 non-null  object
 3   block                176976 non-null  object
 4   street_name          176976 non-null  object
 5   storey_range         176976 non-null  object
 6   floor_area_sqm       176976 non-null  float64
 7   flat_model           176976 non-null  object
 8   lease_commence_date  176976 non-null  int64
 9   remaining_lease      176976 non-null  object
 10  resale_price         176976 non-null  float64
dtypes: float64(2), int64(1), object(8)
memory usage: 14.9+ MB
None
```

Εικόνα 15 : Περιγραφή ανεξάρτητων μεταβλητών

```
NaN values detection : [False False False False False False False False False False False]


Number of duplicates: 566
```

Εικόνα 16 : Εμφάνιση διπλότυπων εγγραφών και κενών τιμών

Σε αυτό το σημείο οφείλουμε να παρατηρήσουμε πως οι περισσότερες ανεξάρτητες μεταβλητές είναι τύπου ‘object’, γεγονός που καθιστά δύσκολη την οπτικοποίησή τους σε διαγράμματα επειδή πολλές από αυτές αξιοποιούνται καλύτερα όταν λαμβάνουν αριθμητικές τιμές. Συγκεκριμένα, η μεταβλητή ‘remaining_lease’ που αντιπροσωπεύει τα εναπομείναντα χρόνια μίσθωσης, αποτελείται από 674 διαφορετικές αλφαριθμητικές τιμές και θα μετατραπεί σε μεταβλητή δεκαδικής μορφής (float). Η μεταβλητή ‘month’ που αντιπροσωπεύει τον μήνα και έτος πώλησης, θα μετατραπεί σε μεταβλητή τύπου ‘datetime’, δίνοντας μας την δυνατότητα να αποσπάσουμε το έτος και τον μήνα χωριστά.

```
Converted remaining lease :
0    61.333333
1    60.583333
2    62.416667
Name: remaining_lease, dtype: float64


Converted month :
0   2017-01-01
1   2017-01-01
2   2017-01-01
Name: month, dtype: datetime64[ns]
```

Εικόνα 17 : Μετατροπή μεταβλητών

Στη συνέχεια θα δημιουργήσουμε μια νέα μεταβλητή ‘Region’, η οποία θα απορρέει από τις τιμές της μεταβλητής ‘town’. Ουσιαστικά η νέα μεταβλητή αποτελείται από 5 διακριτές τιμές που αντιπροσωπεύουν την ευρύτερη περιοχή της Σιγκαπούρης στην οποία ανήκει η εκάστοτε πόλη. Με αυτόν τον τρόπο έχουμε την δυνατότητα να εξάγουμε χρήσιμα συμπεράσματα για τα διαμερίσματα από μια διαφορετική όψη , ενώ ταυτόχρονα μειώνουμε τις διακριτές τιμές από 26 σε μόλις 5. Στην επόμενη φωτογραφία εμφανίζεται το σύνολο τιμών της νέας μεταβλητής ‘Region’ και οι περιοχές που ανήκουν οι πρώτες πέντε εγγραφές. Στη συγκεκριμένη περίπτωση τα πρώτα διαμερίσματα τυγχάνουν να ανήκουν στην κεντρική περιοχή της Σιγκαπούρης.

```
Region
0    Central
1    Central
2    Central
3    Central
4    Central
Name: Region, dtype: object

Unique values of 'Region' are :
['Central' 'East' 'West' 'South' 'North']
```

Εικόνα 18 : Δημιουργία μεταβλητής 'Region'

Η κατάταξη των πόλεων στις εκάστοτε κλάσεις γίνεται με βάση το υπόμνημα [22] της παρακάτω φωτογραφίας :

Region	Town Centres / Areas
North	Admirality, Kranji, Woodlands, Sembawang, Yishun, Yio Chu Kang, Seletar, Sengkang
South	Holland, Queenstown, Bukit Merah, Telok Blangah, Pasir Panjang, Sentosa, Bukit Timah, Newton, Orchard, City, Marina South
East	Serangoon, Punggol, Hougang, Tampines, Pasir Ris, Loyang, Simei, Kallang, Katong, East Coast, Macpherson, Bedok, Pulau Ubin, Pulau Tekong
West	Lim Chu Kang, Choa Chu Kang, Bukit Panjang, Tuas, Jurong East, Jurong West, Jurong Industrial Estate, Bukit Batok, Hillview, West Coast, Clementi
Central	Thomson, Marymount, Sin Ming, Ang Mo Kio, Bishan, Serangoon Gardens, MacRitchie, Toa Payoh

Εικόνα 19 : Υπόμνημα περιοχών Σιγκαπούρης

Τέλος τροποποιούμε τα ονόματα των στηλών και τα ονόματα των τιμών της μεταβλητής 'flat_model' αφαιρώντας περιττά σημεία στίξης όπου είναι επιτρεπτό. Με αυτό τον τρόπο βελτιώνεται η αναγνωσιμότητα της βάσης.

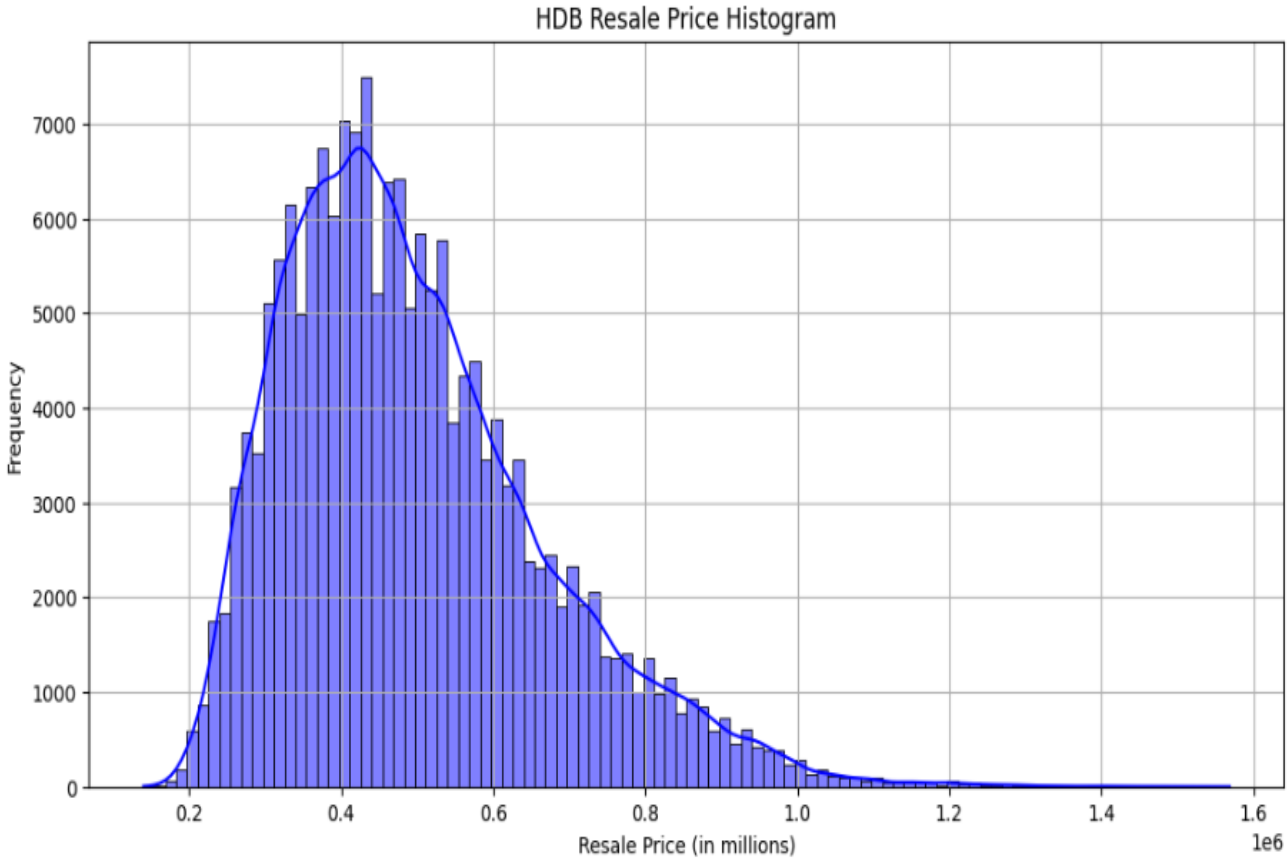
```
Renamed flat_model values...
['IMPROVED' 'NEW GENERATION' 'DBSS' 'STANDARD' 'APARTMENT' 'SIMPLIFIED'
 'MODEL A' 'PREMIUM APARTMENT' 'ADJOINED FLAT' 'MODEL A MAISONETTE'
 'MAISONETTE' 'TYPE S1' 'TYPE S2' 'MODEL A2' 'TERRACE'
 'IMPROVED MAISONETTE' 'PREMIUM MAISONETTE' 'MULTI GENERATION'
 'PREMIUM APARTMENT LOFT' '2 ROOM' '3GEN']


Renamed columns.....
Index(['Month', 'Town', 'Flat_Type', 'Block', 'Street_Name', 'Storey_Range',
       'Floor_Area_sqm', 'Flat_Model', 'Lease_Commence_Date',
       'Remaining_Lease', 'Resale_Price'],
      dtype='object')
```

Εικόνα 20 : Ενημέρωση ονόματος στηλών και τιμών μεταβλητών

5.2 Οπτικοποίηση Δεδομένων

Το κομμάτι της οπτικοποίησης δεδομένων μας βοηθά να κατανοήσουμε τα μοτίβα και τις τάσεις των δεδομένων της βάσης που αναπαρίστανται μέσω διαγραμμάτων. Συγκεκριμένα, θα εντοπιστούν τάσεις σε δείκτες όπως η τιμή πώλησης ακινήτων, η γεωγραφική κατανομή, ο όροφος και πολλά ακόμη τα οποία συντελούν στην αναγνώριση σχέσεων και συσχετίσεων μεταξύ των ανεξαρτήτων μεταβλητών.

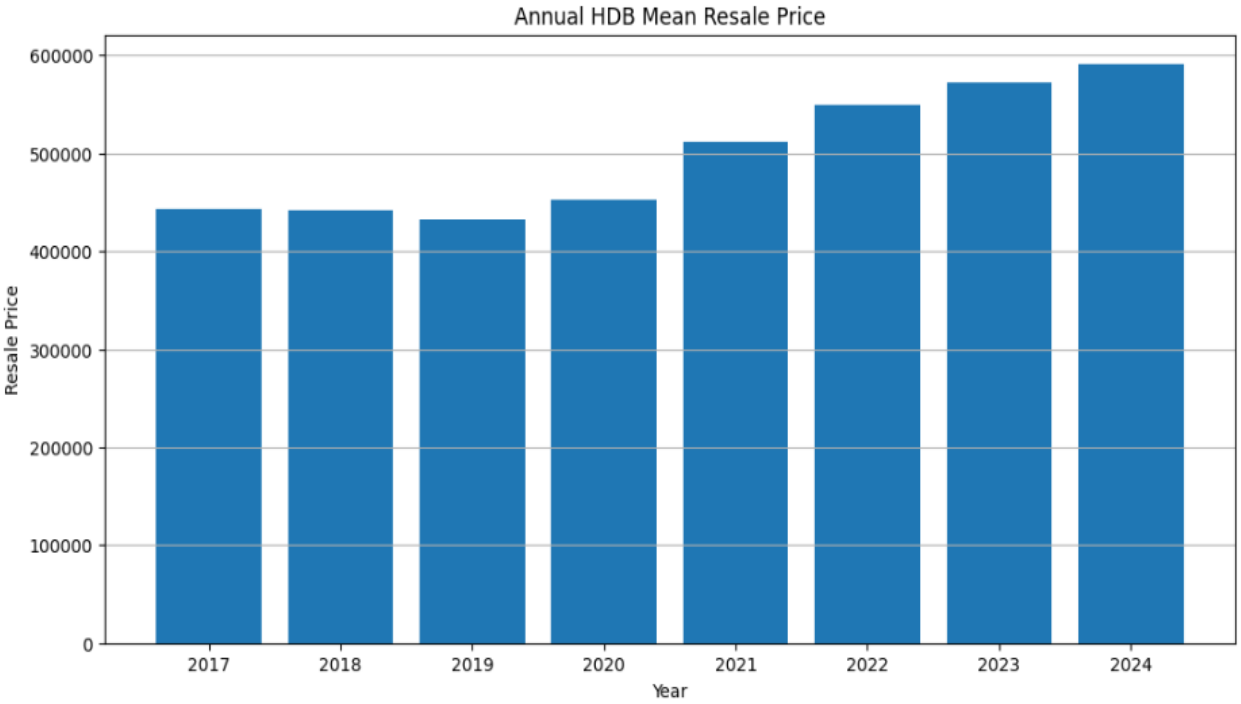
Ξεκινώντας , στο παρακάτω διάγραμμα παρουσιάζεται ένα ιστόγραμμα που οπτικοποιεί τις τιμές πώλησης των ακινήτων, με τον κατακόρυφο άξονα να δείχνει τον αριθμό των συναλλαγών και τον οριζόντιο άξονα να αντιπροσωπεύει τις τιμές πώλησης των ακινήτων σε εκατομμύρια δολάρια, οι οποίες κυμαίνονται από 0.1 έως και 1.6 εκατομμύρια δολάρια.



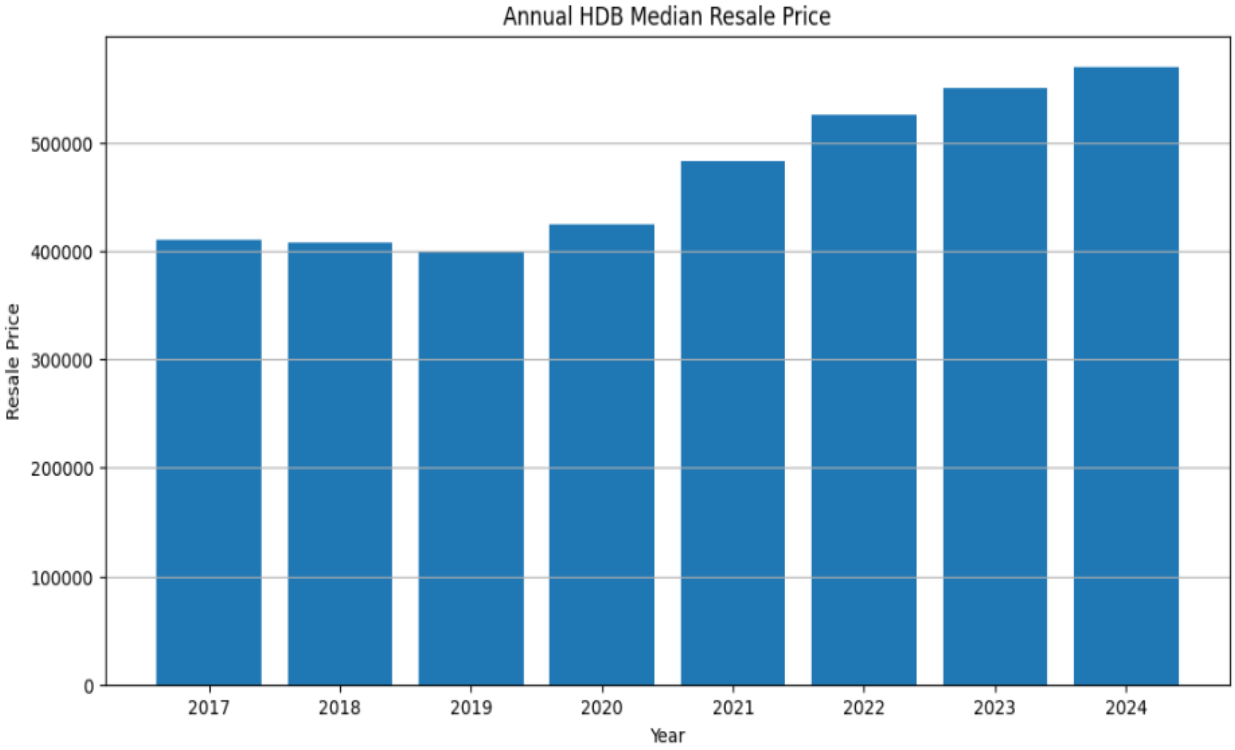
Εικόνα 21 : Ιστόγραμμα τιμών μεταπώλησης

Παρατηρώντας, στο διάγραμμα φαίνεται ξεκάθαρα μια ασύμμετρη δεξιά κατανομή καθώς η καμπύλη που διατρέχει το διάγραμμα κορυφώνεται στα αριστερά όπου το εύρος τιμών είναι χαμηλό και η συχνότητα των πωλήσεων υψηλή. Αναλυτικότερα, παρατηρείται πως τα περισσότερα διαμερίσματα πωλούνται σε τιμές που κυμαίνονται μεταξύ 0,2 και 0,6 εκατομμύρια δολάρια, με την συχνότητα να κορυφώνεται κοντά στα 0,4 εκατομμύρια δολάρια αποτελώντας έτσι την συνηθέστερη τιμή πώλησης. Έπειτα , παρατηρείται σταδιακή μείωση της συχνότητας όσο η τιμή αυξάνεται με συνέπεια τις ελάχιστες πωλήσεις σε τιμές άνω το ενός εκατομμυρίου.

Βάση του διαγράμματος συμπεραίνουμε πως η συντριπτική πλειοψηφία των πωλήσεων πραγματοποιείται σε προσιτές τιμές, ενώ παράλληλα φαίνεται πως σημειώνονται κάποιες ελάχιστες πωλήσεις σε υψηλές τιμές , κάτι που οφείλεται σε παράγοντες που θα αναλύσουμε στη συνέχεια.
Οι επόμενες δύο εικόνες που θα παρατίθενται αφορούν την ετήσια μέση τιμή μεταπώλησης και την ετήσια μεσαία τιμή μεταπώλησης από το 2017 έως το 2024 αντίστοιχα, με τον οριζόντιο άξονα να αναπαριστά τις χρονολογίες που σημειώθηκαν πωλήσεις και τον κατακόρυφο άξονα να αναπαριστά την τιμή μεταπώλησης των διαμερισμάτων.

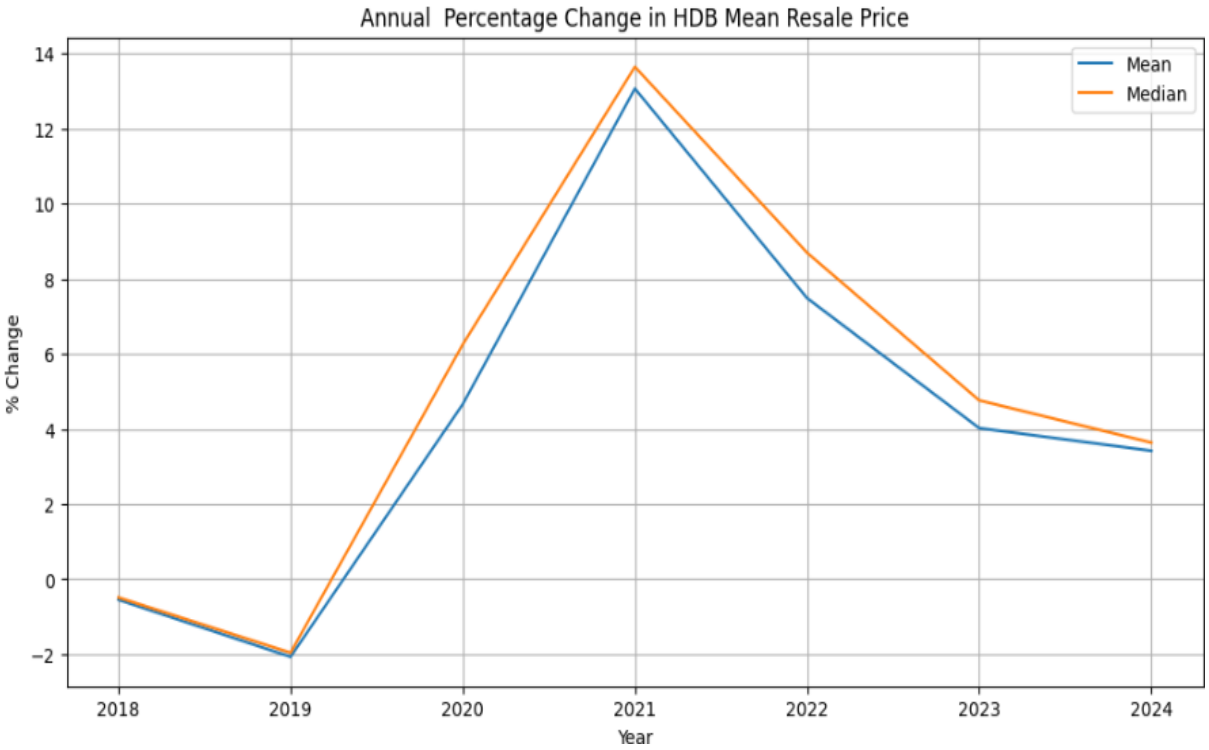


Εικόνα 22 : Ετήσια μέση τιμή μεταπώλησης



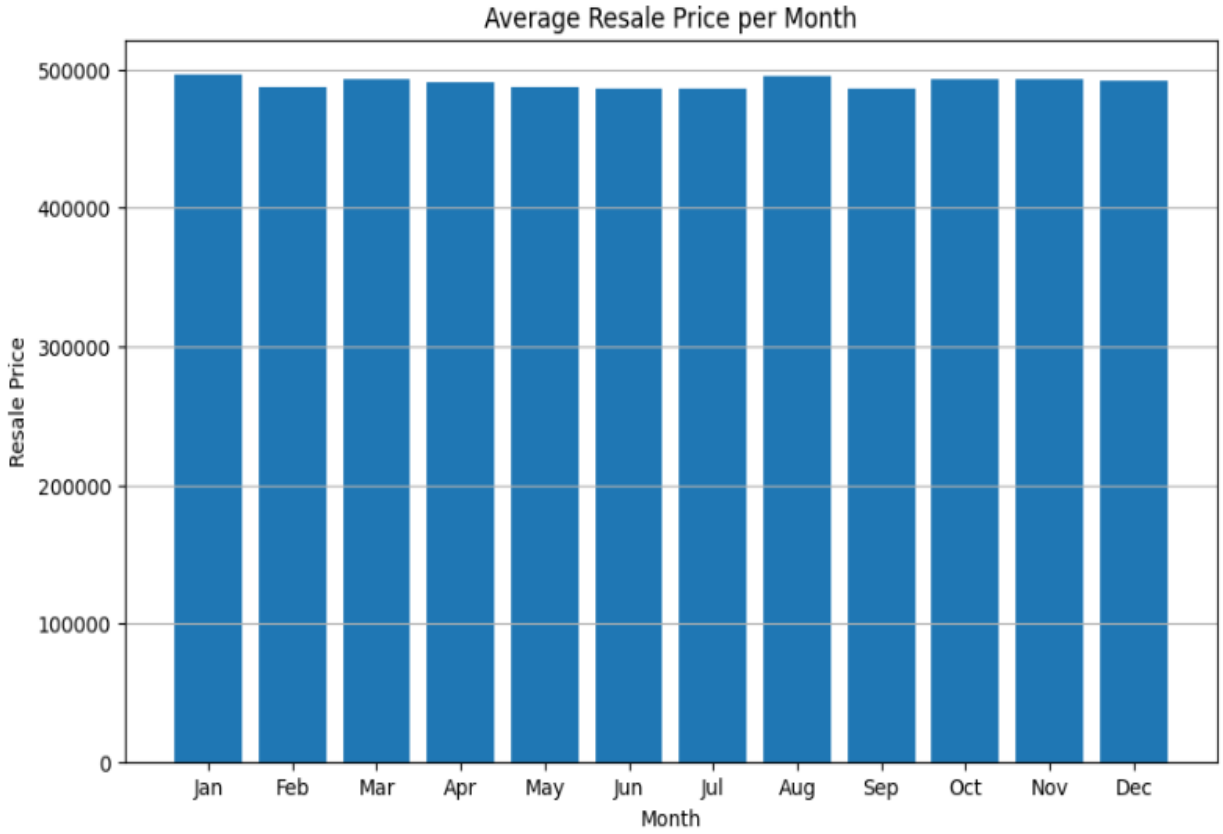
Εικόνα 23 : Ετήσια μεσαία τιμή μεταπώλησης

Συγκεκριμένα τόσο η μέση, όσο και η διάμεση ή μεσαία τιμή παρουσιάζει γενικά μια ανοδική τάση με την πάροδο των ετών. Σε ετήσια ανάλυση όμως, από το 2017 έως και το 2020 οι τιμές είναι σχετικά σταθερές με το μέγεθος των διακυμάνσεων να είναι αρκετά μικρό. Το 2021 υπάρχει ένα αισθητό άλμα στην μέση και μεσαία τιμή μεταπώλησης συγκριτικά με το προηγούμενο έτος, όπως και στην τριετία 2022 έως 2024 όπου η αύξηση μεγαλώνει και αποκτά σταθερό ρυθμό με την πάροδο του χρόνου.
Σε ποσοτική ανάλυση, η μέση τιμή το 2017 έως το 2020 αγγίζει ελαφρώς τα 450.000$, σημειώνοντας αισθητή αύξηση το έτος 2021 στα 490.000$, ενώ η μεσαία τιμή την πρώτη τριετία κυμαίνεται κοντά στα 400.000$ και αυξάνεται κατά 50.000$ με 60.000$ το 2021 . Την τελευταία τριετία, όπως προαναφέρθηκε, η μέση τιμή αυξάνεται σταθερά από 530.000$ το 2022 σε σχεδόν 600.000$ το 2024, ενώ η διάμεση τιμή ξεκινά περίπου στα 485.000$ και ανεβαίνει σταδιακά στα 560.000$.
Η διαφορά αυτή μεταξύ διάμεσης τιμής και μέσης τιμής επαληθεύει την αρχική εκτίμηση περί συμμετρικής κατανομής με δεξιά κλίση , καθώς ο μέσος όρος είναι μεγαλύτερος από τη διάμεσο , υποδηλώνοντας ότι υπάρχουν ακραίες τιμές στα δεξιά της κατανομής.



Εικόνα 24 : Ποσοστιαία μεταβολή μέση και μεσαίας τιμής

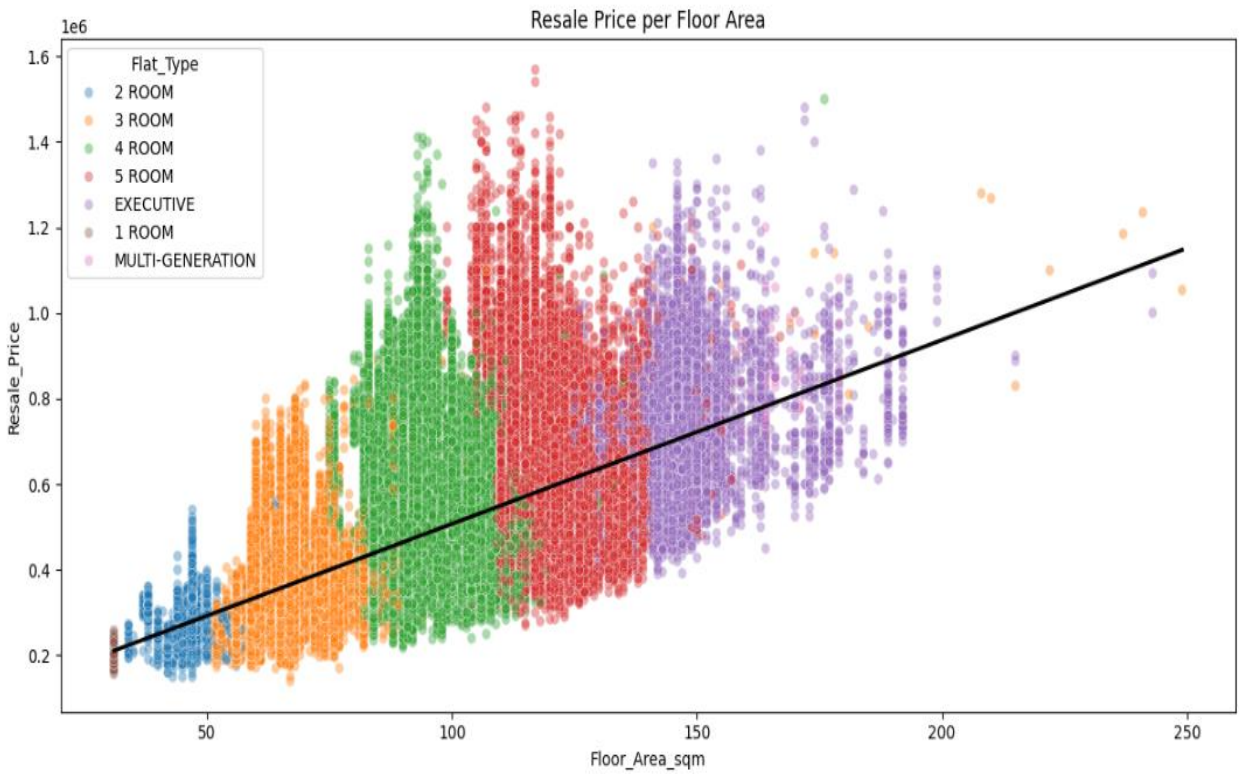
Η παραπάνω εικόνα δείχνει από διαφορετική σκοπιά τις παρόμοιες τάσεις της μεσαίας και μέσης τιμής σε ποσοστιαίο επίπεδο με τις αποκλίσεις των 2 γραμμών να είναι μικρές. Αν όμως αναλύσουμε τις πωλήσεις σε μηνιαία απόδοση δημιουργώντας ένα ραβδόγραμμα που αντιπροσωπεύει την μέση τιμή ανά μήνα, όπου κάθε ράβδος απεικονίζει τον μέσο όρο των τιμών μεταπώλησης του εκάστοτε μήνα συλλέγοντας τα δεδομένα των τελευταίων έξι ετών, θα παρατηρήσουμε τα εξής :



Εικόνα 25 : Μέσο όρος τιμής μεταπώλησης ανά μήνα

Σε αντίθεση με την ετήσια ανάλυση, εδώ παρατηρούμε πως η αγορά παρουσιάζει μια σταθερότητα με ελάχιστες διακυμάνσεις, τόσο μικρές που δεν μπορούμε να θεωρήσουμε ότι η τιμή επηρεάζεται από εποχιακούς παράγοντες. Συνεπώς η ανεξάρτητη μεταβλητή “Month” θα αφαιρεθεί από το σύνολο των μεταβλητών, καθώς δεν φαίνεται να έχει ουσιώδη επιρροή στην μεταβλητή στόχο.

Στη συνέχεια, φτάνουμε στο σημείο να εντάξουμε περισσότερες ανεξάρτητες μεταβλητές για να δούμε πως θα επηρεάσουν την τιμή μεταπώλησης. Λόγου χάριν, στο παρακάτω διάγραμμα διασποράς απεικονίζεται η σχέση μεταξύ του εμβαδού δαπέδου των διαμερισμάτων τα οποία είναι ομαδοποιημένα ανά τύπο διαμερίσματος και της τιμής μεταπώλησης.



Εικόνα 26 : Τιμή μεταπώλησης ανά επιφάνεια δαπέδου

Αναλυτικότερα, στον κατακόρυφο άξονα τοποθετείται η τιμή μεταπώλησης σε εκατομμύρια δολάρια , με τιμές που προσεγγίζουν από τα 0.2 εκατομμύρια δολάρια έως τα 1.6 εκατομμύρια δολάρια. Στον οριζόντιο άξονα αντιπροσωπεύεται το εμβαδό επιφάνειας σε τετραγωνικά μέτρα το οποίο κυμαίνεται από 20 έως και 250 τετραγωνικά μέτρα.
Κάθε σημείο του διαγράμματος αφορά και ένα διαμέρισμα το οποίο είναι χρωματισμένο ανάλογα με τον τύπο στον οποίο ανήκει (ο τύπος διαμερισμάτων θα αναλυθεί σε επερχόμενο γράφημα) :

- Καφέ : Διαμερίσματα ενός δωματίου τα οποία έχουν αρκετά μικρή επιφάνεια περίπου στα 25 τετραγωνικά μέτρα με συνέπεια την χαμηλή μεταπωλητική τους αξία που κυμαίνεται από 150.000$ έως 300.000$.

- Μπλε : Διαμερίσματα 2 δωματίων με εξίσου μικρή επιφάνεια που κατά μέσο όρο αγγίζει τα 40 με 50 τετραγωνικά μέτρα και αξία κοντά στα 300.000$. Ωστόσο παρατηρείται πως στα διαμερίσματα των 45 τετραγωνικών μέτρων η αξία σε πολλές περιπτώσεις μπορεί να φτάσει στα 600.000$.

- Πορτοκαλί : Διαμερίσματα 3 δωματίων με επιφάνεια που κυμαίνεται από 50 έως 80 τετραγωνικά μέτρα στις περισσότερες περιπτώσεις. Η αξία τους οριοθετείται από 200.000$ έως 800.000$.

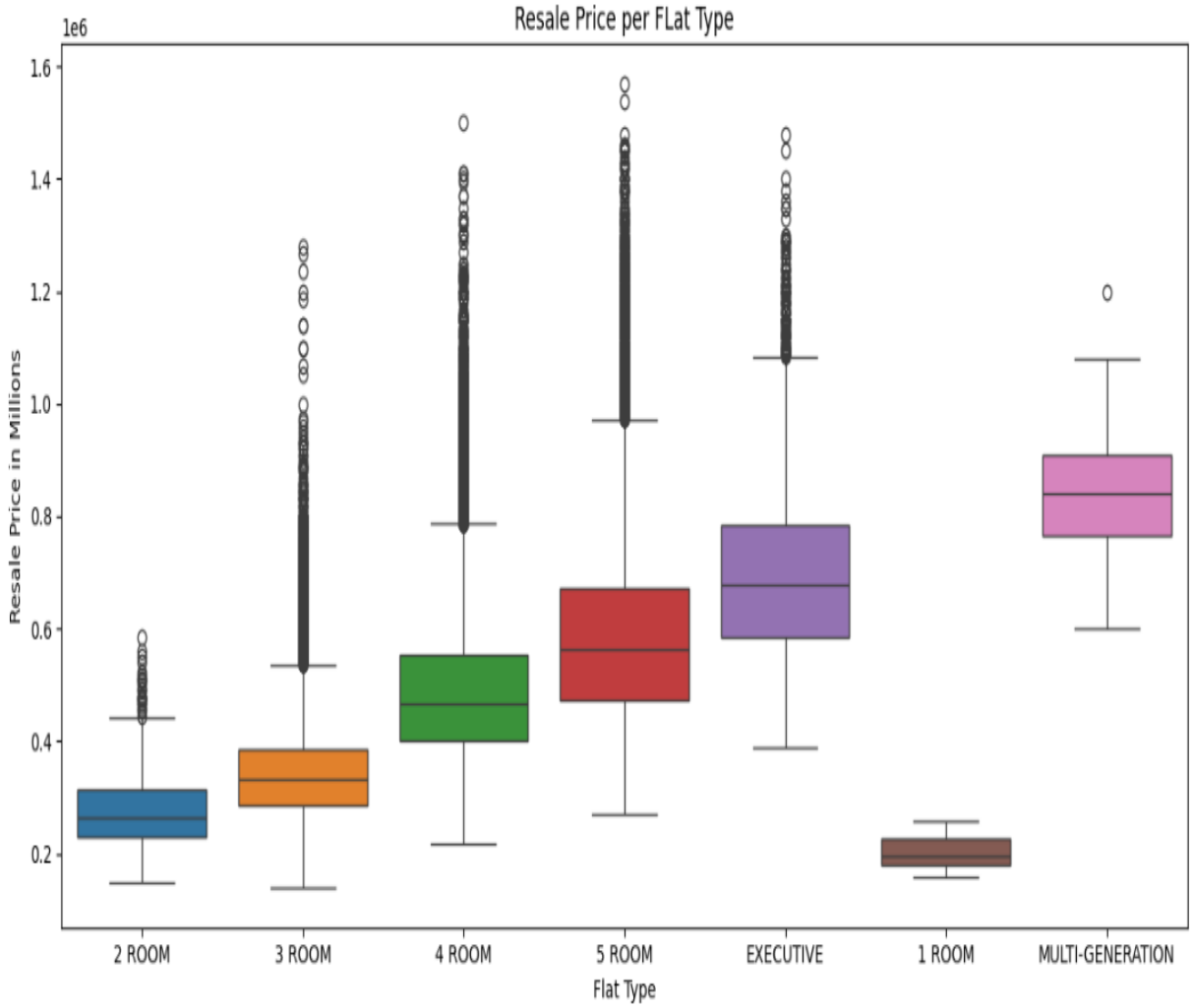
- Πράσινο : Διαμερίσματα 4 δωματίων με εμβαδό από 75 έως 125 τετραγωνικά μέτρα και κυμαινόμενη αξία μεταπώλησης από 200.000$ έως 1.100.000$. Ωστόσο εξαίρεση αποτελούν τα διαμερίσματα των 90 τετραγωνικών μέτρων καθώς, πολλά δείγματα προσεγγίζουν ακόμη και τα 1.500.000$. Τιμή αρκετά μακριά από τον μέσο όρο.

- Κόκκινο : Διαμερίσματα 5 δωματίων με εμβαδό από 115 έως 140 τετραγωνικά μέτρα και αξία παρόμοια με αυτή των πράσινων δειγμάτων.

- Μωβ : Διαμερίσματα τύπου “Executive” 2 έως 3 δωματίων με αρκετά μεγάλη διακύμανση στην επιφάνειά τους καθώς ξεκινάνε από 140 τετραγωνικά μέτρα και υπό εξαιρέσεις αγγίζουν τα 200 τετραγωνικά μέτρα, Η μεταπωλητική τιμή τους κυμαίνεται μεταξύ 500.000$ και 1.000.000$ κατά μέσο όρο.

- Ροζ : Διαμερίσματα τύπου “Multi-generation” 4 δωματίων με πολύ μεγάλη διασπορά στις τιμές της επιφάνειας. Σημειώνονται δείγματα με εμβαδόν που προσεγγίζει τα 250 τετραγωνικά μέτρα αλλά και δείγματα που φτάνουν χαμηλότερα από 150 τετραγωνικά μέτρα. Η μεταπωλητική τους αξία αποκτά κατώφλι στα 800.000$ με μέγιστη τιμή περίπου 1.300.000 δολάρια. Η μεγάλη αυτή διακύμανση του εμβαδού οφείλεται στους διάφορους χώρους που διαθέτει αυτός ο τύπος διαμερίσματος διότι προορίζονται για πολύτεκνες οικογένειες.

Με βάση το διάγραμμα η σχέση που συνδέει τους δύο δείκτες είναι γραμμική, καθώς η τιμή μεταπώλησης αυξάνεται για κάθε τετραγωνικό μέτρο γεγονός που υποδεικνύει την θετική κλίση της γραμμής. Παρόλα αυτά, παρουσιάζεται σημαντική διασπορά των δειγμάτων πέριξ της γραμμής, ιδιαίτερα στα σημεία με μεγάλες επιφάνειες υποδηλώνοντας πως η μεταπωλητική αξία δεν επηρεάζεται μόνο από τον παράγοντα της επιφάνεια. Παράλληλα , παρατηρείται πως υπάρχει γραμμική σχέση τιμής και τύπου διαμερίσματος από την οποία μπορούμε να αντλήσουμε χρήσιμες πληροφορίες μέσω του παρακάτω διαγράμματος boxplot :



Εικόνα 27 : Τιμή ανά τύπο διαμερίσματος

Στο διάγραμμα απεικονίζεται στον κατακόρυφο άξονα η τιμή μεταπώλησης σε εκατομμύρια δολάρια και στον οριζόντιο άξονα οι έξι τύποι διαμερίσματος. Αναλύοντας λεπτομερώς, προκύπτουν οι εξής πληροφορίες :

- Τα διαμερίσματα ενός δωματίου έχουν μικρή διακύμανση στην τιμή τους , της οποίας η διάμεσος είναι περίπου 200.000$, χωρίς να παρατηρείται κάποια ακραία τιμή.

- Τα διαμερίσματα δύο δωματίων συντελούν μια συμπαγή κατανομή με ελάχιστες ακραίες τιμές που φτάνουν τα 600.000$. Η διάμεσος της μεταπωλητικής τιμής είναι περίπου 300.000$.

- Στα διαμερίσματα τριών δωματίων υπάρχουν αρκετές ακραίες τιμές που ξεπερνούν τα 1.200.000$, με την πλειοψηφία αυτών να περιορίζεται κάτω από 1.000.000$. Η διάμεσος είναι περίπου 400.000$.

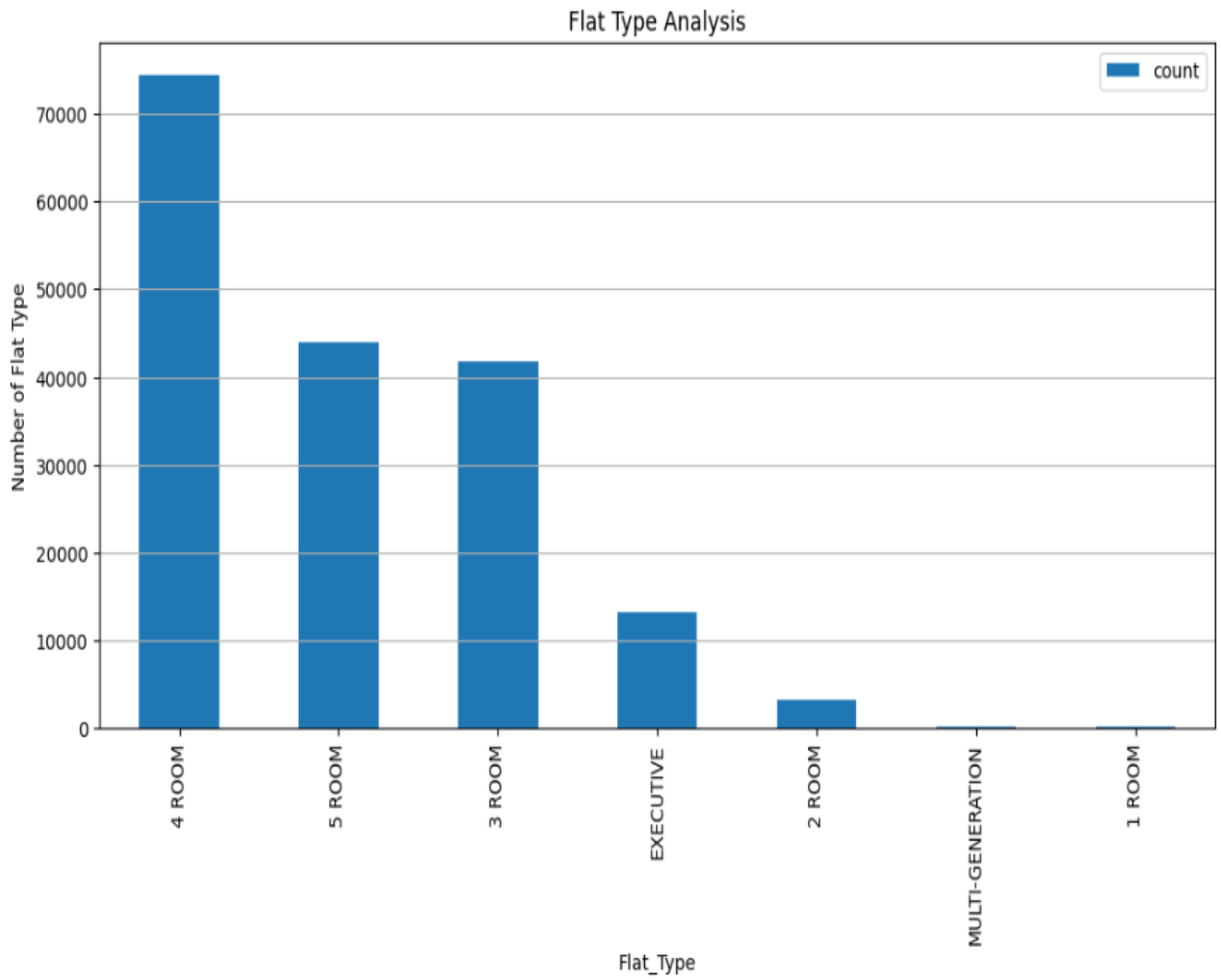
- Τα διαμερίσματα τεσσάρων δωματίων σημειώνουν ακόμη μεγαλύτερη αύξηση στις ακραίες τιμές που προσεγγίζουν τα 1.500.000$, τιμή αρκετά αποστασιοποιημένη από την διάμεσο των 500.000$.

- Παρόμοια τάση ακολουθούν τα διαμερίσματα πέντε δωματίων με διάμεσο ελαφρώς αυξημένη κατά περίπου 100.000$ και ακραίες τιμές που ξεπερνάνε τα 1.500.000$.

- Στα διαμερίσματα τύπου “Executive” παρουσιάζεται ένα μεγάλο εύρος τιμών με διάμεσο κοντά στα 800.000$ και μειωμένο πλήθος ακραίων τιμών.

- Τα διαμερίσματα τύπου “Multi Generation” διαθέτουν μόνο ένα ακραίο σημείο και σχετικά μικρότερη διακύμανση στην τιμή μεταπώλησης η οποία έχει διάμεσο σχεδόν 1.000.000$.

Σύμφωνα με τις παραπάνω αναλύσεις εξάγουμε πολύτιμες πληροφορίες, σχετικά με την διακύμανση στις τιμές των διαμερισμάτων. Για να αποκτηθεί όμως μια ολοκληρωμένη γνώμη σχετικά με την ανεξάρτητη μεταβλητή που αφορά τον τύπο του διαμερίσματος, παρακάτω παρουσιάζεται ένα ραβδόγραμμα που καταγράφει το πλήθος κάθε τύπου διαμερίσματος σε ολόκληρη την βάση δεδομένων.



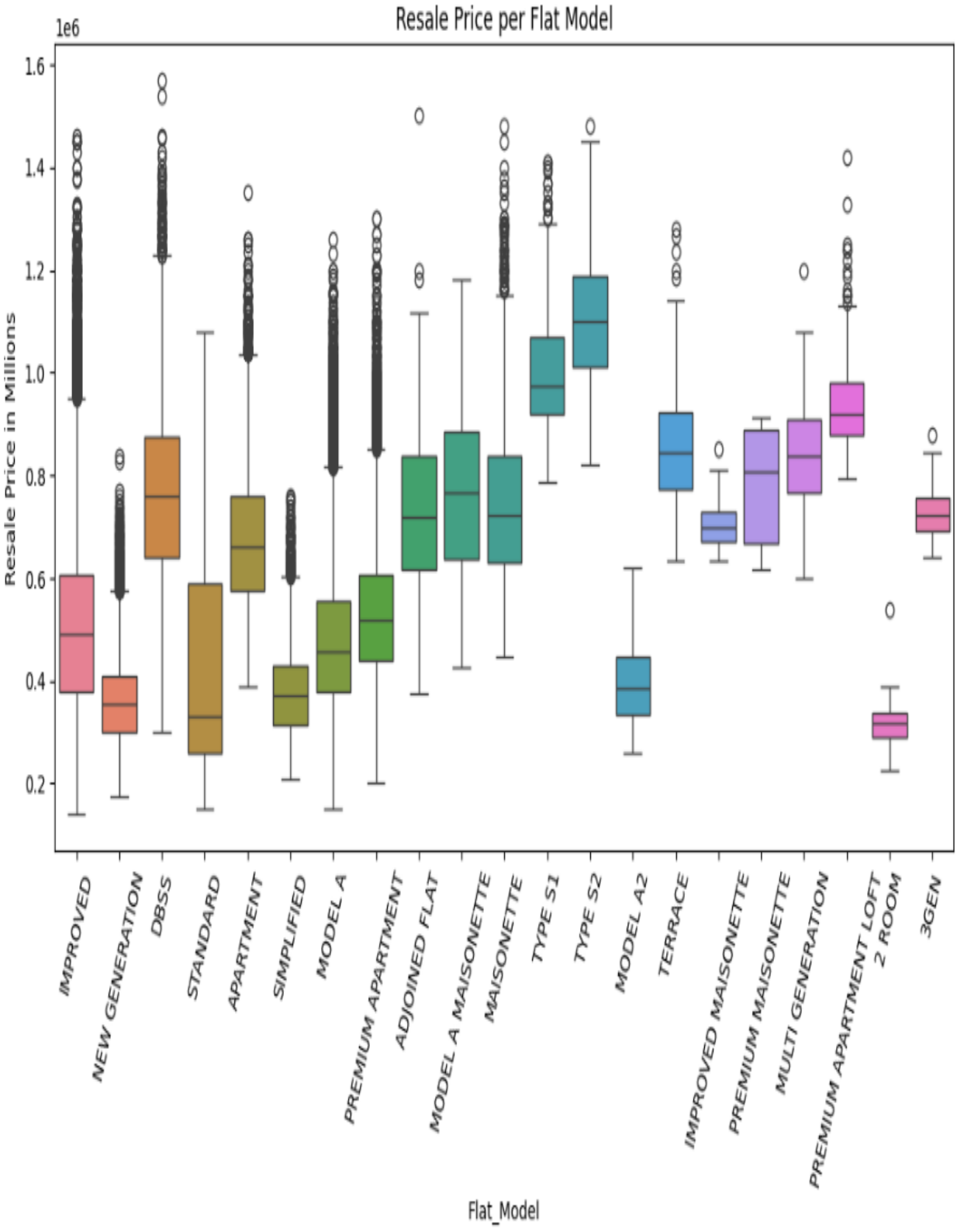
Εικόνα 28 : Πλήθος ανά τύπο διαμερίσματος

Από το παραπάνω διάγραμμα καθίσταται σαφές, πως η συντριπτική πλειοψηφία των διαμερισμάτων υπάγεται στην κατηγορία των τεσσάρων δωματίων με πάνω από 70.000 δείγματα. Ακολουθούν τα διαμερίσματα των πέντε δωματίων με 45.000 δείγματα, ξεπερνώντας ελαφρώς τύπο των τριών δωματίων με 43.000 εμφανίσεις στην βάση δεδομένων. Έπειτα δημιουργείται ένα μεγάλο χάσμα στο πλήθος των δειγμάτων διότι τα διαμερίσματα τύπου “Executive” εμφανίζονται μόλις 10.000 φορές ακολουθούμενα από τον τύπο των δύο δωματίων που σημειώνει μόνο 5.000 δείγματα. Τέλος, σπανίζουν τα διαμερίσματα ενός δωματίου και “Multi Generation” τα οποία δεν ξεπερνάνε τις 50 μονάδες στο διάγραμμα.

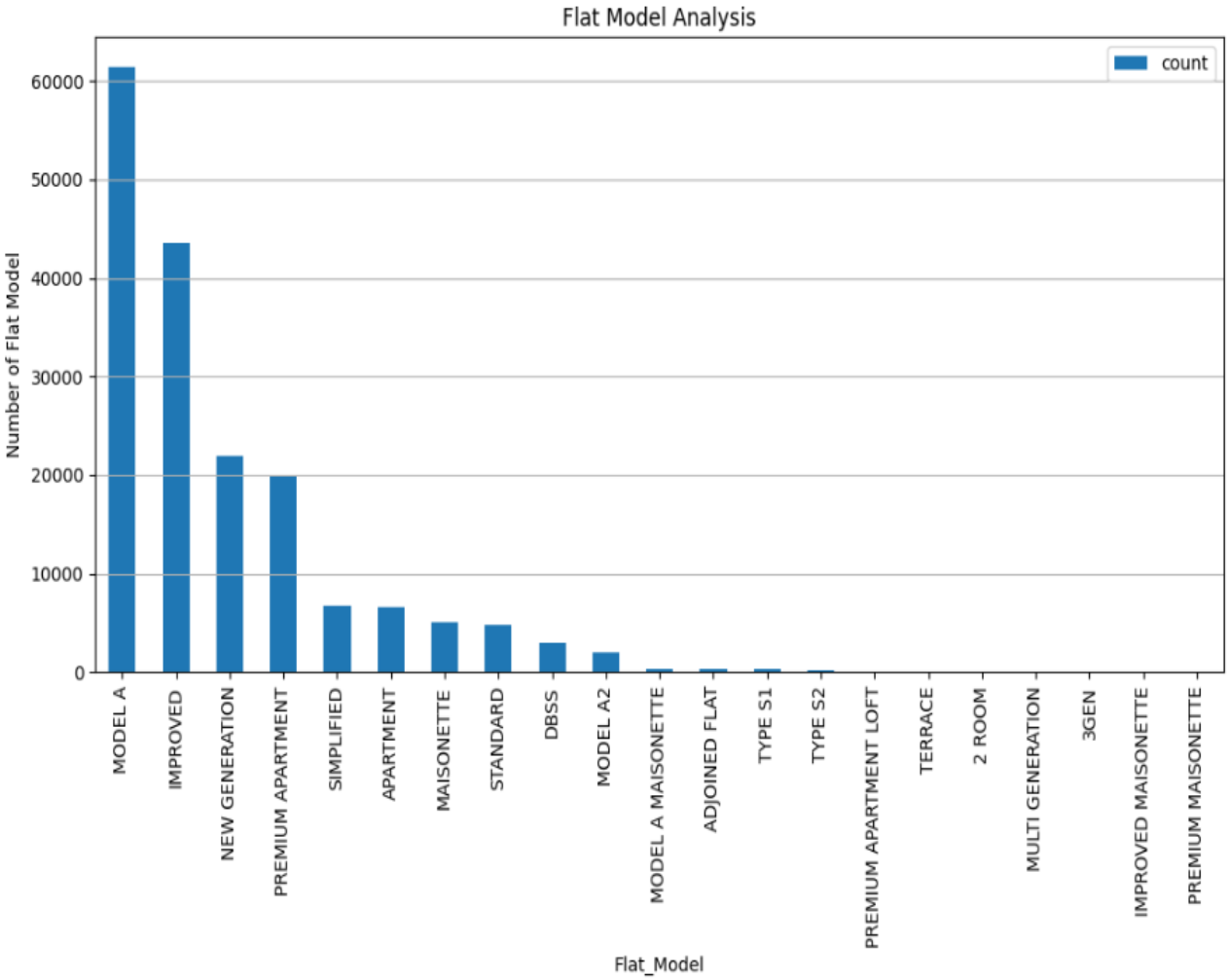
Συνεπώς, συνδυάζοντας τα τρία τελευταία διαγράμματα οδηγούμαστε στα εξής συμπεράσματα για κάθε μία κατηγορία του τύπου διαμερίσματος. Με βάση την δημοτικότητα και τις προτιμήσεις, τα διαμερίσματα τεσσάρων και πέντε δωματίων πρωταγωνιστούν με διαφορά αιτιολογώντας το μεγάλος εύρος της διακύμανσης των μεταπωλητικών τιμών τους, υποδηλώνοντας έτσι πως οι αγοραστές προτιμούν μεγάλους χώρους που είναι πρόσφοροι για την φιλοξενία οικογένειας. Με βάση την οικονομική προσιτότητα τα διαμερίσματα μέχρι και τριών δωματίων με μέτριο μέγεθος είναι τα πιο

προσιτά και ελκυστικά για άτομα που έχουν χαμηλό οικονομικό προϋπολογισμό. Στην συγκεκριμένη περίπτωση ο τύπος των τριών δωματίων είναι ο δημοφιλέστερος, εξασφαλίζοντας μια ισορροπία τιμής και χώρου. Τέλος, αν λάβουμε ως πρώτιστο κριτήριο την πολυτέλεια και την άνεση τότε τα εξειδικευμένα διαμερίσματα τύπου “Executive” και “Multi Generation” καλύπτουν πλήρως τους αγοραστές που είναι πρόθυμοι να δαπανήσουν αρκετά χρήματα για να εξαγοράσουν μεγάλους χώρους με υψηλές παροχές.
Την σκυτάλη της ανάλυσης αναλαμβάνει η ανεξάρτητη μεταβλητή “Flat_Model” η οποία σχετίζεται με το μοντέλο του διαμερίσματος. Δεν πρέπει να συγχέεται με την ανεξάρτητη μεταβλητή “Flat_Type” που αναλύσαμε πριν, καθώς εκείνη αφορά την διαρρύθμιση και την επιφάνεια του διαμερίσματος.
Παρακάτω, παρουσιάζονται δύο διαγράμματα που εξάγουν χρήσιμες πληροφορίες σχετικά με το μοντέλο των διαμερισμάτων.

Το πρώτο διάγραμμα είναι ένα boxplot που περιέχει τα εικοσιένα μοντέλα διαμερισμάτων στον οριζόντιο άξονα, συσχετιζόμενα με την τιμή μεταπώλησης που τοποθετείται στον κατακόρυφο άξονα. Στο δεύτερο διάγραμμα απεικονίζεται ένα ραβδόγραμμα το οποίο αναλύει το πλήθος των διαμερισμάτων ανά μοντέλο διαμερίσματος στον κατακόρυφο και οριζόντιο άξονα αντίστοιχα.



Εικόνα 29 : Τιμή ανά μοντέλο διαμερίσματος



Εικόνα 30 : Πλήθος ανά μοντέλο διαμερίσματος

Στη συνέχεια ακολουθούν οι αναλύσεις όλων των μοντέλων που εμφανίζονται στην βάση δεδομένων :

- Τα “MODEL A” διαμερίσματα είναι το συνηθέστερο μοντέλο που συναντάται στην βάση δεδομένων με περισσότερα από 60.000 δείγματα και διάμεσο τιμής μεταπώλησης πάνω από 0.4 εκατομμύρια δολάρια.

- Τα “IMPROVED” διαμερίσματα αποτελούν μια εξίσου δημοφιλή επιλογή καθώς σημειώνουν πάνω από 40.000 εγγραφές και μεσαία τιμή μεταπώλησης κοντά στα 0.5 εκατομμύρια δολάρια.

- Τα “NEW GENERATION ” διαμερίσματα αγγίζουν τις 22.000 μονάδες έχοντας ως μεσαία τιμή τα 0.35 εκατομμύρια δολάρια.

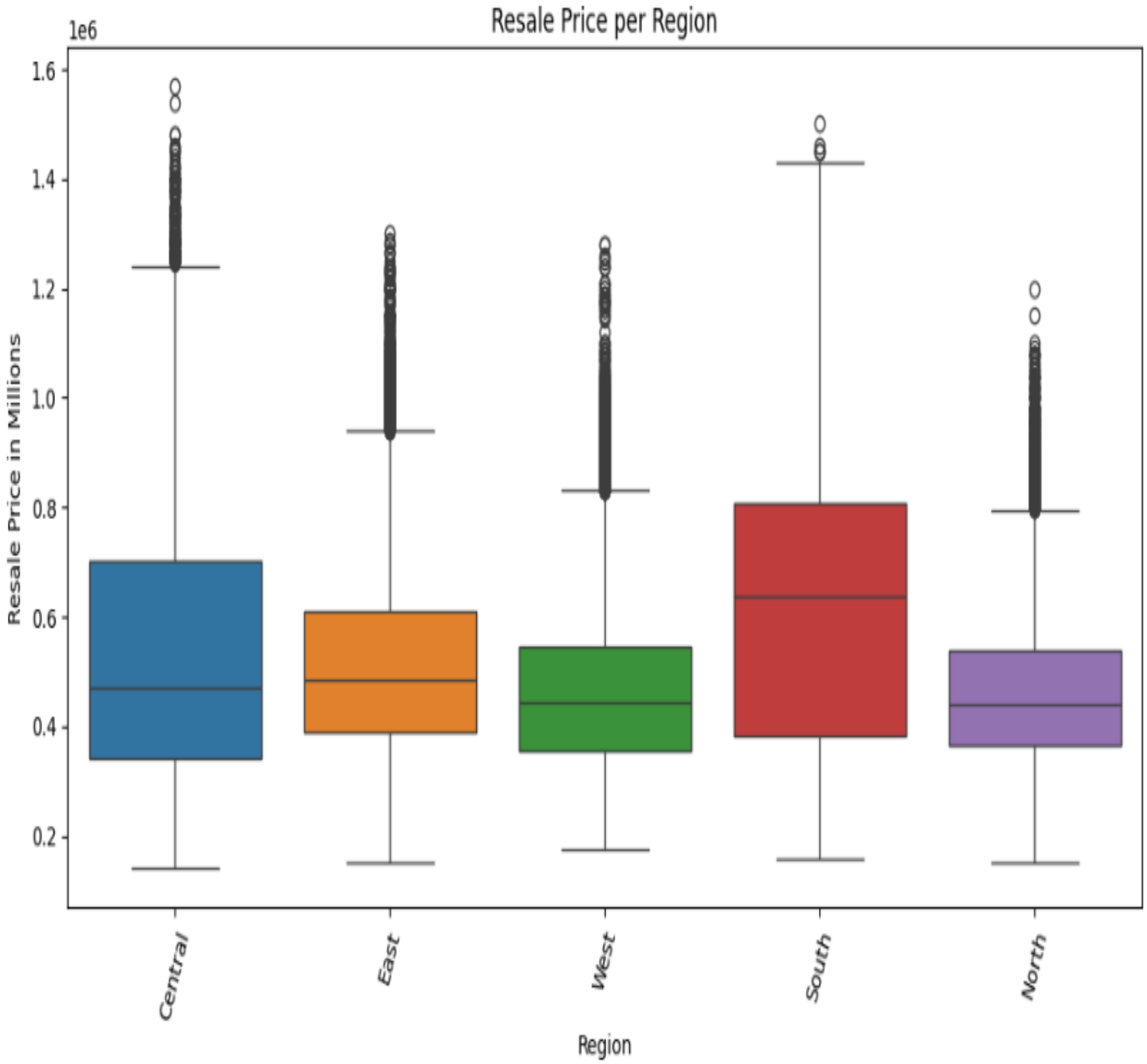
- “PREMIUM APARTMENT” διαμερίσματα είναι το τελευταίο μοντέλο που σημειώνει μεγάλη προτίμηση με μέγεθος που φτάνει τις 20.000 μονάδες. Συναντώνται πολλές ακραίες τιμές που ξεπερνούν τα 1.2 εκατομμύρια δολάρια, ποσό διπλάσιο από την μεσαία τιμή τους.

- Τα μοντέλα “SIMPLIFIED”, “APARTMENT”, “DBSS”, “MAISONETTE” και “STANDARD” εμφανίζονται περίπου 6 έως 4 χιλιάδες φορές στην βάση δεδομένων. Η διάμεση τιμή για τα περισσότερα μοντέλα αγγίζει τα 0.3 έως 0.4 εκατομμύρια δολάρια.

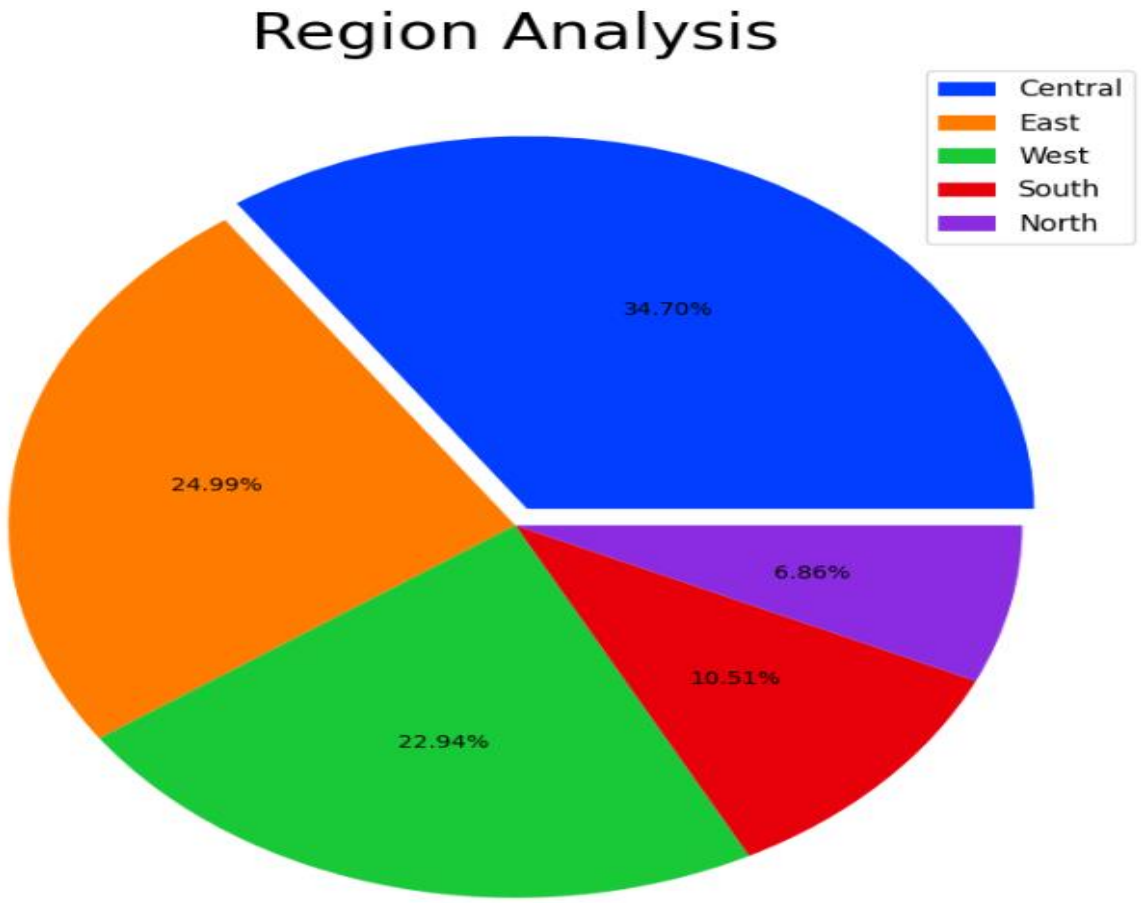
- Αμελητέες μονάδες συγκεντρώνουν τα μοντέλα “MODEL A2”, “MODEL A MAISONETTE”, “ADJOINED FLAT”, “TYPE S1”, “TYPE S2”, “PREMIUM APARTMENT LOFT”, "TERRACE", "2 ROOM", "MULTI GENERATION", "3GEN", "IMPROVED MAISONETTE", "PREMIUM MAISONETTE" με την πλειοψηφία αυτών να σημειώνει μεσαία τιμή μεταπώλησης κυμαινόμενη στα 0.8 εκατομμύρια δολάρια. Οφείλουμε να σημειώσουμε πως σε αυτές τις περιπτώσεις δεν συναντάμε πολλές ακραίες τιμές λόγω του πολύ μικρού αριθμού δειγμάτων.

Συνδυάζοντας τα δύο διαγράμματα παρατηρούμε πως μοντέλα διαμερισμάτων με χαμηλή διαθεσιμότητα και πολύ υψηλές τιμές μεταπώλησης είναι συνήθως πολυτελή, με υψηλή ποιότητα παροχών και καλύτερες ανέσεις. Από την άλλη όψη, διαμερίσματα με τιμές που κυμαίνονται λιγότερο από 0.6 εκατομμύρια δολάρια εμφανίζουν μεγάλη διαθεσιμότητα, υποδηλώνοντας πως τα μοντέλα που παρέχουν τις βασικές ανάγκες χωρίς την προσθήκη πολυτελών χαρακτηριστικών καθίστανται περισσότερο προσιτά για τον μέσο αγοραστή.

Στην συνέχεια θα παρατηρήσουμε πως αντιδρά η μεταπωλητική αξία με βάση την τοποθεσία του διαμερίσματος. Οι ανεξάρτητες μεταβλητές όπως “Town”, “Block”, “Street_Name” παρέχουν πληροφορίες σχετικά με την τοποθεσία. Ωστόσο το μεγάλο πλήθος διακριτών τιμών που διαθέτουν δεν βοηθάει ιδιαίτερα στην ανάλυση. Για αυτό τον λόγο έχουμε δημιουργήσει ήδη μια νέα ανεξάρτητη μεταβλητή “Region” στο κεφάλαιο 5.1, η οποία έχει ομαδοποιημένα τα διαμερίσματα σε πέντε ευρύτερες περιοχές.
Έπειτα ακολουθούν δυο εικόνες που αντιπροσωπεύουν ένα διάγραμμα boxplot και ένα διάγραμμα πίτας οι οποίες προβάλλουν αντίστοιχα, την κατανομή της εμπορικής τιμής ανά περιοχή και την ποσοστιαία κατανομή των διαμερισμάτων ανά περιοχή.



Εικόνα 31 : Εμπορική τιμή ανά περιοχή



Εικόνα 32 : Ποσοστό ανά περιοχή

Όπως έχει αναφερθεί, η ανεξάρτητη μεταβλητή "Region" απαρτίζεται από πέντε διακριτές τιμές οι οποίες αντιπροσωπεύουν ένα τμήμα της πίτας εκάστη. Αναλύοντας τα δεδομένα για κάθε μια περιοχή προκύπτουν οι εξής πληροφορίες :

- Τα διαμερίσματα που ανήκουν στο κέντρο της Σιγκαπούρης αντιπροσωπεύουν το 34,70% της βάσης, καθιστώντας την περιοχή με το μεγαλύτερο ποσοστό. Η διάμεσος της εμπορικής αξίας αγγίζει σχεδόν τα 500.000$, ενώ οι ακραίες τιμές φτάνουν έως 1.600.000$. Παράλληλα υπάρχει μια μικρή ασσυμετρία καθώς τα δεδομένα που είναι μεγαλύτερα από την διάμετρο, έχουν μεγαλύτερη διασπορά σε σύγκριση με τα δεδομένα που είναι μικρότερα.

- Τα διαμερίσματα που ανήκουν στην ανατολική περιοχή καταλαμβάνουν το 24,99% των συνολικών δειγμάτων της βάσης. Η μεσαία τους τιμή είναι λίγο ανώτερη από 450.000$ και οι ακραίες τιμές τους κυμαίνονται από 1.000.000$ έως 1.300.000$.

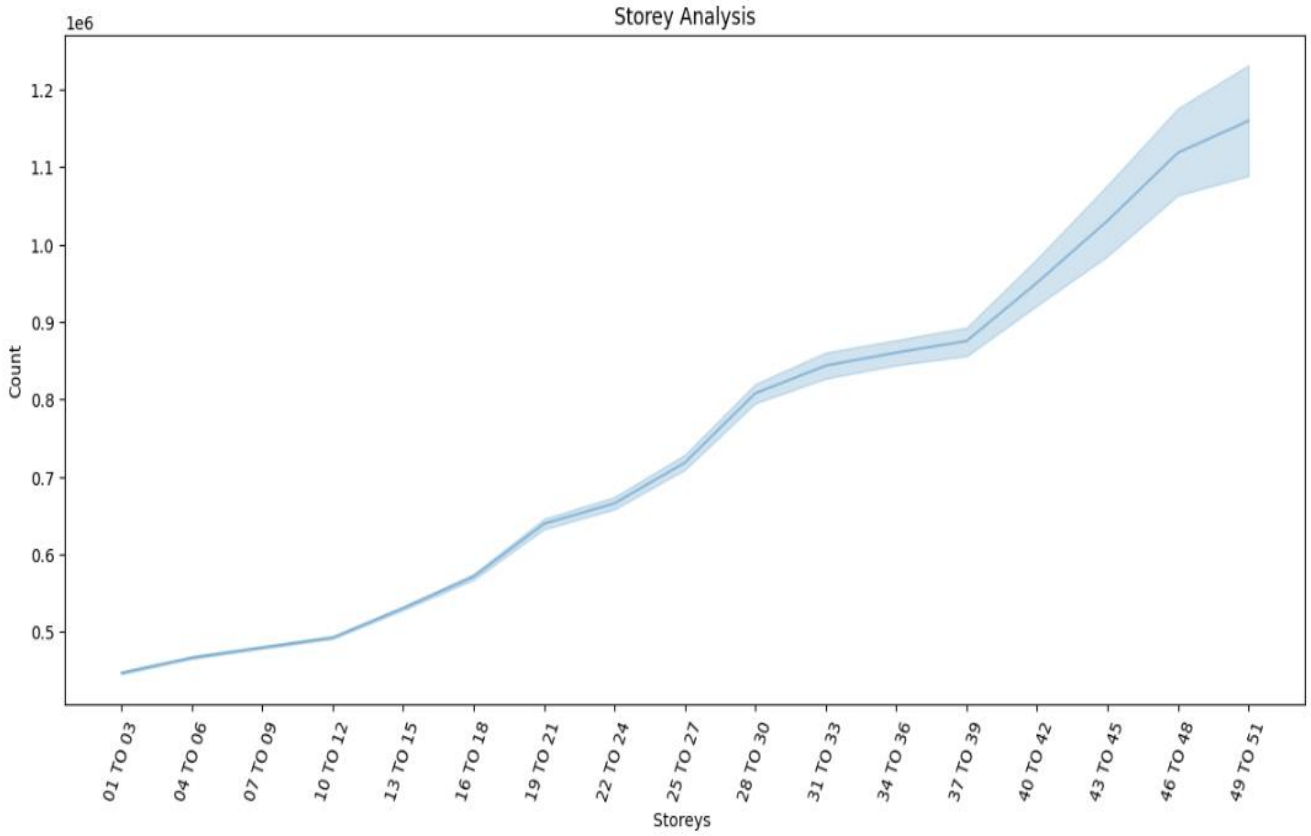
- Η τρίτη δημοφιλέστερη περιοχή είναι η δυτική που αποτελεί μία εξίσου σημαντική περιοχή για κατοικίες με ποσοστό της τάξης του 22,94%.. Οι ακραίες τιμές φτάνουν τα 1.300.000$ και η μεσαία τιμή μεταπώλησης περίπου 450.000$.

- Οι κατοικίες της νότιας περιοχής ανέρχονται στο 10,51% με διάμεσο που φτάνει τα 700.000$. Παρατηρείται επίσης μια μικρή ασσυμετρία αντίθετη με εκείνη της

κεντρικής περιοχής διότι τα δεδομένα που βρίσκονται κάτωθι της διαμέσου προβάλλουν μεγαλύτερη διασπορά από τα δεδομένα που βρίσκονται άνω της διαμέσου. Στην συγκεκριμένη περιοχή εμφανίζονται μόνο τρεις ακραίες τιμές.

- Η περιοχή που καταλαμβάνει το μικρότερο ποσοστό είναι η βόρεια. Με ποσοστό 6,86% την καθιστά λιγότερο δημοφιλή σε σχέση με τις υπόλοιπες. Επιπρόσθετα έχει τις χαμηλότερες τιμές μεταπώλησης, με διάμεσο κάτω από 500.000$ και πολλές ακραίες τιμές που ξεπερνάν το 1.000.000$.

Συνοψίζοντας, με βάση την αξία, οι κατοικίες που βρίσκονται στην νότια και κεντρική περιοχή έχουν την μεγαλύτερη αξία. Ακολουθούν η ανατολική και δυτική περιοχή οι οποίες έχουν ισοδύναμο εύρος τιμών , ακολουθούμενες από την βόρεια περιοχή που έχει την χαμηλότερη αξία. Ωστόσο, αν θέσουμε ως κριτήριο την συγκέντρωση, με μεγάλη διαφορά η κεντρική περιοχή της Σιγκαπούρης διαθέτει υψηλή συγκέντρωση. Ακολουθούν η νότια, η ανατολική και δυτική περιοχή, οι οποίες συνθέτουν μια ισορροπημένη κατανομή μεταξύ τους, και ουραγός στην ιεραρχία είναι ξανά η βόρεια περιοχή, της οποίας η χαμηλή αξία φαίνεται να αντανακλά τη μικρή συγκέντρωση σε αριθμό διαμερισμάτων.
Από το κομμάτι της οπτικοποίησης δεδομένων δε θα μπορούσε να παραλειφθεί η ανάλυση της ανεξάρτητης μεταβλητής “Storey_Range”, που αφορά τους ορόφους. Αναλυτικότερα, στο διάγραμμά που ακολουθεί εκφράζεται η αξία των διαμερισμάτων ανά συγκεκριμένο εύρος ορόφων με τον κατακόρυφο άξονα να αναπαριστά την τιμή μεταπώλησης και τον οριζόντιο άξονα το εκάστοτε εύρος ορόφων.



Εικόνα 33 : Τιμή μεταπώλησης ανά εύρος ορόφων

Με μια πρώτη ματιά φαίνεται ξεκάθαρα η θετική σχέση μεταξύ τιμής και ορόφων καθώς η οπτικοποίηση δείχνει πως η τιμή τείνει να αυξάνεται με την αύξηση των ορόφων. Από τον πρώτο έως και τον εικοστό πρώτο όροφο η αύξηση των τιμών είναι μικρή με ελάχιστη μεταβλητότητα , όπως υποδηλώνεται από την πολύ μικτή σκίαση. Όμως, καθώς αυξάνονται οι όροφοι, ο ρυθμός με τον οποίο η τιμή αυξάνεται, μεγαλώνει έχοντας ως συνέπεια την ύπαρξη μεγάλης μεταβλητότητας, ιδιαίτερα στους τελευταίους δέκα ορόφους. Αυτό υποδηλώνει πως οι μεγαλύτεροι όροφοι έχουν ένα διαφοροποιημένο εύρος τιμών το οποίο στην χειρότερη περίπτωση φτάνει τα 100.000$

Σε αυτό το σημείο οπτικοποιήθηκαν όλες οι ανεξάρτητες μεταβλητές που θα συντελέσουν αργότερα το μοντέλο μηχανικής μάθησης. Μέσα από τα διαγράμματα αντλήθηκαν σημαντικές πληροφορίες που υποδηλώνουν όχι μόνο την σχέση των ανεξάρτητων μεταβλητών με την μεταβλητή στόχο ,αλλά τις μεταξύ τους αλληλεπιδράσεις και τον αντίκτυπο στις προτιμήσεις της αγοράς.

5.3 Επεξεργασία Μεταβλητών

Στο υποκεφάλαιο της επεξεργασίας μεταβλητών θα προετοιμάσουμε την βάση δεδομένων για εκπαίδευση αφαιρώντας τις περιττές στήλες και τις ακραίες τιμές. Παράλληλα θα τροποποιηθούν κατάλληλα οι τύποι των ανεξάρτητων μεταβλητών, ώστε να αποκτήσουν αριθμητικές τιμές σύμφωνα με τους κανόνες που θα θέσουμε.

Αρχικά, θα αξιοποιήσουμε την ανεξάρτητη μεταβλητή “Flat_Type”, η οποία έχει 7 διαφορετικές αλφαριθμητικές τιμές. Η κωδικοποίηση θα γίνει σύμφωνα με την εικόνα 26, με την νέα ανεξάρτητη μεταβλητή “Rooms” να λαμβάνει ακέραιες τιμές από 1 έως και 6, αντιπροσωπεύοντας η κάθε μία, τον αριθμό των δωματίων. Συνεπώς οι αλφαριθμητικές τιμές της “Flat_Type” : “EXECUTIVE” και “MULTI GENERATION” θα αντικατασταθούν από τον αριθμό 6, καθώς σύμφωνα με την εικόνα 26 και τις πληροφορίες της πηγής [22].

```
Transformed Room column unique values...
[2 3 4 5 6 1]
```

Εικόνα 34 : Τιμές μεταβλητής "Rooms"

Στη συνέχεια, θα τροποποιηθεί η μεταβλητή “Flat_Model” η οποία διαθέτει 22 διαφορετικές αλφαριθμητικές μεταβλητές. Αυτό θα συμβεί διότι πολλά μοντέλα διαμερισμάτων έχουν αρκετές ομοιότητες, τις οποίες θα αξιοποιήσουμε για να δημιουργήσουμε 6 ομάδες μοντέλων, όπου κάθε μια θα περικλείει τα μοντέλα διαμερισμάτων που είναι σχετικά όμοια. Με αυτόν το τρόπο, η μεταβλητή “Flat_Model”, διαθέτει 18 λιγότερες διαφορετικές τιμές με συνέπεια την καλύτερη εκπαίδευση του μοντέλου.

```
Transformed Flat Model column unique values...
['Standard' 'New Generation' 'Exclusive' 'Apartment' 'A model'
 'Maisonette']
```

Εικόνα 35 : Τιμές μεταβλητής "Flat_Model"

Τέλος , θα δημιουργηθεί μια νέα ανεξάρτητη μεταβλητή “Storey_Level” η οποία απορρέει από την ανεξάρτητη μεταβλητή “Storey_Range”. Η νέα αυτή μεταβλητή θα διαθέτει μόνο 3 διακριτές τιμές που εκφράζουν πόσο ψηλά βρίσκεται το διαμέρισμα, διότι η ανεξάρτητη μεταβλητή “Storey_Range” διαθέτει 21 διακριτές τιμές, γεγονός που καθιστά το μοντέλο περισσότερο περίπλοκο.

```
Transformed Storey Range column into Storey Level, unique values...
['Medium' 'High' 'Low']
```

Εικόνα 36 : Τιμές μεταβλητής "Storey_Level"

Αφού κατασκευάσαμε τις νέες ανεξάρτητες μεταβλητές, θα διαγράψουμε τις αρχικές μεταβλητές, με την βάση δεδομένων να παίρνει την εξής μορφή :

Daraframe :

	Year	Region	Rooms	Flat_Model	Floor_Area_sqm	Storey_Level	Lease_Commence_Date	Resale_Price
0	2017	Central	2	Standard	44.0	Medium	1979	232000.0
1	2017	Central	3	New Generation	67.0	High	1978	250000.0
2	2017	Central	3	New Generation	67.0	High	1980	262000.0
3	2017	Central	3	New Generation	68.0	Low	1980	265000.0
4	2017	Central	3	New Generation	67.0	High	1980	265000.0
...	...	...	...	...	...	...	...	...
176971	2024	North	5	Standard	112.0	Medium	2018	725000.0
176972	2024	North	5	Standard	121.0	Medium	1987	668888.0
176973	2024	North	6	Apartment	181.0	Medium	1992	1080000.0
176974	2024	North	6	Apartment	146.0	Low	1988	830000.0
176975	2024	North	6	Maisonette	146.0	Low	1988	880000.0

Εικόνα 37 : Τροποποιημένη βάση δεδομένων

Όπως φαίνεται από την εικόνα, η βάση έχει πλέον 8 μεταβλητές στο σύνολο, 3 λιγότερες από την αρχική της μορφή. Η μεταβλητή "Storey_Range" αντικαταστάθηκε από την μεταβλητή "Storey_Level", η μεταβλητή "Flat_Type" αντικαταστάθηκε από την "Rooms", η μεταβλητή "Remaining_Lease" διαγράφηκε καθώς εξάχθηκε το συμπέρασμα πως όλα τα διαμερίσματα που ανήκουν στην Κυβερνητική Υπηρεσία Σιγκαπούρης, μισθώνονται για 100 χρόνια. Συνεπώς η παρούσα μεταβλητή δεν παρέχει σημαντικές πληροφορίες, με αποτέλεσμα να αφαιρεθεί.

Έπειτα, οι μεταβλητές που κατασκευάσαμε στο κεφάλαιο 5.1 , "Year" και "Region" πρόκειται να αντικαταστήσουν τις μεταβλητές "Month" και "Town", "Block", "Street_Name" αντίστοιχα.

Η εκπαίδευση ενός μοντέλου μηχανικής μάθησης που προβλέπει συνεχείς τιμές, προϋποθέτει την ύπαρξη ανεξάρτητων μεταβλητών με αριθμητικές τιμές. Αναλυτικότερα , όλες κατηγορικές τιμές θα αντιστοιχηθούν με αριθμητικές τιμές. Οι μεταβλητές "Flat_Model" και "Storey_Level", θα λάβουν αριθμητικές τιμές σύμφωνα με το μέγεθος που εκφράζει η κατηγορική τους τιμή. Η μεταβλητή "Region" ,θα κωδικοποιηθεί με την τεχνική "One-Hot" , δημιουργώντας έτσι 5 νέες στήλες, που δηλώνουν εάν το συγκεκριμένο διαμέρισμα ανήκει στην εκάστοτε περιοχή. Πλέον, με την ύπαρξη των 5 στηλών, η μεταβλητή "Region" θα αφαιρεθεί, για να υπάρχουν μόνο αριθμητικές τιμές.

Encoded data without scaling...

	Year	Rooms	Flat_Model	Floor_Area_sqm	Storey_Level	Lease_Commence_Date	Resale_Price	Region_Central	Region_East	Region_North	Region_South	Region_West
0	2017	2	4	44.0	1	1979	232000.0	1	0	0	0	0
1	2017	3	0	67.0	2	1978	250000.0	1	0	0	0	0
2	2017	3	0	67.0	2	1980	262000.0	1	0	0	0	0
3	2017	3	0	68.0	0	1980	265000.0	1	0	0	0	0
4	2017	3	0	67.0	2	1980	265000.0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
176971	2024	5	4	112.0	1	2018	725000.0	0	0	1	0	0
176972	2024	5	4	121.0	1	1987	668888.0	0	0	1	0	0
176973	2024	6	3	181.0	1	1992	1080000.0	0	0	1	0	0
176974	2024	6	3	146.0	0	1988	830000.0	0	0	1	0	0
176975	2024	6	5	146.0	0	1988	880000.0	0	0	1	0	0

Εικόνα 38 : Κωδικοποιημένη βάση δεδομένων

Στο επόμενο βήμα θα χρησιμοποιηθεί η τεχνική της τυποποίησης για να προσαρμοστούν οι τιμές των ανεξάρτητων μεταβλητών και να αποφευχθούν οι τιμές μεγάλης κλίμακας. Για την τυποποίηση των δεδομένων χρησιμοποιήθηκε βιβλιοθήκη scikit-learn.

Scaled data...

	Lease_Commence_Date	Floor_Area_sqm	Year	Rooms	Flat_Model	Storey_Level	Resale_Price	Region_Central	Region_East	Region_North	Region_South	Region_West
0	-1.211795	-2.214295	-1.622632	-2.316164	1.198147	-0.187470	232000.0	1	0	0	0	0
1	-1.283195	-1.257101	-1.622632	-1.227108	-1.384194	1.122269	250000.0	1	0	0	0	0
2	-1.140395	-1.257101	-1.622632	-1.227108	-1.384194	1.122269	262000.0	1	0	0	0	0
3	-1.140395	-1.215484	-1.622632	-1.227108	-1.384194	-1.497209	265000.0	1	0	0	0	0
4	-1.140395	-1.257101	-1.622632	-1.227108	-1.384194	1.122269	265000.0	1	0	0	0	0

Εικόνα 39 : Τυποποιημένη βάση δεδομένων

Στο τελευταίο κομμάτι της επεξεργασίας των μεταβλητών , θα αφαιρέσουμε τις ακραίες τιμές με την μέθοδο IQR.

```
The percentage of outliers removed from the  dataframe is: 22.941%

Dataset shape before IQR methon : (176693, 12)

Dataset shape after IQR methon : (136157, 12)

Insances deleted : 40536
```

Εικόνα 40 : Ακραίες τιμές

Η μέθοδος αφαίρεσε 40.536 τιμές, αριθμός που αποτελεί σχεδόν το 23% της βάσης. Ύστερα από την ολοκλήρωση αυτού του βήματος, η βάση είναι έτοιμη να τροφοδοτήσει με δεδομένα το επιβλεπόμενο μοντέλο που θα κατασκευάσουμε.

5.4 Υλοποίηση Μοντέλων

Οι αλγόριθμοι που επιλέχθηκαν, βρίσκονται στην βιβλιοθήκη Scikit_Learn. Αρχικά, μέσω της συνάρτησης train_test_split, το σύνολο δεδομένων της βάσης χωρίστηκε σε δεδομένα εκπαίδευσης και σε δεδομένα τεστ. Με αυτόν τον τρόπο, οι αλγόριθμοι των μοντέλων θα εκπαιδευτούν πάνω στα δεδομένα εκπαίδευσης και έπειτα θα προσπαθήσουν να προβλέψουν την τιμή της μεταβλητής στόχου (Resale_Price) στο σύνολο δεδομένων τεστ.

Στην συνέχεια, έχοντας δημιουργήσει ένα αντίγραφο του συνόλου δεδομένων, εφαρμόζουμε τυποποίηση και υπερ-παραμετροποίηση. Για κάθε ένα μοντέλο θέτουμε διαφορετικές τιμές στις παραμέτρους , μέσω ενός πλέγματος που ορίζει σε κάθε μοντέλο τις καταλληλότερες τιμές, για να πετύχουμε μεγαλύτερη απόδοση.

5.4.1 Ridge Regression

Ο πρώτος αλγόριθμος είναι η παλινδρόμηση Ridge, με τις βασικότερες αρχικές τιμές παραμέτρων να είναι οι εξής:

- alpha: 1.0. Μη αρνητική σταθερά που πολλαπλασιάζει τον όρο L2 και ελέγχει την επιρροή της κανονικοποίησης.

- max_iter: None. Ο μέγιστος αριθμός των επαναλήψεων που χρειάζονται για την σύγκλιση των λυτών.

- tol: 1e-4. Παράμετρος που αντιπροσωπεύει την ανεκτικότητα στα κριτήρια τερματισμού.

- copy_X: True. Παράμετρος που καθορίζει αν τα αρχικά δεδομένα εισόδου θα αντιγραφούν κατά την εκπαίδευση ή θα τροποποιηθούν στη συνέχεια.

- solver: ‘auto’. Ο λύτης που αξιοποιείται για τις υπολογιστικές ρουτίνες.

Με τις αρχικές τιμές παραμέτρων το μοντέλο εμφανίζει τα παρακάτω αποτελέσματα :

```
MSE without Stardadization and Hypetuning: 8035000920.669171
RMSE without Stardadization and Hypetuning: 89638.16665165107
MAE without Stardadization and Hypetuning: 68075.91397889901
R-squared without Stardadization and Hypetuning:0.74174
```

Εικόνα 41: Αρχικά αποτελέσματα παλινδρόμησης Ridge.

Η διαδικασία υπολογισμού διήρκησε 0,3 δευτερόλεπτα. Σύμφωνα με την μετρική R-squared, φαίνεται πως το μοντέλο δεν προσαρμόζεται πολύ καλά στα δεδομένα , επιτυγχάνοντας ένα ποσοστό 74,17%.

Η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) που δείχνει την απόκλιση των πραγματικών τιμών μεταπώλησης με τις τιμές μεταπώλησης που προβλέπει το μοντέλο, ισούται με 89638 δολάρια.

Τέλος το μέσο απόλυτο σφάλμα (MAE) ισούται με 68076 δολάρια.

Ακολουθεί η διαδικασία της παραμετροποίησης και η δημιουργία του πλέγματος με τις πιθανές τιμές που μπορεί να λάβει η παλινδρόμηση Ridge.

```
# Ridge Regresson with Stardadization and Hypetuning
parameters = {'alpha' : [1.0, 1.5, 2.0, 2.5],
              'max_iter' : [400,500,600,700,800]}
```

Εικόνα 42: Πλέγμα τιμών παραμέτρων

Οι βέλτιστες τιμές παραμέτρων που επιλέχθηκαν είναι οι εξής :

```
The best estimator across ALL searched params:
Ridge(alpha=2.5, max_iter=400)
```

Εικόνα 43 : Βέλτιστες τιμές παλινδρόμησης Ridge.

Χρόνος υπολογισμού : 5,4 δευτερόλεπτα.

Αφού πραγματοποιήθηκε υπερ-παραμετροποίηση, εφαρμόστηκε το νέο μοντέλο στο τυποποιημένο σύνολο δεδομένων χωρίς να σημειώνεται κάποια βελτίωση στην απόδοσή του.

```
Mean MSE with Stardadization and Hypetuning: 8034959729.320453
Mean RMSE with Stardadization and Hypetuning: 89637.93688679171
Mean MAE with Stardadization and Hypetuning: 68075.13561721808
Mean R-squared with Stardadization and Hypetuning:0.74175
```

Εικόνα 44 : Αποτελέσματα έπειτα από υπερ-παραμετροποίηση και τυποποίηση.

5.4.2 Lasso Regression

Ο δεύτερος αλγόριθμος που επιλέχθηκε είναι η παλινδρόμηση Lasso με τις βασικότερες προκαθορισμένες παραμέτρους να είναι οι εξής :

- alpha: 1.0. Μη αρνητική σταθερά που πολλαπλασιάζει τον όρο L2 και ελέγχει την επιρροή της κανονικοποίησης.

- tol: 1e-4. Παράμετρος που αντιπροσωπεύει την ανεκτικότητα στα κριτήρια τερματισμού.

- max_iter: 1000. Ο μέγιστος αριθμός των επαναλήψεων που χρειάζονται για την σύγκλιση των λυτών.

- selection: ‘cyclic’. Καθορίζει τον τρόπο με τον οποίο ενημερώνονται οι συντελεστές κατά την βελτιστοποίηση. Στην συγκεκριμένη περίπτωση, ενημερώνονται με κυκλική σειρά μέχρι να συγκλίνει ο αλγόριθμος.

Με τις αρχικές τιμές παραμέτρων το μοντέλο εφαρμόστηκε στο μη τυποποιημένο σύνολο δεδομένων εμφανίζοντας τα εξής αποτελέσματα :

```
Mean MSE without Stardadization and Hypetuning: 8035004378.351105
Mean RMSE without Stardadization and Hypetuning: 89638.1859385335
Mean MAE without Stardadization and Hypetuning: 68076.02857866902
Mean R-squared without Stardadization and Hypetuning:0.74174
```

Εικόνα 45: Αρχικά αποτελέσματα παλινδρόμησης Lasso.

Ο υπολογισμός διήρκησε 4,8 δευτερόλεπτα, με την προσαρμοστικότητα του μοντέλου στα δεδομένα να προσεγγίζει τα 74,17%.
Παράλληλα, η ρίζα του μέσου τετραγωνικού σφάλματος ισούται με 89638 δολάρια. Ενώ η απόλυτη διαφορά μεταξύ πραγματικών τιμών και τιμών που προβλέφθηκαν αγγίζει τα 68076 δολάρια.

Ακολουθεί η διαδικασία της παραμετροποίησης και η δημιουργία του πλέγματος με τις πιθανές τιμές που μπορεί να λάβει η παλινδρόμηση Lasso.

```
# Lasso Regresson with Stardadization and Hypetuning
parameters = { 'alpha' : [1.0, 1.5, 2.0, 2.5],
               'max_iter' : [1500, 2000, 2500]}
```

Εικόνα 46: Πλέγμα τιμών παραμέτρων.

Οι βέλτιστες τιμές που επιλέχθηκαν είναι οι εξής :

```
The best estimator across ALL searched params:
Lasso(alpha=2.5, max_iter=1500)
```

Εικόνα 47: Βέλτιστες τιμές παλινδρόμησης Lasso.

Έπειτα από τυποποίηση, με μέγιστο όριο επαναλήψεων ‘max_iter=1500’ και ‘alpha=2.5’ , το μοντέλο δεν παρουσιάζει βελτίωση , με τα αποτελέσματα των μετρικών να παραμένουν ίδια.

```
Mean MSE with Stardadization and Hypetuning: 8034958467.280991
Mean RMSE with Stardadization and Hypetuning: 89637.9298471411
Mean MAE with Stardadization and Hypetuning: 68074.9831630295
Mean R-squared with Stardadization and Hypetuning:0.74175
```

Εικόνα 48: Αποτελέσματα έπειτα από υπερ-παραμετροποίηση και τυποποίηση.

5.4.3 Random Forest

Ο τρίτος αλγόριθμος που παρουσιάζεται ως λύση είναι ο αλγόριθμος των τυχαίων δασών με τις βασικότερες παραμέτρους να είναι οι εξής:

- n_estimators: 100. Το πλήθος των δένδρων.
- criterion: ‘gini’. Κριτήριο με το οποίο καθορίζεται η αξιολόγηση της ποιότητας των διαχωρισμών κατά την εκπαίδευση του δένδρου.
- max_depth: None. Μέγιστο βάθος δένδρου. Σε περίπτωση που δεν εισαχθεί τιμή, οι κόμβοι αναπτύσσονται μέχρι να είναι όλα τα φύλλα καθαρά ή έως ότου κάθε φύλλο να έχει αριθμό δειγμάτων, μικρότερο από min_samples_split.
- min_samples_split: 2. Αφορά τον ελάχιστο δυνατό αριθμό δειγμάτων που απαιτείται για να πραγματοποιηθεί διαχωρισμός του εσωτερικού κόμβου.
- min_samples_leaf: 1. Παράμετρος που καθορίζει τον ελάχιστο αριθμό δειγμάτων σε κάθε φύλλο του δένδρου.

Με τις αρχικές τιμές παραμέτρων το μοντέλο εφαρμόστηκε στο μη τυποποιημένο σύνολο δεδομένων εμφανίζοντας σε χρόνο 33,4 δευτερολέπτων τα εξής αποτελέσματα :

```
MSE without Stardadization and Hypetuning: 2650439907.700197
RMSE without Stardadization and Hypetuning: 51482.42328892646
MAE without Stardadization and Hypetuning: 34535.29238508548
R-squared without Stardadization and Hypetuning:0.91481
```

Εικόνα 49 : Αρχικά αποτελέσματα τυχαίων δασών.

Σύμφωνα με τα αποτελέσματα το μοντέλο φαίνεται να προσαρμόζεται κατά 91,4% στα δεδομένα με την ρίζα του μέσου τετραγωνικού σφάλματος να ανέρχεται στα 51482 δολάρια. Ενώ το μέσο απόλυτο σφάλμα ισούται με 34535 δολάρια.

Ακολουθεί η διαδικασία της παραμετροποίησης και η δημιουργία του πλέγματος με τις πιθανές τιμές που μπορεί να λάβει ο αλγόριθμος των τυχαίων δασών

```
#Random Forest with Standardization and HyperParameter Tuning
parameters = {'n_estimators' : [500,1000,1500],
              'max_depth' : [None],
              'min_samples_split' : [20],
              'min_samples_leaf' : [2],
              'max_features' : ['sqrt','log2',None]}
```

Εικόνα 50 : Πλέγμα τιμών παραμέτρων.

Ο βέλτιστος συνδυασμός παραμέτρων που προκύπτει ύστερα από 83 λεπτά είναι :

```
The best estimator across ALL searched params:
RandomForestRegressor(max_features=None, min_samples_leaf=2,
                      min_samples_split=20, n_estimators=1000)
```

Εικόνα 51 : Βέλτιστες τιμές παραμέτρων του αλγορίθμου τυχαίων δασών.

Ύστερα από την επιλογή των βέλτιστων παραμέτρων και μετά από τυποποίηση των ανεξαρτήτων μεταβλητών , το μοντέλο παρουσιάζει :

```
Mean MSE with Stardadization and Hypetuning: 2642107097.1580563
Mean RMSE with Stardadization and Hypetuning: 51401.43088629009
Mean MAE with Stardadization and Hypetuning: 35144.62581621391
Mean R-squared with Stardadization and Hypetuning:0.91508
```

Εικόνα 52: Αποτελέσματα έπειτα από υπερ-παραμετροποίηση και τυποποίηση.

Διακρίνεται μια αμελητέα αύξηση στην προσαρμοστικότητα του μοντέλου κατά 0,1% σε σχέση με το αρχικό μοντέλο. Παράλληλα το μέσο απόλυτο σφάλμα αυξήθηκε κατά 600 περίπου δολάρια και η ρίζα του μέσου τεραγονικού σφάλματος μειώθηκε κατά 80 δολάρια. Ο υπολογισμός διήρκησε 3,40 λεπτά.

5.4.4 Decision Trees

Ο τελευταίος αλγόριθμος που επιλέχθηκε είναι τα δέντρα απόφασης. Μερικές από τις σημαντικότερες παραμέτρους εμφανίζονται παρακάτω :

- criterion: 'squared_error'. Κριτήριο με το οποίο καθορίζεται η αξιολόγηση της ποιότητας των διαχωρισμών κατά την εκπαίδευση του δένδρου.

- max_depth: None. Μέγιστο βάθος δένδρου. Σε περίπτωση που δεν εισαχθεί τιμή, οι κόμβοι αναπτύσσονται μέχρι να είναι όλα τα φύλλα καθαρά ή έως ότου κάθε φύλλο να έχει αριθμό δειγμάτων, μικρότερο από min_samples_split.

- min_samples_split: 2. Αφορά τον ελάχιστο δυνατό αριθμό δειγμάτων που απαιτείται για να πραγματοποιηθεί διαχωρισμός του εσωτερικού κόμβου.

- min_samples_leaf: 1. Παράμετρος που καθορίζει τον ελάχιστο αριθμό δειγμάτων σε κάθε φύλλο του δένδρου.

- max_features : None. Ο αριθμός των στοιχείων που πρέπει να ελεγχθούν για πραγματοποιηθεί ο διαχωρισμός ενός κόμβου σε ένα δένδρο.

Τα αποτελέσματα που προέκυψαν με βάση τις προκαθορισμένες τιμές παραμέτρων δείχνουν πως το μοντέλο προσαρμόζεται καλά στα δεδομένα, πετυχαίνοντας ποσοστό 90,2%. Επίσης, το μέσο απόλυτο σφάλμα ισούται με 36306 δολάρια και η ρίζα του τετραγωνικού σφάλματος προσεγγίζει τα 54943 δολάρια. Ο υπολογισμός διήρκησε 1,7 δευτερόλεπτα.

```
MSE without Stardadization and Hypetuning: 3018701990.550933
RMSE without Stardadization and Hypetuning: 54942.7155367382
MAE without Stardadization and Hypetuning: 36306.84978831339
R-squared without Stardadization and Hypetuning:0.90297
```

Εικόνα 53: Αρχικά αποτελέσματα δένδρων απόφασης.

Έπειτα, το πλέγμα δημιουργήθηκε το πλέγμα με τις επιτρεπόμενες τιμές :

```
# Decision Tree with Standardization and HyperParameter Tuning

parameters = {'splitter' : ['best', 'random'],
              'max_depth' : [15,20,22,25],
              'min_samples_split' : [20,25,30],
              'min_samples_leaf' : [2,3,4],
              'max_features' : ['sqrt','log2',None]}
```

Εικόνα 54 : Πλέγμα τιμών παραμέτρων.

Προκύπτει πως ο καλύτερος συνδυασμός είναι :

```
The best estimator across ALL searched params:
DecisionTreeRegressor(max_depth=20, min_samples_leaf=2, min_samples_split=20)
```

Εικόνα 55: Βέλτιστες τιμές παραμέτρων του αλγορίθμου των δένδρων απόφασης.

Τέλος , με τις βέλτιστες τιμές που δόθηκαν στις παραμέτρους δεν φαίνεται να υπάρχει κάποια βελτίωση στην απόδοση του μοντέλου, με τον χρόνο υπολογισμού να ισούται με 0,7 δευτερόλεπτα.

```
MSE with Stardadization and Hypetuning: 3030409630.1768966
RMSE with Stardadization and Hypetuning: 55049.156489240566
MAE with Stardadization and Hypetuning: 36362.45737007994
R-squared with Stardadization and Hypetuning:0.90260
```

Εικόνα 56 : Αποτελέσματα έπειτα από υπερ-παραμετροποιηση και τυποποίηση

6.Σύγκριση αποτελεσμάτων των μοντέλων

Στους παρακάτω πίνακες απεικονίζονται συγκεντρωτικά τα αποτελέσματα των μοντέλων, με βάση την μετρική αξιολόγησης. Με πράσινο χρώμα σημειώνεται το μοντέλο που πέτυχε την καλύτερη απόδοση σε κάθε κατηγορία.

ΜΈΣΟ ΤΕΤΡΑΓΩΝΙΚΌ ΣΦΑΛΜΑ				
	Παλινδρόμηση Ridge	Παλινδρόμηση Lasso	Τυχαία Δάση	Δένδρα Απόφασης
MSE με αρχικές τιμές παραμέτρων	**8.034.958.467**	**8.034.959.729**	**2.650.439.907**	**3.0304.09.630**
MSE έπειτα από βελτιστοποίηση και τυποποίηση	**8.035.000.920**	**8.034.959.729**	**2.642.107.097**	**3.030409..630**

Πίνακας 2 : Τιμές Μέσου Τετραγωνικού Σφάλματος

Με βάση το μέσο τετραγωνικό σφάλμα (MSE), ο αλγόριθμος των τυχαίων δασών έχει την μικρότερη τιμή σε σχέση με το υπόλοιπα μοντέλα

ΡΙΖΑ ΜΈΣΟΥ ΤΕΤΡΑΓΩΝΙΚΟΎ ΣΦΑΛΜΑΤΟΣ				
	Παλινδρόμηση Ridge	Παλινδρόμηση Lasso	Τυχαία Δάση	Δένδρα Απόφασης
RMSE με αρχικές τιμές παραμέτρων	**89.638**	**89.638**	**51.482**	**55.049**
RMSE έπειτα από βελτιστοποίηση και τυποποίηση	**89.637**	**89.637**	**51.401**	**55.049**

Πίνακας 3: Τιμές Ρίζας Μέσου Τετραγωνικού Σφάλματος

Η Ρίζα του μέσου τετραγωνικού σφάλματος εμφανίζει καλύτερα την απόκλιση μεταξύ πραγματικών τιμών και τιμών πρόβλεψης, με το μοντέλο του αλγορίθμου των τυχαίων δασών να εμφανίζει την μικρότερη απόκλιση που ισούται με 51.482 δολάρια στις αρχικές τιμές παραμέτρων, ενώ η τυποποίηση και η υπερ-παραμετροποίηση οδήγησε σε μια μικρή βελτίωση κατά 79 δολάρια.

Δεύτερος στην κατάταξη έρχεται ο αλγόριθμος τυχαίων δασών με απόκλιση που ισούται με 55.049 δολάρια. Αντίθετα, τα υπόλοιπα δύο μοντέλα παλινδρόμησης δε φαίνονται ανταγωνιστικά καθώς παρουσιάζουν αρκετά χειρότερη απόδοση που αγγίζει τα 90.000 δολάρια και στις 2 περιπτώσεις.

ΜΈΣΟ ΑΠΟΛΥΤΟ ΣΦΑΛΜΑ				
	Παλινδρόμηση Ridge	Παλινδρόμηση Lasso	Τυχαία Δάση	Δένδρα Απόφασης
MAE με αρχικές τιμές παραμέτρων	**68.075**	**68.076**	**34.535**	**36.362**
MAE έπειτα από βελτιστοποίηση και τυποποίηση	**68.075**	**68.074**	**35.144**	**36.362**

Πίνακας 4: Τιμές Μέσου Απόλυτου Σφάλματος

Το μέσο απόλυτο σφάλμα δείχνει πως, τα τυχαία δάση ηγούνται σε απόδοση. Η απόκλιση με τις προκαθορισμένες παραμέτρους ανέρχεται στα 34.535 δολάρια. Η απόδοση της μετρικής όμως μειώθηκε με στο δεύτερο στάδιο του μοντέλου , με αποτέλεσμα να αυξηθεί το μέσο απόλυτο σφάλμα κατά 600 δολάρια περίπου.

Στην δεύτερη θέση έρχονται τα δένδρα απόφασης με διαφορά λίγο μικρότερη από 1.900 δολάρια η οποία δεν αλλάζει καθώς δεν σημειώθηκε αλλαγή στην τιμή κατά την βελτιστοποίηση.

Τέλος, οι αλγόριθμοι Lasso και Ridge, σημειώνουν σχεδόν διπλάσιο μέσο απόλυτο σφάλμα ,με τιμή που προσεγγίζει τα 68.000 δολάρια και στα δύο στάδια των μοντέλων.

R-SQUARED				
	Παλινδρόμηση Ridge	Παλινδρόμηση Lasso	Τυχαία Δάση	Δένδρα Απόφασης
R^2 με αρχικές τιμές παραμέτρων	**74,1%**	**74,1%**	**91,4%**	**90,26%**
R^2 έπειτα από βελτιστοποίηση και τυποποίηση	**74,1%**	**74,1%**	**91,5%**	**90,26%**

Πίνακας 5: Ποσοστά μετρικής R-squared

Η μετρική που αντιπροσωπεύει τον βαθμό προσαρμοστικότητας του μοντέλου στα δεδομένα αναδεικνύει τα τυχαία δάση στην πρώτη θέση με το υψηλό ποσοστό της τάξης του 91,4%, το οποίο βελτιώνεται αμυδρά κατά +0,1% στο στάδιο της βελτιστοποίησης.

Ως δεύτερο μοντέλο, έρχονται και πάλι τα δένδρα απόφασης με εξίσου καλό ποσοστό που ισούται με 90,26%. Η βελτιστοποίηση δεν απέφερε καμία βελτίωση στην προσαρμοστικότητα του μοντέλου.

Αναξιόπιστα καθίστανται ξανά τα μοντέλα παλινδρόμησης με ποσοστό 74%. Τιμή αρκετά αποκλίνουσα από τους δεντρικούς αλγόριθμους.

ΤΥΧΑΙΑ ΔΕΝΤΡΑ				
	MSE	RMSE	MAE	R^2
Απόδοση με αρχικές τιμές παραμέτρων	**2.650.439.907**	**51.482**	**34.535**	**91,4%**
Απόδοση έπειτα από βελτιστοποίηση και τυποποίηση	**2.642.107.097**	**51.401**	**35.144**	**91,5%**

Πίνακας 6 : Πίνακας απόδοσης Τυχαίων Δέντρων

7. Συμπεράσματα

Η επιστήμη των δεδομένων είναι ένας ραγδαία εξελισσόμενος τομές με τεράστια συμβολή σε πολλούς τομείς της καθημερινότητας. Ένας από αυτούς τους τομείς είναι η αγορά ακινήτων, η οποία αξιοποιεί τις τεχνικές και τα εργαλεία που παρέχει η επιστήμη δεδομένων για να αναλύσει και να προβλέψει την τάση της αγοράς.

Στην παρούσα πτυχιακή εργασία, πραγματοποιήθηκε ανάλυση δεδομένων της βάσης που ανήκει στο Υπουργείο Ανάπτυξης και Στέγασης της Σιγκαπούρης. Μέσα από την οπτικοποίηση των δεδομένων εντοπίστηκαν διάφορες τάσεις στην αγορά που σχετίζονται με την αξία, την ζήτηση, την διαρρύθμιση των διαμερισμάτων αλλά και τον εντοπισμό των ευρύτερων περιοχών που βρίσκονται σε ανάπτυξη. Παράλληλα , κατασκευάστηκαν τέσσερα μοντέλα, βασισμένα σε διαφορετικούς αλγορίθμους επιβλεπόμενης μηχανικής μάθησης. Οι αλγόριθμοι που επιλέχθηκαν είναι οι : Παλινδρόμηση Lasso, Παλινδρόμηση Ridge, Τυχαία Δάση (Random Forest) και Δέντρα Απόφασης (Decision Trees). Οι μετρικές αξιολόγησης : R^2, μέσο απόλυτο σφάλμα(MAE), μέσο τετραγωνικό σφάλμα (MSE) και η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) βοήθησαν στην ανάδειξη του καλύτερου μοντέλου. Συνεπώς, σύμφωνα με τα παραπάνω, καταλήγουμε στα εξής τελικά συμπεράσματα :

- Η πλειοψηφία των τιμών πώλησης κυμαίνεται σε προσιτά επίπεδα κοντά στα 500.000 δολάρια, με την συχνότητα πώλησης να μειώνεται σε μεγάλο βαθμό καθώς οι τιμές προσεγγίζουν τα 800.000 δολάρια. Μάλιστα, ελάχιστες πωλήσεις σημειώνονται σε τιμές που ξεπερνάνε το ένα εκατομμύριο δολάρια.

- Η τιμή πώλησης των ακινήτων ακολουθεί ανοδική πορεία με το πέρασμα των χρόνων σημειώνοντας άνοδο +40% κατά μεσαία τιμή, μέσα σε επτά χρόνια (2017-2024). Αυτό ενδεχομένως να οφείλεται σε πολλούς παράγοντες , όπως η πανδημία του covid-19 η οποία δημιούργησε τεράστιες ανισορροπίες στην αγορά, ο πληθωρισμός που μείωσε την αγοραστική δύναμη του χρήματος με συνέπεια την καταβολή μεγαλύτερων χρηματικών ποσών για την αγορά του ίδιου ακινήτου. Επίσης η συνεχής αύξηση του βιοτικού επιπέδου της Σιγκαπούρης, προσελκύει μεγάλο πλήθος αγοραστών με την ζήτηση να αποκτά αυξητικούς ρυθμούς.

- Τα δημοφιλέστερα διαμερίσματα είναι αυτά των τεσσάρων δωματίων, τα οποία καλύπτουν σχεδόν το 40% του συνολικού πλήθους των δειγμάτων της βάσης. Δεύτερα ιεραρχικά ακολουθούν τα διαμερίσματα πέντε και τριών δωματίων με ισόποσο αριθμό δειγμάτων ο οποίος αθροιστικά καλύπτει το 50% της βάσης δεδομένων, ενώ εξαιρετικά πολύ μικρή συχνότητα εμφανίζουν διαμερίσματα κάτω των δύο δωματίων. Η προσκόμιση της παραπάνω ανάλυσης συμπεραίνει ότι υπάρχει μεγάλη ζήτηση για διαμερίσματα ευρύχωρα που μπορούν να φιλοξενήσουν οικογένειες, καθώς η γενικότερη κατανομή των διαμερισμάτων απεικονίζει την μεγάλη ανάγκη της αγοράς για μεγάλους χώρους με πολλά δωμάτια.

- Όσον αφορά το μοντέλο διαμερίσματος , η τεράστια ύπαρξη διαμερισμάτων τύπου “MODEL_A” και “IMPROVED”, τα οποία καλύπτουν αθροιστικά πάνω από το 60% των δειγμάτων της βάσης, υποδηλώνει πως τα περισσότερα διαμερίσματα είναι χτισμένα κατά τις δεκαετίες του 1970 με 1990, ενώ ταυτόχρονα επαληθεύουν το προηγούμενο συμπέρασμα διότι , πρόκειται για διαμερίσματα που διαθέτουν τρία έως και πέντε δωμάτια.

- Με βάση την γεωγραφική κατανομή των κατοικιών, τα διαμερίσματα που βρίσκονται στην κεντρική περιοχή της Σιγκαπούρης σε συνδυασμό με τα διαμερίσματα που ανήκουν στην νότια περιοχή, κατέχουν τις υψηλότερες τιμές μεταπώλησης. Αυτό συμβαίνει διότι το κέντρο της πόλης βρίσκεται στον νότο με αποτέλεσμα η περιβάλλουσα περιοχή να αναπτυχθεί ιδιαίτερα νωρίς. Η βορειοδυτική περιοχή διαθέτει τις οικονομικότερες τιμές ,αυτή η τάση ανακλάται από το χαμηλό βοιωτικό επίπεδο της συγκεκριμένης περιοχής καθώς η ύπαρξη τροπικών δασών δεν επέτρεψαν την ανάπτυξή της.

- Με βάση την απόδοση των μοντέλων εξάγεται το συμπέρασμα πως οι δενδροειδείς αλγόριθμοι αποδίδουν καλύτερα σε σχέση με τους γραμμικούς. Αναλυτικότερα, η μετρική R^2 των γραμμικών αλγορίθμων ισούται με 74,1% ,ενώ οι δενδροειδείς παρουσιάζουν πολύ μεγαλύτερη αποτελεσματικότητα με 90,26% και 91,5%. Πρώτο σε κατάταξη μοντέλο με το ποσοστό μετρικής R^2 να ισούται με 91,5% είναι το μοντέλο των Τυχαίων δασών. Δεύτερο, ακολουθεί το μοντέλο των Δέντρων Αποφάσεων με διαφορά -1,24%. Στην συνέχεια παρατηρείται μεγάλη διαφορά στην απόδοση των μοντέλων Παλινδρόμησης Ridge και Lasso με ποσοστό R^2 ίσο με 74,1% αντίστοιχα.

- Η απόδοση των μοντέλων δεν βελτιώθηκε με την τυποποίηση των ανεξαρτήτων μεταβλητών και την υπερ-παραμετροποίηση. Εξαίρεση αποτελεί το μοντέλο των Τυχαίων Δασών με μια πολύ μικρή βελτίωση κατά +0,1% στην μετρική R^2.

- Η μη βελτίωση της απόδοσης κατά το στάδιο της βελτιστοποίησης, ενδεχομένως να οφείλεται στην έλλειψη διαχωρισμού των δεδομένων σε πολλά μέρη με την εντολή cross_val_score που διαθέτει η βιβλιοθήκη Scikit_Learn. Ο διαχωρισμός των δεδομένων βοηθά στην καταπολέμηση της υπερπροσαρμογής και στην βελτίωση της απόδοσης των μετρικών.

8.Βιβλιογραφία

[1] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, “An Unsupervised Machine Learning Algorithms: Comprehensive Review,” *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023, doi: 10.12785/ijcds/130172.

[2] Λιανός Δημ. Γεώργιος, “Τεχνητή Νοημοσύνη και Εφοδιαστική αλυσίδα,” 2023.

[3] Τ. Ψηφιακων Συστηματων and Ε. Καραγιάννης, “ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ Μεταπτυχιακή εργασία Μελέτη τεχνικών εξόρυξης δεδομένων και μηχανικής μάθησης για χρήση σε συστήματα ανίχνευσης εισβολών,” 2017.

[4] C. Sedeslis, “Μηχανική Μάθηση στην Υγεία,” Thessaloniki, 2022. [Online]. Available: www.kaggle.com

[5] M. Z. Rodriguez *et al.*, “Clustering algorithms: A comparative approach,” *PLoS One*, vol. 14, no. 1, Jan. 2019, doi: 10.1371/journal.pone.0210236.

[6] K. A. Abdul Nazeer and M. P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. 2009.

[7] Markou Giorgos, “Τεχνικές ανάλυσης δεδομένων με classification και clustering,” 2021.

[8] Κ. Αγγελική and Σ. Αθηνά, “ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ.”

[9] Δ. Σωτηρόπουλος, Γ. Τσιχριντζής, Ε. Σ. Επίκουρος, Κ. Καθηγητής, and Α. Καθηγητής, “Εξομοίωση και Σύγκριση Αλγορίθμων Ενισχυτικής Μάθησης,” 2022.

[10] Κούτσης Χαρίλαος, “ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΕ ΙΑΤΡΙΚΑ ΔΕΔΟΜΕΝΑ,” ΑΘΗΝΑ, 2022.

[11] E. Liaras, “School of Social Sciences Master in Business Administration (MBA) Postgraduate Dissertation Machine Learning in Accounting and Finance Research: A Literature Review.”

[12] P. T. R, “A Comparative Study on Decision Tree and Random Forest Using R Tool,” *IJARCCE* , pp. 196–199, Jan. 2015, doi: 10.17148/ijarcce.2015.4142.

[13] A. Carlsson, “Predictive Regression Model Evaluation: Evaluating Predictive Machine Learning Models to Reduce Food Waste in the Dairy Industry; Predictive Regression Model Evaluation: Evaluating Predictive Machine Learning Models to Reduce Food Waste in the Dairy Industry; Utvärdering av prediktiva regressionsmodeller – Utvärdering av prediktiva maskininlärningsmodeller för att minska matsvinn i mejeriindustrin.”

[14] D. T. Soon. Heng and S. Muhd. K. Aljunied, *Singapore in global history*. 2011.

[15] L. Lin-Heng, “Public Housing In Singapore: A Success Story In Sustainable Development,” 2020. [Online]. Available: http://law.nus.edu.sg/apcel/wps.htmlElectroniccopyavailableat:https://ssrn.com/abstract=3595956

[16] https://beta.data.gov.sg/datasets/d_8b84c4ee58e3cfc0ece0d773c8ca6abc/view

[17] https://www.udemy.com/course/the-complete-supervised-machine-learning-models-in-python

[18] https://www.udemy.com/course/complete-machine-learning-and-data-science-zero-to-mastery
[19] https://youtu.be/W-0-u6XVbE4?si=a0ASggQzqfLkJ4s6
[20] https://youtu.be/Q81RR3yKn30?si=MZTw7QQLwcVH0hrm
[21] https://www.kaggle.com/code/misterkix/prediction-of-singapore-hdb-price-machine-learning#Machine-Learning-Models-(Regression)
[22] https://www.teoalida.com