Πρόβλεψη Τιμών και Ανάλυση Συναισθήματος για καταλύματα Airbnb της Αθήνας

Κετσέας Σταύρος [1] Λαμπρόπουλος Ιωάννης[2] Σίλου Ηλιάνα[3]

[1] School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytechneiou 9, Athens, 15780, Greece.
[2] School of Chemical Engineering, National Technical University of Athens, Iroon Polytechneiou 9, Athens, 15780, Greece.
[3] School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Iroon Polytechneiou 9, Athens, 15780, Greece.

1. Abstract

Price prediction has proven to be a difficult problem for hosts using the platform of Airbnb for accommodation renting since the usual criteria used are empirical and subjective. In this study, we compare different algorithms for their ability to effectively predict the suitable price for any given accommodation based on characteristics such as its location and facilities. Furthermore, through the use of text mining, sentiment analysis is performed. Multiple algorithms are once again compared for their ability to utilize the extracted information. Using cross-validation schemes and performing hyperparameter tuning, it was concluded that the Random Forest algorithm yielded the best results in terms of price prediction while Logistic Regression proved to be the optimal choice for sentiment analysis.

2. Εισαγωγή

Η μίσθωση ακινήτων μέσω της πλατφόρμας Airbnb είναι μια πρακτική η οποία συνεχώς κερδίζει έδαφος στη αγορά της βραχυχρόνιας ενοικίασης. Η Airbnb αποτελεί τον δίαυλο επικοινωνίας μεταξύ ενοικιαστή και ενδιαφερομένου για την μίσθωση καταλυμάτων ανά τον κόσμο.

Ένα συχνό πρόβλημα που αντιμετωπίζουν οι ιδιοκτήτες των ακινήτων είναι ο καθορισμός της τιμής ενοικίασης για το ακίνητο τους. καθώς, έχουν στην διάθεσή τους, ως επί το πλείστον, εμπειρικά κριτήρια για τον καθορισμό της τιμής που εν τέλει ορίζουν. Συνεπώς, υπάρχει ο κίνδυνος λανθασμένης εκτίμησης, με αποτέλεσμα, είτε να αποθαρρύνουν εν δυνάμει ενοίκους, είτε να αποτυγχάνουν να μεγιστοποιούν το κέρδος της επιχειρηματικής τους δραστηριότητας.

Πέραν της επίλυσης του προαναφερθέντος προβλήματος, η εν λόγω μελέτη καταπιάνεται και με το πρόβλημα της ανάλυσης συναισθήματος, δηλαδή την εξόρυξη ανθρώπινου συναισθήματος από μια γραπτή κριτική για ένα συγκεκριμένο κατάλυμα. Αυτό επιτυγχάνεται με την κατάλληλη αξιοποίηση των δεδομένων και εφαρμόζεται σε συστήματα ανατροφοδότησης ως προς την δημοτικότητα (popularity) του εκάστοτε καταλύματος. Η πληροφορία αυτή μπορεί να αξιοποιηθεί, τόσο από την ίδια την Airbnb για σκοπούς, όπως η αναγνώριση μοτίβων που υποδεικνύουν την καταλληλόλητα ή μη ενός καταλύματος, όσο και από τους ιδιοκτήτες, ώστε να αυτοαξιολογηθούν. Ακόμα, η ανάλυση συναισθήματος συνεπικουρεί στην διαμόρφωση μιας πιο εύπεπτης πληροφορίας για τον χρήστη της πλατφόρμας, κάτι που με την σειρά του επιτρέπει την ευκολότερη εκτίμηση της ποιότητας του εκάστοτε καταλύματος.

Στην παρούσα μελέτη, αντλήθηκαν πληροφορίες από την ιστοσελίδα “Inside Airbnb” (1) με τα δεδομένα των καταχωρημένων καταλυμάτων για την ευρύτερη περιοχή των Αθηνών και αξιοποιήθηκαν κατάλληλα για την βέλτιστη διεκπεραίωση των ανωτέρω δύο βασικών αξόνων της μελέτης.

Όσον αφορά την πρόβλεψη τιμών, επιλέχθηκαν ως δεδομένα εισόδου χαρακτηριστικά του συνόλου δεδομένων που σχετίζονται άμεσα με την διαμόρφωση τιμής ενός καταλύματος και πραγματοποιήθηκε ανάλυση παλινδρόμησης με τη χρήση και σύγκριση των αλγορίθμων KNN,SVR, Decision Tree, Random Forest και Regression, για την επιλογή του βέλτιστου μοντέλου. Η αξιολόγηση του καλύτερου αλγορίθμου βασίστηκε σε δύο μετρικές, το μέσο τετραγωνικό σφάλμα και το R2 .

Στο κομμάτι της ανάλυσης συναισθήματος, τα δεδομένα εισόδου κατηγοριοποιήθηκαν με τεχνικές ταξινόμησης σε τρεις κλάσεις, οι οποίες εκφράζουν το επίπεδο ικανοποίησης του χρήστη από τις παρεχόμενες υπηρεσίες με βάση τις κριτικές των χρηστών σε μορφή σχολίων, αλλά και την βαθμολογία τους. Εφαρμόστηκαν, ακόμη, τεχνικές εξόρυξης κειμένου για την εξαγωγή υψηλής ποιότητας πληροφορίας από το περιεχόμενο της εκάστοτε κριτικής. Χρησιμοποιήθηκαν οι ταξινομητές Decision Tree, Naive-Bayes, Logistic Regression και Random Forest. Τα αποτελέσματα αξιολογήθηκαν μέσω των μετρικών ορθότητα (accuracy) και F1-score.

3. Βιβλιογραφική Ανασκόπηση

Το φαινόμενο της βραχυπρόθεσμης μίσθωσης ακινήτων χαίρει ταχύτατης επέκτασης στον ευρωπαϊκό χώρο. Περιοχές με τουριστικό ενδιαφέρον βρίσκονται στο επίκεντρο αυτής της νέας μορφής διάθεσης καταλυμάτων. Χαρακτηριστικό παράδειγμα αποτελεί η Αθήνα όπου, σύμφωνα με τα ευρήματα των Gourzis et al. (2), αυτή η νέα μορφή αστικού τουρισμού κερδίζει συνεχώς έδαφος, μεταμορφώνοντας ακόμα και περιοχές οι οποίες ήταν, μέχρι πρότινος, ανεκμετάλλευτες. Σε αυτό το πνεύμα, οι Alexandridis et al (3) εφάρμοσαν μια μεθοδολογία για να χαρτογραφήσουν τα προσφερόμενα καταλύματα στον Αθηναϊκό χώρο και να εξορύξουν μοτίβα και άλλες χρήσιμες πληροφορίες από πολλαπλές αναλύσεις που αφορούν στην θέση του εκάστοτε ακινήτου.

Όπως είναι συνεπώς αναμενόμενο, το πρόβλημα της εκτίμησης κατάλληλης τιμής μίσθωσης έχει απασχολήσει πολλούς μελετητές, καθώς η εύρεση ενός αξιόπιστου αλγορίθμου βρίσκει άμεση εφαρμογή στον χώρο του τουρισμού. Μια πολύ ενδιαφέρουσα προσέγγιση πραγματοποίησαν οι Kalehbasti et al. (4). Χρησιμοποίησαν μεθόδους μηχανικής μάθησης και βαθιάς μάθησης, όπως αλγορίθμους SVR, k-Means και νευρωνικά δίκτυα με σκοπό την πρόβλεψη ενδεδειγμένης τιμής μίσθωσης. Επιπλέον, προχώρησαν στην υλοποίηση ανάλυσης

συναισθήματος μέσω της κειμενικής εξόρυξης (text mining).
Τεχνικές μηχανικής και βαθιάς μάθησης συνδύασαν όμως και οι Zhu et al. (5). Συγκεκριμένα, δοκιμάστηκαν τεχνικές μηχανικής και βαθιάς μάθησης, αλλά και τεχνικές boosting και bagging (συνδυασμός μεθόδων) για να προβλέψουν την κατάλληλη τιμή μίσθωσης για την πόλη της Νέας Υόρκης. Αποδείχθηκε ότι οι τεχνικές bagging, XGBoost και Random Forest προκρίθηκαν ως οι βέλτιστες επιλογές για την επίλυση του προβλήματος.
Σε συνδυασμό μεθόδων συσταδοποίησης κατέφυγαν και οι Tang and Sangani (6) οι οποίοι, όχι μόνο επέλεξαν να δώσουν πρωτεύουσα σημασία σε χαρακτηριστικά που αφορούν την τοποθεσία του καταλύματος (νυχτερινή ζωή, κουλτούρα κ.α.) αλλά, αποφάσισαν να χρησιμοποιήσουν και τις εικόνες των καταλυμάτων στις καταχωρήσεις της πλατφόρμας με σκοπό να ανιχνεύσουν κοινά σημεία. Για το σκοπό αυτό κατήρτισαν ένα λεξικό ώστε να κατηγοριοποιήσει κάθε εικόνα σε μια από τις 1000 συστάδες του οι οποίες δημιουργήθηκαν με την βοήθεια του αλγορίθμου k-Means. Τέλος, αφού διαμόρφωσαν τον πίνακα των χαρακτηριστικών, εμπλουτίζοντάς τον με το επιπλέον χαρακτηριστικό που σχετίζεται με την εικόνα κατά την καταχώρηση, προχώρησαν στον διαχωρισμό των καταλυμάτων σε ζώνες κοστολόγησης μέσω του αλγορίθμου SVM.
Οι Li et al. (7) προχώρησαν στην υλοποίηση μεθόδου Multi-Scale Affinity Propagation (MSAP) για να κατηγοριοποιήσουν τα καταλύματα με βάση την τοποθεσία και τις προσφερόμενες παροχές, προτού χρησιμοποιήσουν γραμμική παλινδρόμηση LRNN (Linear Regression with Normal Noise). Η παραπάνω μέθοδος επιτυγχάνει τον διαχωρισμό των καταλυμάτων σε ζώνες προτεινόμενων τιμών μίσθωσης, λαμβάνοντας υπόψη τα οικονομικά δεδομένα κάθε πόλης.
Οι Priambodo and Sihabuddin (8) υιοθέτησαν μια διαφορετική τακτική αφού, χρησιμοποίησαν για τον ίδιο σκοπό την προσέγγιση των ELM (Extreme Learning Machines) με δεδομένα από την αγορά του Λονδίνου, επιτυγχάνοντας αυξημένη ακρίβεια και μεγαλύτερη ταχύτητα στο στάδιο εκπαίδευσης του μοντέλου συγκριτικά με το υφιστάμενο μοντέλο.
Τέλος, οι Cai and Han (9) συνέκριναν διάφορες μεθόδους παλινδρόμησης με σκοπό την πρόβλεψη της κατάλληλης τιμής μίσθωσης, αντλώντας πληροφορίες από τις αγγελίες στην πλατφόρμα Airbnb για την Μελβούρνη. Το μοντέλο τους εμφάνισε ικανοποιητικά αποτελέσματα για την πρόβλεψη τιμών σε οικονομικότερα καταλύματα ενώ, έτεινε να υποεκτιμά την τιμή μίσθωσης πολυτελέστερων καταλυμάτων.

4. Σύνολο δεδομένων και Χαρακτηριστικά

Στα πλαίσια της παρούσας ερευνητικής εργασίας μελετήθηκαν δύο διαφορετικά σύνολα δεδομένων. Το πρώτο αφορά το κομμάτι της πρόβλεψης της ενδεδειγμένης τιμής μιας νέας εγγραφής καταλύματος στην ιστοσελίδα της Airbnb ενώ, το δεύτερο συμβάλει στην περάτωση της ανάλυσης συναισθήματος των καταναλωτών της πλατφόρμας.
Στην παρούσα παράγραφο θα παρουσιαστεί η διαδικασία προετοιμασίας των δύο συνόλων δεδομένων ξεχωριστά.

4.1. Σύνολο Δεδομένων για τον καθορισμό ενδεδειγμένης τιμής μίσθωσης

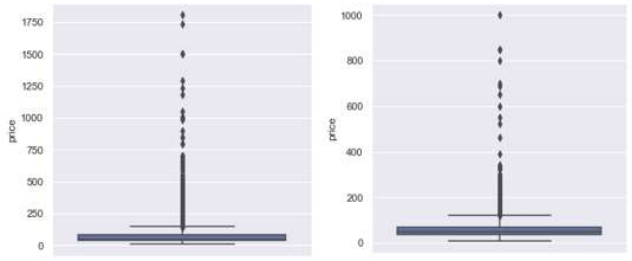
Το μελετώμενο σύνολο δεδομένων αποτελείται από 9674 εγγραφές καταλυμάτων στις οποίες καταγράφονται 75 χαρακτηριστικά. Τα χαρακτηριστικά αυτά καλύπτουν ένα μεγάλο εύρος πληροφορίας που αφορούν το ίδιο το κατάλυμα αλλά και τον ενοικιαστή.

4.1.1. Προ επεξεργασία Δεδομένων

4.1.1.1. Καθαρισμός Δεδομένων

Για την ανάγκη πρόβλεψης τιμής νέων εγγραφών επιβίωσαν δεκαέξι χαρακτηριστικά που κρίθηκαν άμεσα σχετιζόμενα με την διαμόρφωση της τιμής μίσθωσης ενός ακινήτου, μειώνοντας έτσι τη διαστατικότητα του πίνακα. Αναλυτικότερα, τα χαρακτηριστικά αυτά είναι η γειτονιά, η ευρύτερη περιοχή, το γεωγραφικό πλάτος και μήκος, ο βραβευμένος ή μη οικοδεσπότης, το είδος ενός δωματίου, ο μέγιστος αριθμός καλεσμένων, ο αριθμός των μπάνιων, κλινών και δωματίων, η τιμή, η διαθεσιμότητα ανά χρόνο, η ελάχιστη και μέγιστη διαμονή, βαθμολογία και ο αριθμός αξιολογήσεων ανά μήνα.
Στα χαρακτηριστικά που αφορούν την ευρύτερη περιοχή, το βραβευμένο ή μη οικοδεσπότη, τον αριθμό των μπάνιων, δωματίων και κλινών, τη βαθμολογία και των αριθμό αξιολογήσεων ανά μήνα, εντοπίστηκαν απουσιάζουσες τιμές. Για την εξάλειψή τους εφαρμόστηκε διαφορετική τεχνική για το εκάστοτε χαρακτηριστικό. Συγκεκριμένα, για τον αριθμό κλινών και δωματίων και τη βαθμολογία συμπληρώθηκε στις κενές εγγραφές η μέση τιμή του αντίστοιχου χαρακτηριστικού. Για το βραβευμένο ή μη οικοδεσπότη αντικαταστάθηκαν οι απουσιάζουσες τιμές με τη συχνότερα εμφανιζόμενη. Στη συνέχεια επιλέχθηκε ως καταλληλότερη τεχνική η συμπλήρωση μηδενικών τιμών στην περίπτωση των αριθμών αξιολογήσεων ανά μήνα.
Οι αναγραφόμενες γειτονιές ομαδοποιήθηκαν με σκοπό την συλλογικότερη παρουσίαση των αποτελεσμάτων, καταρτίζοντας τη στήλη με το χαρακτηριστικό της ευρύτερης περιοχής, το οποίο αποτελούνταν αποκλειστικά από απουσιάζουσες τιμές.
Αφότου ολοκληρώθηκε το κομμάτι κατάλληλης διαχείρισης των απουσιαζουσών τιμών, κρίθηκε αναγκαίος ο εντοπισμός ακραίων τιμών. Για παράδειγμα, αρχικά εντοπίστηκαν ακραίες τιμές στο χαρακτηριστικό ελάχιστες νύχτες διαμονής.
Παρατηρείται ότι υπάρχει μικρός αριθμός εγγραφών με ελάχιστη διαμονή 1000 νύχτες, κατάσταση που δεν συνάδει με μια πλατφόρμα βραχυχρόνιας μίσθωσης, όπως η Airbnb. Στα πλαίσια της ανάλυσής μας θα περιορίσουμε την τιμή των ελάχιστων διανυκτερεύσεων σε 3 κατά μέγιστο, καθώς αυτός ο αριθμός εκφράζει το σύνολο των δεδομένων μας και πιο συγκεκριμένα περιγράφει περίπου το 99,8% αυτών.
Στη συνέχεια πραγματοποιήθηκε έλεγχος ακραίων τιμών στο χαρακτηριστικό της τιμής. Στο Σχήμα (1) παρουσιάζονται αυτές οι ακραίες τιμές, οι οποίες τελικά αφαιρέθηκαν από το σύνολο δεδομένων. Αντίστοιχα, η μέγιστη τιμή ανά διανυκτέρευση περιορίστηκε στα 2500 ευρώ πάνω από την μέση τιμή, δηλαδή στα 2572,53 ευρώ. Παρόλο που τα καταλύματα με τις υψηλότερες τιμές μίσθωσης δεν απευθύνονται στον μέσο ενδιαφερόμενο, η πλατφόρμα παρέχει λύσεις μίσθωσης κατοικιών υψηλού κόστους που δικαιολογούν αυτή τη μέγιστη τιμή. Ως εκ τούτου, θεωρήθηκε λάθος η αφαίρεσή τους από το σύνολο δεδομένων. Οι χειρισμοί που αναλύθηκαν παραπάνω συνέβαλαν στην ομαλοποίηση του συνόλου δεδομένων. Ο Άγιος Ελευθέριος παρουσιαζόταν ως η ακριβότερη περιοχή της Αθήνας πριν την επεξεργασία των δεδομένων, πράγμα που φαντάζει αδύνατο. Η συμπεριφορά αυτή προκλήθηκε καθώς, στην περιοχή του Αγίου Ελευθερίου υπάρχουν εγγεγραμμένα καταλύματα με τιμή διανυκτέρευσης

έως και 9000 ευρώ, τα οποία δεν αποτελούν ενδεδειγμένα παραδείγματα για την περιοχή, Συνεπώς, θεωρήθηκαν ακραίες τιμές και αφαιρέθηκαν. Κατόπιν της αφαίρεσης ακραίων τιμών, παρατηρήθηκε αισθητή μείωση της ασυμμετρίας (skewness) των δεδομένων, γεγονός το οποίο συνεπάγεται ότι η ουρά στα δεδομένα, η οποία εκφράζει την ύπαρξη ακραίων τιμών, περιορίστηκε και πλέον τα δεδομένα κατανέμονται πιο ομοιόμορφα.



(α) (β)

Σχήμα 1: Θηκόγραμμα για το χαρακτηριστικό της τιμής (α) πριν και (β) μετά την αφαίρεση των ακραίων τιμών.

4.1.1.2. Μετασχηματισμός Χαρακτηριστικών

Για την πιο εύστοχη πρόβλεψη της τιμής μιας νέας εγγραφής καταλύματος στην ιστοσελίδα του Airbnb, χρειάστηκαν κατάλληλες τεχνικές μετασχηματισμού των χαρακτηριστικών που αφορούν την προετοιμασία των δεδομένων ώστε να είναι πλήρως συμβατά με τους επιλεγμένους αλγόριθμους παλινδρόμησης, αλλά και με την βελτιστοποίηση των μοντέλων μηχανικής μάθησης.

Το χαρακτηριστικό βραβευμένος ή μη οικοδεσπότης μετατράπηκε από κατηγορικό σε δυαδικό, ώστε να διευκολυνθεί η ανάλυση παλινδρόμησης η οποία δέχεται αριθμητικές τιμές ως ορίσματα. Μια ακόμη μετατροπή είναι η μετατροπή των χαρακτηριστικών αριθμός μπάνιων και τιμή από αλφαριθμητικά σε αριθμητικά.

Επιπλέον, δημιουργήθηκε το χαρακτηριστικό δημοτικότητα (popularity) από το γινόμενο των χαρακτηριστικών βαθμολογία και αριθμός αξιολογήσεων ανά μήνα. Αφού μετρηθεί η εκάστοτε δημοτικότητα, επιλέγονται τα καταλύματα που ανήκουν στο 25% του συνόλου με την υψηλότερη δημοτικότητα. Έπειτα, εκπαιδεύεται το επιλεγμένο μοντέλο βασισμένο στις δημοφιλέστερες εγγραφές. Αν το μοντέλο εκπαιδευτεί με βάση τα καταλύματα τα οποία κατατάσσονται στα πιο δημοφιλή, θα δώσει ως αποτέλεσμα μια ανταγωνιστική προτεινόμενη τιμή, επιλέγοντας κοντινή εκτιμώμενη αξία. Αναζητώντας τις εγγραφές που ανήκουν στο 25% των πλέον δημοφιλών, καταλήξαμε σε 2416 εγγραφές που πληρούν τις προδιαγραφές.

Στη συνέχεια πραγματοποιήθηκε έλεγχος και σε υπόλοιπα χαρακτηριστικά όπως ο αριθμός μπάνιων (bathrooms_text), το οποίο αποτελείται από έναν υπερβολικά μεγάλο αριθμό διαφορετικών τιμών.

Στην προσπάθεια να εξηγηθεί αυτή η συμπεριφορά, ανακαλύφθηκε ότι τα δεδομένα, που στο αρχικό σύνολο αποτυπώνονταν σε μορφή κειμένου (string), ήταν μια περιγραφή του αριθμού και του τύπου του μπάνιου για κάθε εγγραφή, η οποία περιλάμβανε τόσο τα ιδιωτικά όσο και τα κοινόχρηστα μπάνια.

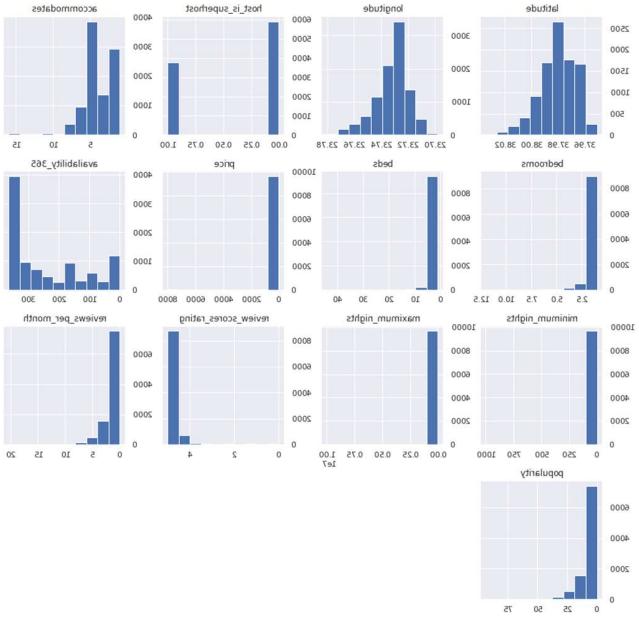
Είναι προφανές ότι στη λήψη της απόφασης θα διαδραματίσει σημαντικό ρόλο η προτίμηση ενός ιδιωτικού σε σχέση με ένα κοινόχρηστο μπάνιο. Συνεπώς, τελευταίο στάδιο του μετασχηματισμού των δεδομένων αποτελεί η δημιουργία ξεχωριστών στηλών τόσο για τον αριθμό των μπάνιων όσο και για τον τύπο τους. Οι απουσιάζουσες τιμές συμπληρώθηκαν με μηδενικές και κατασκευάστηκε το σύνολο δεδομένων με τίτλο new_pop, το οποίο χρησιμοποιήθηκε στην εξερεύνηση των δεδομένων και στην εφαρμογή τεχνικών μηχανικής μάθησης.

4.1.1.3. Εξερεύνηση των Δεδομένων

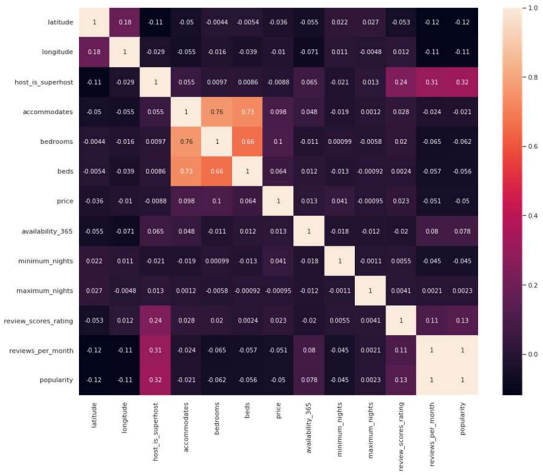
Σε αυτή την ενότητα παρουσιάζονται και οπτικοποιούνται τα δεδομένα, με στόχο την ανακάλυψη πιθανών συσχετίσεων μεταξύ των διαφόρων μεταβλητών.

Μερικές από αυτές τις οπτικοποιήσεις αφορούν την κατανομή κάθε χαρακτηριστικού, τη συσχέτισή του με την τιμή μέσω heatmaps και την πιθανή συσχέτιση μεταξύ των αριθμητικών τιμών. Επιπλέον, παρουσιάζονται ραβδογράμματα που κατατάσσουν τις φθηνότερες και ακριβότερες περιοχές της Αθήνας, το πλήθος των εγγραφών ανά γειτονιά, την κατανομή αυτών σε ευρύτερες περιοχές και των τύπων δωματίων ανά γειτονική περιοχή. Τέλος, παρουσιάζεται χάρτης που απεικονίζει την κατανομή των εγγραφών στην Αθήνα.

Στο Σχήμα (2) παρουσιάζεται η κατανομή κάθε χαρακτηριστικού ενώ, το Σχήμα (3) αποτυπώνει τον χάρτη συσχέτισης (heatmap) μεταξύ των χαρακτηριστικών, αλλά και η συσχέτισή τους με το χαρακτηριστικό τιμή.



Σχήμα 2: Κατανομή των χαρακτηριστικών του μελετώμενου συνόλου δεδομένων.

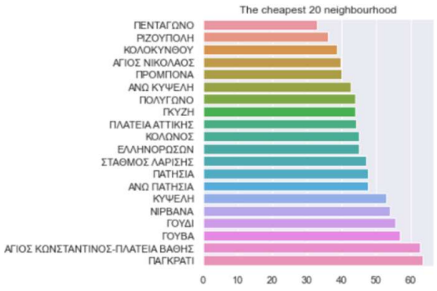


Σχήμα 3: Χάρτης συσχέτισης των χαρακτηριστικών του μελετώμενου συνόλου δεδομένων.

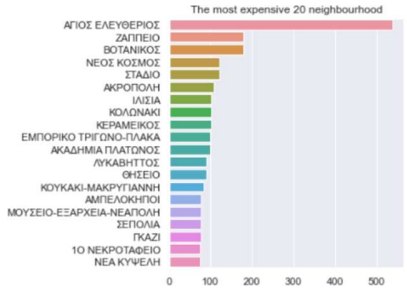
Η απόλυτη τιμή των αριθμών δείχνει την ισχύ της συσχέτισης μεταξύ των χαρακτηριστικών. Θετικοί αριθμοί εκφράζουν θετικές συσχετίσεις μεταξύ χαρακτηριστικών και αντίστοιχα οι αρνητικοί αριθμοί τις αρνητικές. Ξεχωρίζει η έντονη θετική συσχέτιση μεταξύ των χαρακτηριστικών μέγιστος αριθμός φιλοξενούμενων, μπάνια και κρεβάτια, κάτι που είναι απόλυτα λογικό, αλλά και συσχετίσεις όπως αυτή της τιμής με τον αριθμό υπνοδωμάτια και το μέγιστο αριθμό φιλοξενούμενων. Όσον αφορά το χαρακτηριστικό της δημοτικότητας, είναι επίσης πολύ λογικό να σχετίζεται με τον αν ο οικοδεσπότης είναι βραβευμένος ή όχι και με την ετήσια διαθεσιμότητα του καταλύματος. Καταλύματα με μικρή διαθεσιμότητα δεν θα μπορούσαν άλλωστε να είναι δημοφιλή. Αξιοσημείωτη είναι και η αρνητική συσχέτιση μεταξύ δημοτικότητας και ελαχίστων διανυκτερεύσεων. Αυτό καταδεικνύει ότι καταλύματα με πολλές ελάχιστες διανυκτερεύσεις δεν είναι δημοφιλή, γεγονός που αποδίδεται στον περιορισμό των υποψήφιων πελατών σε αυτούς που είναι διατεθειμένοι να προχωρήσουν σε μια πολυήμερη διαμονή.

Στο Σχήμα (4α) παρατίθενται, σε αύξουσα σειρά, οι φθηνότερες περιοχές της Αθήνας και στο Σχήμα (4β), σε φθίνουσα σειρά, οι ακριβότερες. Εδώ επαληθεύεται η προηγούμενη διαπίστωση ότι ο Άγιος Ελευθέριος έχει ακραίες τιμές καταλυμάτων που επηρεάζουν τη μέση τιμή του και τον κατατάσσουν λανθασμένα ως την ακριβότερη περιοχή.

(α)



(β)

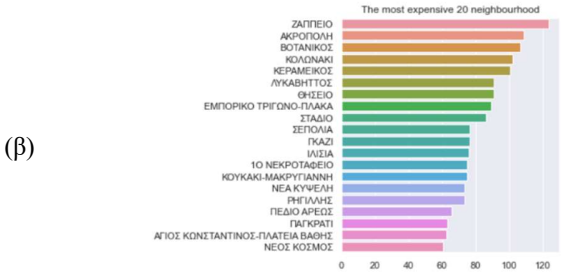


Σχήμα 4: Κατάταξη των (α) 20 περιοχών με τις φθηνότερες μισθώσεις καταλυμάτων στην πλατφόρμα Airbnb σε αύξουσα σειρά και (β) 20 περιοχών με τις ακριβότερες μισθώσεις καταλυμάτων σε φθίνουσα σειρά πριν την αφαίρεση των ακραίων τιμών.

Κατόπιν, βέβαια, της τροποποίησης που προαναφέρθηκε, συνοψίζεται στο Σχήμα (5) η πιο αντικειμενική εικόνα των φθηνότερων και ακριβότερων περιοχών.

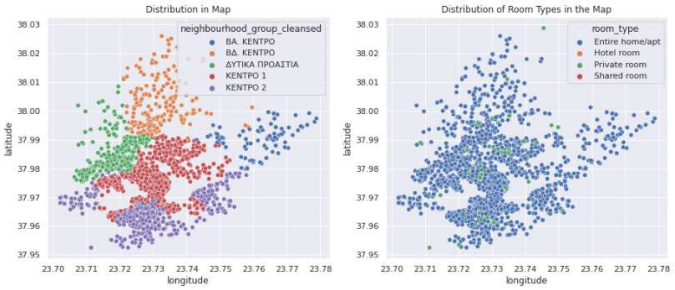
(α)

(β)



Σχήμα 5: Κατάταξη των (α) 20 περιοχών με τις φθηνότερες μισθώσεις καταλυμάτων στην πλατφόρμα Airbnb σε αύξουσα σειρά και (β) 20 περιοχών με τις ακριβότερες μισθώσεις καταλυμάτων σε φθίνουσα σειρά κατόπιν της τροποποίησης του συνόλου δεδομένων.

Στο Σχήμα (6) παρουσιάζεται η κατανομή των γειτονικών περιοχών και των τύπων δωματίων ανά γειτονική περιοχή στην Αθήνα.



Σχήμα 6: Κατανομή των καταλυμάτων ανά (α) περιοχή και (β) τύπο καταλύματος.

Στο Σχήμα (7) παρουσιάζεται ο δυναμικός χάρτης των συστάδων των εγγεγραμμένων καταλυμάτων της Αθήνας. Χαρακτηρίζεται δυναμικός καθώς, εστιάζοντας όλο και περισσότερο, οι συστάδες αναλύονται σε υποσυστάδες και ο χάρτης παράγει ανάλογα αποτελέσματα. Με μια γρήγορη ματιά παρατηρούμε ότι τα περισσότερα καταλύματα συγκεντρώνονται σε περιοχές γύρω από το κέντρο της Αθήνας, όπως Αττική, Μεταξουργείο, Κεραμεικός, Μοναστηράκι, Πλάκα , Ακρόπολη, Ζάππειο, Μακρυγιάννη και Νέος Κόσμος.

Σχήμα 7: Δυναμικός χάρτης των συστάδων των καταλυμάτων στην ευρύτερη περιοχή των Αθηνών.

4.2. Σύνολο Δεδομένων για την Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος είναι η διαδικασία ανίχνευσης θετικού ή αρνητικού συναισθήματος από δεδομένο κείμενο και συχνά χρησιμοποιείται από επιχειρήσεις για την διερεύνηση της κοινής γνώμης. Η ανάλυση συναισθήματος είναι εξαιρετικά

σημαντική γιατί βοηθά τις επιχειρήσεις να κατανοήσουν γρήγορα τις απόψεις των πελατών τους. Για να επιτευχθεί αυτό, απαραίτητη είναι η ταξινόμηση του συναισθήματος το οποίο ποσοτικοποιείται μέσω κριτικών και σχολίων.

Το δεύτερο σύνολο δεδομένων που αξιοποιήθηκε στα πλαίσια της παρούσας εργασίας, αφορά κριτικές (reviews) χρηστών στην πλατφόρμα αναφορικά με καταλύματα στην περιοχή των Αθηνών. Το εν λόγω σύνολο δεδομένων αποτελείται από 41471 κριτικές και 6 χαρακτηριστικά. Συγκεκριμένα, τα χαρακτηριστικά αυτά είναι ο κωδικός της καταχώρησης ενός καταλύματος, της κριτικής και του κριτή, η ημερομηνία της κριτικής και η ίδια η κριτική σε μορφή σχολιασμού. Είναι φανερό, πως χωρίς την βαθμολογία και μόνο με τα σχόλια των κριτών, δεν είναι εφικτή η ταξινόμηση ενός συναισθήματος. Για αυτό το λόγο, συγχωνεύτηκε το πρώτο σύνολο δεδομένων με το δεύτερο, με βάση το χαρακτηριστικό του κωδικού καταχώρησης, το οποίο είναι κοινό και στα δύο σύνολα.

4.2.1. Προ επεξεργασία Δεδομένων

4.2.1.1. Καθαρισμός Δεδομένων

Από τα δύο σύνολα δεδομένων, επιλέχθηκαν τα χαρακτηριστικά αυτά που διαδραματίζουν σημαντικό ρόλο στην ανάλυση του συναισθήματος. Αυτά είναι οι βαθμολογίες από το πρώτο σύνολο δεδομένων και οι κριτικές σε μορφή σχολιασμού από το δεύτερο.

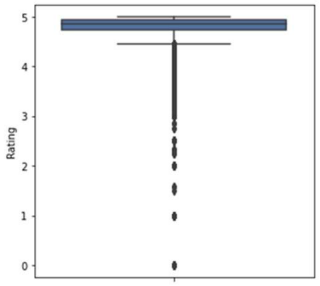
Έπειτα από την επιλογή χαρακτηριστικών, διαχειρίστηκαν οι απουσιάζουσες τιμές. Αν σε μία καταχώρηση υπάρχει μόνο η βαθμολογία ή μόνο η κριτική, δεν μπορεί να πραγματοποιηθεί ανάλυση συναισθήματος. Για αυτό το λόγο αφαιρέθηκαν οι καταχωρήσεις με απουσιάζουσες τιμές.

Το επόμενο βήμα είναι η αφαίρεση των σημείων στίξης καθώς και λέξεων που δεν προσφέρουν σημασιολογική αξία στο κείμενο (stopwords) όπως "the", "a", "to", "and", "he", "she" στα αγγλικά και "έναν", "άλλο", "από", "ας" στα ελληνικά.

4.2.1.2. Μετασχηματισμός Χαρακτηριστικών

Για την ταξινόμηση του συναισθήματος, δημιουργήθηκε ένα επιπλέον χαρακτηριστικό, το συναίσθημα (Sentiment). Το χαρακτηριστικό αυτό παίρνει τιμές 0,1,2 ανάλογα με το πόσο ικανοποιημένος έμεινε ο ένοικος μετά την διαμονή του στο εκάστοτε κατάλυμα. Η τιμή 0 ερμηνεύεται ως μη ικανοποιημένος (unsatisfied) , η τιμή 1 ως ουδέτερος/ικανοποιημένος (neutral/satisfied) και η τιμή 2 ως ενθουσιασμένος (excited). Ως "2" ταξινομήθηκαν τα καταλύματα με βαθμολογία πάνω από τον μέσο όρο ο οποίος είναι 4.81/5. Ως "0" ταξινομήθηκαν τα καταλύματα τα οποία ήταν στα χειρότερα 1%, δηλαδή είχαν βαθμολογία κάτω από 4.22/5. Τα καταλύματα με βαθμολογία ανάμεσα στις παραπάνω τιμές ταξινομήθηκαν ως "1".

Στο Σχήμα (8) παρουσιάζεται θηκόγραμμα (box plot) των βαθμολογιών.



Σχήμα 8: Θηκόγραμμα των βαθμολογιών (ratings) για τα υπό μελέτη καταλύματα του συνόλου δεδομένων.

Στον Πίνακα (1) παρουσιάζεται ενδεικτικό μέρος του συνόλου δεδομένων έχοντας εισάγει το συναίσθημα (sentiment). Επιπλέον, παρουσιάζεται συγκριτικά το κείμενο πριν και μετά την αφαίρεση των σημείων στίξης και των λέξεων που δεν προσφέρουν σημασιολογική αξία.

Πίνακας 1: Παραδείγματα κριτικών του συνόλου δεδομένων (α) πριν και (β) μετά την αφαίρεση stopwords καθώς, και την ανάθεση του αντίστοιχου συναισθήματος.

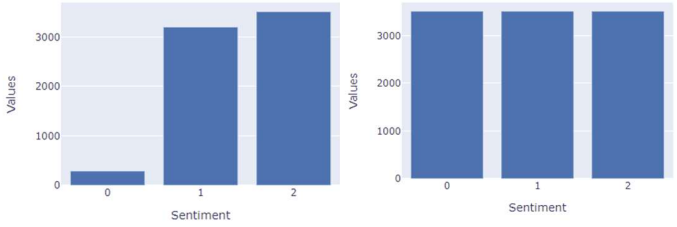
(α)

	Review	Rating
414213	When we first stumbled upon this flat on AirBn...	5.0
397736	Nous avons eu un accueil irréprochable ! L...	5.0
397830	Best Apartment I have ever had. Such a great v...	5.0
397802	Το σπίτι είναι σε πολύ βολική τοποθεσία, ακριβ...	5.0
397801	Πολύ όμορφο διαμέρισμα με όλα τα βασικά, σε κα...	5.0

(β)

	Review	Rating	Sentiment
414213	when we first stumbled upon this flat on airbn...	5.0	2
397736	nous avons eu un accueil irréprochable brlappa...	5.0	2
397830	best apartment i have ever had such a great vi...	5.0	2
397802	σπίτι είναι πολύ βολική τοποθεσία ακριβώς απέν...	5.0	2
397801	πολύ όμορφο διαμέρισμα όλα βασικά καλή πολυκατ...	5.0	2

Στο Σχήμα (9) παρουσιάζεται η συχνότητα των κλάσεων του χαρακτηριστικού sentiment. Η ταξινόμηση με μοναδικό κριτήριο τις βαθμολογίες δεν είναι η βέλτιστη λύση. Χρειάζεται και η ποσοτικοποίηση της συχνότητας εμφάνισης κάθε όρου. Για αυτο το λόγο χρησιμοποιήθηκε η μέθοδος TF-IDF (Term Frequency - Inverse Document Frequency). Το TF-IDF αποτελείται από δύο όρους. Ο πρώτος είναι το Term Frequency (TF) και είναι η συχνότητα με την οποία εμφανίζεται ο κάθε όρος στο κείμενο. Οι λέξεις με μεγάλη συχνότητα είναι σημαντικότερες για το κείμενο από ό,τι λέξεις με μικρή. Ο δεύτερος όρος είναι το Inverse Document Frequency (IDF), που είναι ένας δείκτης της πληροφορίας που δίνει η κάθε λέξη. Αν η λέξη εμφανίζεται σε όλα τα κείμενα τότε αυτή δε δίνει πολύτιμη πληροφορία. Αντίθετα, όσο πιο σπάνια εμφανίζεται η λέξη, τόσο πιο μεγάλη πληροφορία παρέχει. Στο Σχήμα (9α) φαίνεται η συχνότητα των κλάσεων μετά το TF-IDF.



(α) (β)

Σχήμα 9: Κατανομή των δειγμάτων του συνόλου δεδομένων σε κλάσεις (α) πριν την εφαρμογή της τυχαίας υπερδειγματοληψίας και (β) κατόπιν αυτής.

Τέλος, επειδή τα δεδομένα δεν ήταν ισορροπημένα, εφαρμόστηκε η τεχνική της υπερδειγματοληψίας (oversampling) με σκοπό την εξισορρόπηση τους. Η τεχνική της τυχαίας υπερδειγματοληψίας (random oversampling), η οποία και επιλέχθηκε, επιλέγει τυχαία ορισμένα παραδείγματα από τις λιγότερο συχνές κατηγορίες και τα επαναλαμβάνει. Με αυτό το τρόπο τα δείγματα όλων των κλάσεων καθίστανται ισάριθμα. Ο καταμερισμός των κλάσεων μετά την εφαρμογής υπερδειγματοληψίας παρουσιάζεται στο Σχήμα (9β).

5. Μεθοδολογία – Μοντελοποίηση

5.1. Μεθοδολογία - Μοντελοποίηση για την πρόβλεψη τιμής

Για την επιτυχή διεκπεραίωση του προβλήματος της πρόβλεψης τιμής εφαρμόστηκαν εννέα προβλεπτικά μοντέλα. Η εφαρμογή του εκάστοτε μοντέλου δεν επιλέχθηκε τυχαία, αλλά βασίστηκε εξ'ολοκλήρου στη σχετική βιβλιογραφία. Μπορεί εν τέλη, μερικές τεχνικές να αποδεικνύονται μη ικανοποιητικές, όμως εντοπίστηκε στη βιβλιογραφία ικανοποιητική εφαρμογή τους σε δεδομένα άλλων πόλεων και κρίθηκε από την ερευνητική ομάδα κατάλληλη η δοκιμή τους, ανεξαρτήτως αποτελέσματος. Πιο συγκεκριμένα:

- KNN
- SVR
- Decision Tree
- Random Forest
- Linear Regression
- Ridge Regression
- Elastic Net Regression

Σκοπός της έρευνας όπως προαναφέρθηκε είναι η όσο το δυνατό ακριβέστερη εκτίμηση της τιμής μιας νέας εγγραφής ενός καταλύματος στην ιστοσελίδα της Airbnb, σύμφωνα πάντα με το σύνολο δεδομένων που αφορά τα καταλύματα της Αθήνας. Για να επιτευχθεί αυτό, δεδομένου ότι το χαρακτηριστικό τιμή αποτελείται από συνεχείς τιμές, θα εφαρμοστεί ανάλυση παλινδρόμησης. Τέτοια παραδοσιακά μοντέλα παλινδρόμησης αποτελούν, για παράδειγμα, οι αλγόριθμοι KNN και SVM.

Για την αξιολόγηση των παραπάνω μεθοδολογιών στο πρόβλημα της πρόβλεψης τιμών χρησιμοποιήθηκαν οι μετρικές R2 και μέσο τετραγωνικό σφάλμα (mean squared error). Είναι ευνόητο ότι μετρώντας κανείς σφάλματα, επιθυμεί την ελαχιστοποίηση τους. Για κάθε μοντέλο, πραγματοποιήθηκε εκπαίδευση (train) και προσαρμογή (fit) των δεδομένων στο εκάστοτε μοντέλο. Στη συνέχεια, για να πραγματοποιηθεί η αξιολόγηση των μοντέλων στις επιλεγμένες μετρικές, αξιοποιήθηκε κατάλληλα η απόκλιση μεταξύ των πραγματικών και των προβλεφθέντων τιμών.

KNN

Ο αλγόριθμος k-Nearest Neighbors μπορεί να χρησιμοποιηθεί είτε σε προβλήματα ταξινόμησης, είτε σε παλινδρόμησης. Χρησιμοποιεί την μετρική της ομοιότητας χαρακτηριστικών για να προβλέψει τιμές οποιασδήποτε νέας εγγραφής σε ένα σύνολο δεδομένων. Αυτό πρακτικά σημαίνει ότι στη νέα αυτή εγγραφή θα ανατεθεί μια τιμή, βασισμένη στο πόσο πολύ μοιάζει με τις υπόλοιπες εγγραφές του συνόλου εκπαίδευσης.

SVR

Η επίλυση του προβλήματος παλινδρόμησης έγκειται στην εύρεση συνάρτησης που χαρτογραφεί προσεγγιστικά από ένα domain εισόδου σε πραγματικούς αριθμούς στη βάση ενός δείγματος εκπαίδευσης. Για να κατανοήσει κανείς βαθύτερα τον τρόπο λειτουργίας ενός SVR, θα μπορούσε να φανταστεί δύο ευθείες που συμβολίζουν το όριο απόφασης και περικλείουν μέσα τους ένα υπερεπίπεδο. Στόχος είναι προχωρώντας με τον αλγόριθμο SVR, να συμπεριληφθούν τα σημεία τα οποία βρίσκονται ανάμεσα στις γραμμές του ορίου απόφασης. Η γραμμή εκείνη που ταιριάζει καλύτερα στα δεδομένα μας (best fit) είναι εκείνο το υπερεπίπεδο με το μέγιστο αριθμό σημείων. Το γραμμικό SVR (Linear Support Vector Regressor) ανήκει στην οικογένεια των SVMs (Support Vector Machines) και είναι κατάλληλο για προβλήματα ανάλυσης παλινδρόμησης και για αυτό το λόγο επιλέχθηκε. Στις περισσότερες εφαρμογές του δίνει ακριβέστερα αποτελέσματα και εφαρμόζεται ταχύτατα.

Decision Tree

Τα Δέντρα Αποφάσεων (Decision Tree) είναι εργαλείο λήψης αποφάσεων το οποίο χρησιμοποιεί ένα διάγραμμα ροής στη δομή ενός δέντρου ή θα μπορούσε να χαρακτηριστεί ως ένα μοντέλο αποφάσεων το οποίο περιέχει κάθε πιθανό αποτέλεσμα, κάθε είσοδο, κάθε κόστος και κάθε χρησιμότητα. Οι αλγόριθμοι δέντρων αποφάσεων ανήκουν στην οικογένεια των αλγορίθμων επιβλεπόμενης μάθησης και δύναται να χρησιμοποιηθούν τόσο σε προβλήματα ταξινόμησης, όσο και παλινδρόμησης. Δουλεύουν εξίσου καλά τόσο με συνεχής όσο και με κατηγορικές μεταβλητές. Στη παρούσα μελέτη χρησιμοποιείται ο αλγόριθμος του δέντρου απόφασης για ανάλυση παλινδρόμησης. Αυτό πρακτικά σημαίνει ότι ο αλγόριθμος παρατηρεί χαρακτηριστικά ενός αντικειμένου και εκπαιδεύει ένα μοντέλο ακολουθώντας τη δομή ενός δέντρου, με σκοπό τη πρόβλεψη μελλοντικών δεδομένων που θα παράξουν χρήσιμες σε πληροφορία συνεχής (μη διακριτές) εξόδους.

Random Forest

Ο αλγόριθμος Random Forest είναι μια τεχνική ανσάμπλ (ensemble) ικανή να εφαρμοστεί τόσο σε προβλήματα ταξινόμησης, όσο και παλινδρόμησης, με την χρήση πολλαπλών δέντρων αποφάσεων και την τεχνική Bootstrap και Aggregation, γνωστή και ως bagging. Εν συντομία, η βασική ιδέα του αλγορίθμου είναι ο συνδυασμός πολλαπλών δέντρων αποφάσεων για τη λήψη της τελικής απόφασης. Για αυτό ακριβώς τον λόγο, αναμένεται στα αποτελέσματα τα σημεία στο χώρο να είναι πιο απλωμένα και η απόδοση του αλγορίθμου να ξεπερνάει σίγουρα αυτή του Decision Tree, προσαρμόζοντας στα δεδομένα καλύτερη ευθεία συσχέτισης από τις υπόλοιπες.

Linear Regression

Η Γραμμική Παλινδρόμηση (Linear Regression) επιχειρεί να μοντελοποιήσει τη σχέση μεταξύ δύο μεταβλητών εφαρμόζοντας μια γραμμική εξίσωση στα παρατηρούμενα δεδομένα. Για το σκοπό αυτό, υπάρχει μια επεξηγηματική μεταβλητή και μια εξαρτημένη μεταβλητή.

Ridge Regression

Η παλινδρόμηση Ridge χαρακτηρίζεται εν συντομία ως ένα μοντέλο που εφαρμόζεται στην ανάλυση δεδομένων, τα οποία υποφέρουν από πολυσυγγραμμικότητα. Αυτή η μέθοδος εφαρμόζει την L2 ομαλοποίηση. Όταν εμφανιστεί το πρόβλημα της πολυσυγγραμμικότητας, τα ελάχιστα τετράγωνα είναι αμερόληπτα (unbiased Least - Squares) και οι διακυμάνσεις είναι μεγάλες, πράγμα που σηματοδοτεί ότι τα αποτελέσματα στις προβλεφθείσες τιμές απέχουν κατά πολύ από τις πραγματικές τιμές. Για την καταπολέμηση της πολυσυγγραμμικότητας η παλινδρόμηση Ridge συρρικνώνει τις παραμέτρους και ελαττώνει την πολυπλοκότητα του μοντέλου μέσω της συρρίκνωσης των συντελεστών.

Elastic Net
Το Elastic Net είναι ένας τύπος ομαλοποιημένης γραμμικής παλινδρόμησης που συνδυάζει δύο συναρτήσεις ποινής (penalty functions), την L1 και την L2.

5.2. Μεθοδολογία - Μοντελοποίηση για την ανάλυση συναισθήματος

Για την επιτυχή διεκπεραίωση του προβλήματος της ανάλυσης συναισθήματος εφαρμόστηκαν τέσσερα προβλεπτικά μοντέλα. Πιο συγκεκριμένα:

- Decision Tree
- Naive Bayes
- Logistic Regression
- Random Forest

Στόχος είναι η εξόρυξη ανθρώπινου συναισθήματος μέσω της ανάγνωσης κριτικής για ένα συγκεκριμένο κατάλυμα. Έχοντας τις κριτικές και τις αντίστοιχες βαθμολογίες, είναι δυνατό να κατηγοριοποιηθούν μερικές λέξεις-κλειδιά ως λέξεις που φανερώνουν ενθουσιασμό, ικανοποίηση ή μη. Για να επιτευχθεί αυτό, εφαρμόστηκε ανάλυση ταξινόμησης, με την χρήση των παραπάνω μοντέλων.
Για το κάθε μοντέλο ταξινόμησης, πραγματοποιήθηκε εκπαίδευση (train) και ταίριασμα (fit) των δεδομένων στο εκάστοτε μοντέλο. Για την αξιολόγηση τους, χρησιμοποιήθηκαν οι μετρικές ακρίβεια (accuracy), F1-Score, Precision και Recall. Ως βέλτιστος, κρίθηκε ο ταξινομητής με τις μέγιστες τιμές των παραπάνω μετρικών.

Decision Tree
Τα Δέντρα Αποφάσεων (Decision Trees) για ταξινόμηση συμπεριφέρονται αντίστοιχα όπως και για την παλινδρόμηση που προαναφέρθηκε.

Naive Bayes
Ο Αφελής Μπεϋζιανός ταξινομητής (Naive Bayes classifier) είναι ο απλούστερος ταξινομητής που μπορεί να εφαρμοστεί σε δεδομένα προς ταξινόμηση. Βασίζεται στο Θεώρημα του Bayes με την υπόθεση ότι υπάρχει κανονικότητα και ανεξαρτησία μεταξύ των χαρακτηριστικών.

Logistic Regression
Η Λογιστική Παλινδρόμηση (Logistic Regression) είναι ένας αλγόριθμος ταξινόμησης που εκτιμά την πιθανότητα μια παρατήρηση να ανήκει σε μια κλάση, μέσω της εκτίμησης μέγιστης πιθανοφάνειας. Εφόσον οι κλάσεις των συναισθημάτων είναι τρεις για την ταξινόμηση χρησιμοποιήθηκε η πολυωνυμική (multinomial) λογιστική παλινδρόμηση.

Random Forest
Ο αλγόριθμος Random Forest για ταξινόμηση συμπεριφέρεται αντίστοιχα όπως και για την παλινδρόμηση που προαναφέρθηκε.

6. Αποτελέσματα και Συζήτηση

6.1. Αποτελέσματα πρόβλεψης τιμής

Στο πρόβλημα της πρόβλεψης ενδεδειγμένης τιμής μίσθωσης, οι αλγόριθμοι που εφαρμόστηκαν, αρχικά χρησιμοποιήθηκαν με τις προεπιλεγμένες τιμές υπερπαραμέτρων. Όμως, σε κάθε δυνατή περίπτωση πραγματοποιήθηκε βελτιστοποίηση με την μέθοδο του grid search και εφαρμόστηκε το σχήμα διασταυρούμενης επικύρωσης. Αναλυτικότερα, περιλαμβάνεται η εκπαίδευση του εκάστοτε μοντέλου στις προκαθορισμένες παραμέτρους του, ο έλεγχος του σφάλματος και της μετρικής R2 και τέλος η σύγκριση των αποτελεσμάτων μεταξύ των μοντέλων. Η δεύτερη ενότητα περιλαμβάνει το κομμάτι μετά τον συντονισμό των υπερπαραμέτρων, δηλαδή τις όσο κατά το δυνατόν βέλτιστες τιμές τους. Πιο συγκεκριμένα, πραγματοποιείται ο καθορισμός της παραμέτρου πλέγματος (grid), εν συνεχεία μια αναζήτηση πλέγματος για την εύρεση των βέλτιστων παραμέτρων (grid search) και τέλος η σύγκριση των αποτελεσμάτων μεταξύ των μοντέλων. Η τεχνική της αναζήτησης πλέγματος για τον συντονισμό των υπερπαραμέτρων, είναι αποδοτική σε χρόνο και χρησιμοποιώντας όσο το δυνατό λιγότερους υπολογιστικούς πόρους και προσφέροντας ευκολία στον αναλυτή, προσπαθεί κάθε φορά να εντοπίσει τις βέλτιστες τιμές τους, δρώντας πάντα εξαντλητικά.
Στον αλγόριθμο KNN επιχειρήθηκε ο συντονισμός και η βελτιστοποίηση παραμέτρων του αλγορίθμου, όπως το πλήθος των γειτόνων και η συνάρτηση εύρεσης της απόστασης (Euclidean, Manhattan). Στη συνέχεια εφαρμόστηκε η τεχνική της διασταυρούμενης επικύρωσης (cross validation) με τη χρήση 5 folds. Ο διαχωρισμός μεταξύ συνόλου εκπαίδευσης και συνόλου ελέγχου έγινε σε ποσοστό 80% και 20% αντίστοιχα. Τέλος, εφαρμόστηκε ανάλυση πλέγματος με διασταυρούμενη επικύρωση για την ανακάλυψη του βέλτιστου μοντέλου, αναζητώντας τον βέλτιστο αριθμό γειτόνων. Η μέθοδος αυτή υπολογίζει όλες τις μετρικές και εντοπίζει το μοντέλο με τις βέλτιστες. Οι βελτιστοποιημένες τιμές των υπερπαραμέτρων για τον αλγόριθμο KNN είναι n_neighbors = 15 και p = 1, με βέλτιστο σκορ είναι περίπου ίσο με 1148.
Με την ίδια λογική που κατασκευάστηκε ο Best KNN από τον Default KNN, δημιουργήθηκε και ο Best SVR από τον Default SVR. Πραγματοποιήθηκαν δοκιμές για τις τιμές των υπερπαραμέτρων loss, C, dual για διαφορετικό εύρος της υπερπαραμέτρου C. Οι πιθανές τιμές του loss function είναι είτε L1, είτε L2 και της dual είναι είτε Αληθές, είτε Ψευδές. Όπως και πριν, μέσω της εφαρμογής αναζήτησης πλέγματος με διασταυρούμενη επικύρωση προέκυψε το βέλτιστο μοντέλο. Οι βελτιστοποιημένες τιμές των υπερπαραμέτρων του αλγορίθμου είναι C = 10, dual = True, loss = squared_epsilon_insensitive και tol = 0.0001, με βέλτιστο σκορ περίπου ίσο με 1229.
Στην περίπτωση των Decision Trees οι υπερπαράμετροι που λήφθηκαν υπόψη είναι η max_depth, η οποία εκφράζει το μέγιστο βάθος ενός δέντρου, η min_samples_leaf, η οποία εκφράζει το ελάχιστο αριθμό δειγμάτων που θα χρειαστούν σε έναν κόμβο φύλλου (leaf node) και η min_samples_split, η οποία εκφράζει τον ελάχιστο αριθμό δειγμάτων που θα χρειαστούν για τη διχοτόμηση ενός εσωτερικού κόμβου (split internal node). Στη συνέχεια, όπως και στις προηγούμενες περιπτώσεις θα εφαρμοστεί αναζήτηση πλέγματος με διασταυρούμενη επικύρωση για τον εντοπισμό του βέλτιστου μοντέλου. Οι βελτιστοποιημένες τιμές των υπερπαραμέτρων του αλγορίθμου είναι max_depth = 6, min_samples_leaf = 3 και min_samples_split = 3, με βέλτιστο σκορ περίπου ίσο με 1361.
Καθώς δεν ήταν εφικτή η άμεση εφαρμογή συντονισμού υπερπαραμέτρων στο μοντέλο της γραμμικής παλινδρόμησης, ο συντονισμός αυτός θα επιτευχθεί μέσω της χρήσης της παλινδρόμησης Ridge.
Στο πρόβλημά μας, δοκιμάστηκε αρχικά η παλινδρόμηση Ridge για τις προκαθορισμένες παραμέτρους, καθώς και διασταυρούμενη επικύρωση για διάφορες τιμές της υπερπαραμέτρου alpha. Τα αποτελέσματα κρίθηκαν μη

ικανοποιητικά και αυτό οδήγησε στη χρήση αναζήτησης πλέγματος με διασταυρούμενη επικύρωση. Τα νέα αποτελέσματα για την τιμή alpha=100 του βέλτιστου μοντέλου είναι σαφώς βελτιωμένα και επιτεύχθηκε ο συντονισμός των παραμέτρων. Και πάλι όμως, δεν παρατηρείται αισθητή διαφορά στα αποτελέσματα της παλινδρόμησης Ridge συγκριτικά με την γραμμική παλινδρόμηση.

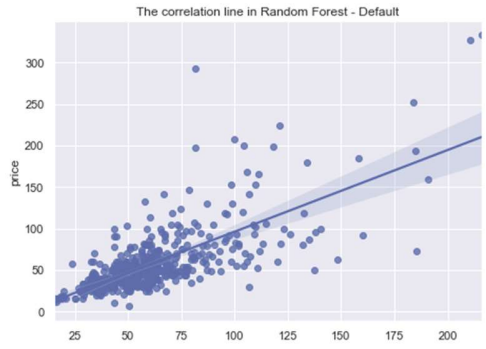
Για την αξιολόγηση των παραπάνω μεθοδολογιών στο πρόβλημα της πρόβλεψης τιμών χρησιμοποιήθηκαν οι μετρικές R2 και μέσο τετραγωνικό σφάλμα (mean squared error). Ο λόγος που επιλέξαμε τις παραπάνω μετρικές είναι ότι αποτελούν μια ένδειξη της επιτυχημένης προσαρμογής του μοντέλου στα δεδομένα.

Στο κομμάτι της εκτίμησης ενδεδειγμένης τιμής καταλύματος, αποδείχθηκε ότι ο καταλληλότερος αλγόριθμος είναι ο Random Forest. Όπως διακρίνεται και στον Πίνακα (2), ο αλγόριθμος αυτός συγκεντρώνει το υψηλότερο R2 και ταυτόχρονα, ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα.

Πίνακας 2: MSE και R2 ανά αλγόριθμο.

	Mean Squared Error	R2 Score
KNN Default	989.586581	0.367491
MSE KNN-Best	871.450551	0.443000
SVR Default	973.490964	0.377779
SVR Best	969.696895	0.380204
Decision Tree - Best	1377.081816	0.119818
Random Forest - Default	777.779012	0.502871
Linear Regression - Default	879.545749	0.437825
Ridge	878.846624	0.438272
ElasticNet	919.994818	0.411972

Στο Σχήμα (10) παρουσιάζεται η προσαρμογή της καμπύλης του μοντέλου στα παρεχόμενα δεδομένα. Τελικά όντως, τα σημεία στο χώρο είναι πιο απλωμένα και η απόδοση του αλγορίθμου να ξεπερνάει σίγουρα αυτή του Decision Tree, αφού πρακτικά τρέχει πολλά Decision Trees, αλλά και από τους υπόλοιπους αλγορίθμους, προσαρμόζοντας στα δεδομένα καλύτερη ευθεία συσχέτισης από τις υπόλοιπες.



Σχήμα 10: Γραμμή συσχέτισης για τον Random Forest.

Στον Πίνακα (3) παρουσιάζονται οι πραγματικές τιμές αξίας ενός καταλύματος σε σχέση με τις προβλέψεις του εκάστοτε αλγορίθμου. Ναι μεν, ο αλγόριθμος Random Forest έχει αστοχίες, αλλά αυτές είναι σίγουρα λιγότερες σε σχέση με τους υπόλοιπους, αφού έχει επιλεχθεί ως ο βέλτιστος, σύμφωνα με τις μετρικές που αναφέραμε. Πρακτικά εάν ο ενοικιαστής ενός καταλύματος αποφασίσει να το κοστολογήσει πάνω από την υποδεικνυόμενη τιμή, τότε κατά πάσα πιθανότητα θα έχει υπερεκτιμήσει την τελική τιμή και θα προσελκύσει λιγότερο κόσμο και στο αντίθετο σενάριο θα έχει υποεκτιμήσει την αξία του καταλύματός του και τελικά δεν θα πετύχει μεγιστοποίηση του κέρδους του.

Πίνακας 3: Πραγματικές έναντι προβλεφθείσες τιμές ανά αλγόριθμο.

Actual Values	KNN-Default	KNN-Best	SVR-Default	SVR-Best	Decision Tree - Best	Random Forest - Default	Linear Regression - Default	Ridge	ElasticNet
65.0	54.8	54.000000	42.865475	42.835552	59.761905	62.18	44.475196	46.338391	52.717573
60.0	55.4	47.666667	19.313208	18.851910	32.777778	32.42	12.081109	11.555485	19.207116
36.0	48.8	55.133333	40.580089	40.731519	43.244898	46.60	43.539798	44.773946	49.398091
35.0	53.2	48.933333	41.138077	41.133135	33.745833	49.31	49.947370	50.098036	49.140490
32.0	61.6	60.866667	43.243872	43.295085	52.327869	59.72	51.824018	51.914298	51.072663
37.0	43.4	39.000000	49.450159	49.615191	43.244898	51.64	58.989951	58.846135	57.164697
40.0	50.4	54.333333	56.497792	56.775092	45.878378	52.57	63.739993	64.271574	65.897711
39.0	31.6	31.400000	43.462886	43.206346	33.745833	36.68	55.618486	55.290869	53.245065
41.0	43.8	45.733333	34.868330	34.609859	43.244898	48.09	37.527518	37.397643	42.181624
75.0	51.6	45.800000	43.110387	43.041736	57.102564	51.15	48.416694	48.914661	51.811563
55.0	63.2	61.666667	65.744951	65.664350	119.869565	100.66	79.977370	77.821954	73.358254
23.0	28.8	23.666667	19.209147	19.074382	20.473684	28.66	15.994453	17.513696	26.241422
44.0	48.0	48.933333	38.332420	38.260448	45.878378	45.44	39.518427	39.751297	44.659418
21.0	24.8	44.600000	43.587008	43.449968	43.244898	47.97	51.582083	51.874385	52.093155
24.0	52.0	55.333333	44.269959	44.390092	33.745833	37.94	51.968031	51.890282	51.629417
25.0	30.8	45.066667	47.358575	47.318103	39.821429	42.47	52.269790	52.906529	55.782294
73.0	73.6	76.000000	51.840291	51.659766	45.878378	63.12	63.136054	63.263146	61.901019
58.0	73.2	77.200000	66.671851	66.812688	91.000000	86.66	80.131575	79.340598	77.443614
49.0	48.0	46.133333	41.540703	41.370673	57.102564	38.38	46.178113	46.238499	49.764108
20.0	37.4	35.666667	28.164538	28.010618	33.745833	34.90	25.497316	25.691207	33.739934

6.2. Αποτελέσματα ανάλυσης συναισθήματος

Όσον αφορά το πρόβλημα της ανάλυσης συναισθήματος, επιλέχθηκαν οι μετρικές ορθότητα (accuracy) και F1-score. Ο λόγος για τον οποίο επιλέχθηκε η ορθότητα είναι ότι στο παρόν πρόβλημα, είναι πρωταρχικής σημασίας η ορθή αξιολόγηση του συναισθήματος του καταναλωτή από τις παρεχόμενες υπηρεσίες ώστε, να παρέχεται η δυνατότητα στον ιδιοκτήτη του καταλύματος να αυτοαξιολογηθεί αλλά και στην πλατφόρμα να ποσοτικοποιεί ορθά την εισερχόμενη πληροφορία. Επιπλέον, η μετρική F1-score προτιμήθηκε με σκοπό να ληφθεί υπόψη τόσο η ακρίβεια (precision), όσο και η ανάκληση (recall) για τα μελετώμενα σύνολα δεδομένων.

Στον Πίνακα (4) συγκεντρώνονται οι τιμές των υπολογιζόμενων μετρικών για τους αλγόριθμους που χρησιμοποιήθηκαν. Παρατηρείται ότι ο αλγόριθμος Logistic Regression παρουσιάζει τα βέλτιστα αποτελέσματα με μικρή διαφορά από τον δεύτερο καλύτερο, Random Forest ο οποίος αποδίδει ικανοποιητικά και στα δύο προβλήματα.

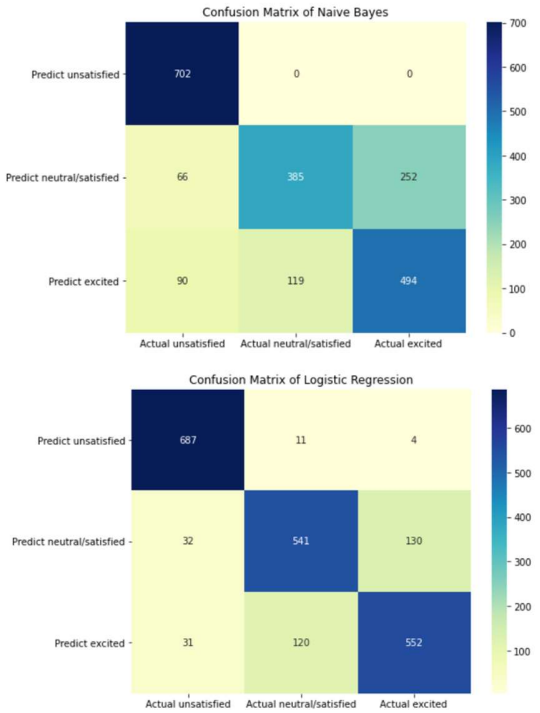
Πίνακας 4: Μετρικές αλγορίθμων ανάλυσης συναισθήματος.

	Accuracy	F1-Score	Precision	Recall
Desicion Tree	0.7903	0.7877	0.7866	0.7904
Naive Bayes	0.7500	0.7399	0.7481	0.7501
Logistic Regression	0.8444	0.8427	0.8419	0.8445
Random Forest	0.8397	0.8396	0.8405	0.8397

Για την οπτικοποίηση των αποτελεσμάτων των διάφορων ταξινομητών, χρησιμοποιήθηκε και ο confusion matrix, ο οποίος είναι ένας πίνακας σφαλμάτων. Στην διαγώνιό του έχει το πλήθος των σωστά ταξινομημένων δειγμάτων και εκτός διαγωνίου το πλήθος των εσφαλμένα ταξινομημένων δειγμάτων. Επιθυμητή είναι η εμφάνιση όσο το δυνατόν μικρότερων τιμών εκτός της διαγωνίου. Στο Σχήμα (11) παρουσιάζονται οι πίνακες σύγχυσης για τους επιλεγμένους αλγορίθμους. Γίνεται αμέσως κατανοητό ότι οι αλγόριθμοι Logistic Regression και Random Forest εμφανίζονται ικανότεροι να διακρίνουν και να κατηγοριοποιήσουν ορθά το συναίσθημα neutral/satisfied και excited. Αντιθέτως, φαίνεται ότι οι ταξινομητές Decision Tree και Naive Bayes συχνά κατηγοριοποιούν λάθος τις παραπάνω κλάσεις.

(α)

(β)



Σχήμα 11: Πίνακας σύγχυσης για τις ταξινομήσεις (α) του Naive Bayes (χειρότερος ταξινομητής) και (β) της Logistic Regression (καλύτερος ταξινομητής).

7. Συμπεράσματα, Προκλήσεις και Μελλοντική Έρευνα

Ένα από τα βασικότερα συμπεράσματα της παρούσας μελέτης είναι ότι τα αποτελέσματά της συμβάλλουν άμεσα στην ικανοποίηση των συμφερόντων και των τριών εμπλεκομένων στην διαδικασία βραχυπρόθεσμης ενοικίασης ενός καταλύματος μέσω της ιστοσελίδας Airbnb, τόσο σε αυτά του ενοικιαστή, όσο και του ενοίκου, αλλά και της ίδιας της εταιρείας. Παρέχεται πρακτικά στον ενοικιαστή μια προτεινόμενη τιμή για την προσθήκη του καταλύματός του στην ιστοσελίδα, όσο το δυνατόν πιο ρεαλιστική και εύστοχη, σύμφωνα βέβαια με τον βέλτιστο προβλεπτικό αλγόριθμο. Επιτυγχάνεται, δηλαδή, παράλληλα η μεγιστοποίηση των εσόδων του ενοικιαστή σε συνδυασμό με τον βαθμό ικανοποίησης του ενοίκου και εν τέλη την μεγιστοποίηση των εσόδων της ίδιας της εταιρείας Airbnb.

Η έρευνα αυτή είναι ευεργετική όσον αφορά και τη δημιουργία μιας ιστοσελίδας στην οποία ένας νέος ένοικος απλώς τοποθετεί ως εισόδους τα διάφορα χαρακτηριστικά του καταλύματός του και με το πάτημα ενός κουμπιού, εκτελούνται υπολογισμοί αυτόματα, των οποίων το αποτέλεσμα εμφανίζεται ως νέα προτεινόμενη τιμή στον ενδιαφερόμενο χρήστη της, ακόμη και σε πραγματικό χρόνο. Σε αυτά τα αποτελέσματα συνδράμουν και τα δύο σκέλη της έρευνάς μας, αυτά της πρόβλεψης τιμής και της ανάλυσης συναισθήματος.

Μια πιθανή πρόκληση αποτελεί το γεγονός ότι μέσω της ευρείας χρήσης του αλγορίθμου μας, θα οδηγηθούμε σε μια πολύ πιο ανταγωνιστική αγορά και συνεπώς όλες οι τιμές θα τοποθετούνται σε παρόμοια εύρη, εξαιτίας της σύγκρισης παρόμοιων σε χαρακτηριστικά καταλυμάτων και τελικά, τα υπόλοιπα χαρακτηριστικά, ενδέχεται να θεωρηθούν πιο σημαντικά από το χρήστη.

Όσον αφορά την μελλοντική έρευνα, αξίζει να σημειωθεί ότι δεν λήφθηκαν υπόψη εξωγενείς παράγοντες, όπως η γειτνίαση ενός καταλύματος σε κάποιο αξιοθέατο ή χώρο τουριστικής σημασίας ή οτιδήποτε επιπρόσθετο που θα μπορούσε εν γένη να προσελκύσει περισσότερους χρήστες και είναι παράμετρος την οποία επισημαίνουμε και αξίζει να μελετηθεί περαιτέρω από τους μελλοντικούς ερευνητές.

Ακόμη, θα ήταν ωφέλιμο να επαληθευτούν και να ξεπεραστούν αυτά τα αποτελέσματα μετά από μερικά χρόνια, έχοντας περισσότερα δεδομένα για νέες εγγραφές καταλυμάτων στην Airbnb, αρκετά ώστε να μην χρειαστεί η χρήση τεχνικών υπερδειγματοληψίας.

Η υλοποίηση και εφαρμογή των αλγορίθμων της παρούσας μελέτης σε νέα σύνολα δεδομένων, τα οποία δεν έχουν ερευνηθεί, είναι ένα απλό παράδειγμα για επιπρόσθετη διερεύνηση.

Τέλος, σημειώνεται ότι στο επιχειρηματικό πρόβλημα στο οποίο εστιάσαμε για αυτή την έρευνα, δεν συμπεριλάβαμε ηθικούς περιορισμούς, καθώς αντιμετωπίστηκε ως ένα καθαρά εμπορικό θέμα.

8. Βιβλιογραφία

1. http://insideairbnb.com/get-the-data.html
2. Gourzis, K., Alexandridis, G., Gialis, S., & Caridakis, G. (2019). Studying the spatialities of short-term rentals' sprawl in the urban fabric: The case of airbnb in Athens, Greece. IFIP Advances in Information and Communication Technology, 196–207. https://doi.org/10.1007/978-3-030-19909-8_17
3. Alexandridis, G., Voutos, Y., Mylonas, P., & Caridakis, G. (2020). A geolocation analytics-driven ontology for short-term leases: Inferring Current Sharing Economy Trends. Algorithms, 13(3), 59. https://doi.org/10.3390/a13030059
4. Rezazadeh Kalehbasti, P., Nikolenko, L., & Rezaei, H. (2021). Airbnb price prediction using machine learning and sentiment analysis. *Lecture Notes in Computer Science*, 173–184. https://doi.org/10.1007/978-3-030-84060-0_11
5. Zhu, A., Li, R., & Xie, Z. (2020). Machine learning prediction of new york airbnb prices. *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*. https://doi.org/10.1109/ai4i49448.2020.00007
6. Tang, E. & Sangani, K. (2015). Neighborhood and Price Prediction for San Francisco Airbnb Listings.
7. Li, Y., Pan, Q., Yang, T., & Guo, L. (2016). Reasonable price recommendation on Airbnb using multi-scale clustering. *2016 35th Chinese Control Conference (CCC)*. https://doi.org/10.1109/chicc.2016.7554467
8. Priambodo, F. N., & Sihabuddin, A. (2020). An extreme learning machine model approach on Airbnb base price prediction. *International Journal of Advanced Computer Science and Applications*, *11*(11). https://doi.org/10.14569/ijacsa.2020.0111123
9. Cai, T., & Han, K.G. (2019). Melbourne Airbnb Price Prediction