ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΙΤΛΟΣ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ: «ΕΚΤΙΜΗΣΗ ΤΙΜΗΣ ΒΡΑΧΥΧΡΟΝΙΑΣ ΜΙΣΘΩΣΗΣ ΑΚΙΝΗΤΟΥ ΜΕ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ»

Ονοματεπώνυμο **Αριθμός Μητρώου**

ΓΙΑΝΝΑΚΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ **Π2013057**

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΘΕΜΙΣΤΟΚΛΗΣ ΈΞΑΡΧΟΣ

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ: ΚΑΤΙΑ ΛΗΔΑ ΚΕΡΜΑΝΙΔΟΥ, ΠΑΝΑΓΙΩΤΗΣ ΚΟΥΡΟΥΘΑΝΑΣΗΣ

ΚΕΡΚΥΡΑ 12/02/2020

Περιεχόμενα

Περίληψη

Σήμερα, οι αλγόριθμοι τεχνητής νοημοσύνης χρησιμοποιούνται ευρέως για τη λύση ποικίλων προβλημάτων σε διαφορετικά πεδία της καθημερινής ζωής. Συγκεκριμένα, χρησιμοποιούνται αλγόριθμοι μηχανικής εκμάθησης για να δημιουργηθεί ένα μοντέλο από γνωστά δεδομένα, το οποίο δέχεται ως είσοδο νέα άγνωστα δεδομένα και μπορεί να πραγματοποιεί προβλέψεις. Επιπλέον, η χρήση νέων τεχνολογιών για την πραγματοποίηση κάποιας δραστηριότητας παράγει ένα μεγάλο όγκο δεδομένων, γνωστό ως "Big Data", τα οποία χρειάζονται εξειδικευμένες τεχνικές για αποτελεσματική διαχείριση.

Οπότε, σε αυτή τη διπλωματική, ο σκοπός μας είναι η μελέτη ενός προβλήματος με χρήση μεθόδων μηχανικής εκμάθησης ώστε να δημιουργηθεί ένα μοντέλο με υψηλή ακρίβεια πρόβλεψης. Για αυτό, μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι μηχανικής εκμάθησης και αν συγκριθούν τα αποτελέσματά τους, όπως για παράδειγμα οι αλγόριθμοι K-Nearest-Neighbors (KNN), Support Vector Machines (SVM), Decision Trees (καθώς και πιο εξελιγμένοι αλγόριθμοι που βασίζονται σε δέντρα αποφάσεων, όπως ο Random Forest) και άλλοι. Για κάθε αλγόριθμο θα εξεταστούν διαφορετικές τιμές στις παραμέτρους εισόδου, ώστε να μεγιστοποιηθεί η ακρίβεια πρόβλεψης και να καταλήξουμε στην καλύτερη μέθοδο για το συγκεκριμένο σύνολο δεδομένων. Η υλοποίηση θα γίνει σε Python, μια γλώσσα προγραμματισμού που διευκολύνει τη χρήση μεθόδων μηχανικής εκμάθησης.

Ως δεδομένα εισόδου, θα χρησιμοποιηθεί κάποιο σύνολο δεδομένων δημόσια διαθέσιμο, από αποθετήρια όπως για παράδειγμα το kaggle.com. Ένα τέτοιο σύνολο δεδομένων είναι το “Berlin AirBnB Data”, όπου ο στόχος είναι να προβλεφθεί η τιμή ενοικίασης ενός διαμερίσματος με βάση τα στοιχεία της πλατφόρμας AirBnB. Το πρόβλημα μπορεί είτε να αντιμετωπιστεί ως πρόβλημα παλινδρόμησης, δηλαδή να προβλέψουμε μια συνεχή μεταβλητή (εδώ την τιμή), είτε ως πρόβλημα κατηγοριοποίησης, όπου οι τιμές θα ομαδοποιηθούν σε κατηγορίες, π.χ. «φτηνό», «ακριβό» και θα γίνει πρόβλεψη της κατηγορίας. Επίσης, θα πραγματοποιηθεί προ-επεξεργασία των μεταβλητών εισόδου των δεδομένων για αποτελεσματικότερη πρόβλεψη.

1. Πρόβλεψη τιμής εκμίσθωσης ακινήτων – Το πρόβλημα

Η ακίνητη περιουσία δεν είναι μόνο η βασική ανάγκη ενός ανθρώπου, αλλά σήμερα αντιπροσωπεύει επίσης το πλούτο και το κύρος ενός ατόμου. Οι επενδύσεις σε ακίνητα γενικά φαίνεται να είναι επικερδείς επειδή οι τιμές των ακινήτων τους δεν μειώνονται γρήγορα. Οι μεταβολές στην τιμή των ακινήτων μπορούν να επηρεάσουν διάφορους επενδυτές, τραπεζίτες, υπεύθυνους χάραξης πολιτικής και πολλά ακόμη άτομα. Οι επενδύσεις στον τομέα των ακινήτων φαίνεται να αποτελούν ελκυστική επιλογή για τους επενδυτές. Οι επενδύσεις είναι μια επιχειρηματική δραστηριότητα για την οποία οι περισσότεροι άνθρωποι ενδιαφέρονται αυτήν την εποχή της παγκοσμιοποίησης. Υπάρχουν διάφορα αντικείμενα που χρησιμοποιούνται συχνά για επενδύσεις, όπως χρυσός, αποθέματα και ακίνητα. Ειδικότερα, οι επενδύσεις σε ακίνητα αυξήθηκαν σημαντικά από το 2011, τόσο στη ζήτηση όσο και στην πώληση ακινήτων (Chouthai et al., 2019).

Σε παγκόσμια κλίμακα, η αύξηση των τιμών των ακινήτων στις μητροπολιτικές περιοχές έχει οδηγήσει στο πρόβλημα της προσιτότητας των κατοικιών που αντιμετωπίζουν οι ομάδες χαμηλού εισοδήματος (Hui et al., 2016 · Zhang et al., 2016). Οι μετανάστες, τα νεαρά άτομα και τα νιόπαντρα νοικοκυριά χρειάζονται όλες τις προσιτές τιμές που μπορεί να προσφέρει η κατοικία τους. Κατά συνέπεια, η ενοικίαση κατοικιών έχει γίνει σταδιακά ένα απαραίτητο συμπλήρωμα στην ιδιοκτησία κατοικίας. Σε ολόκληρο τον κόσμο, περίπου 1.2 δισεκατομμύρια άνθρωποι ζουν σε ενοικιαζόμενα καταλύματα (Gilbert, 2016). Συγκεκριμένα, η τιμή ενοικίασης ξεπέρασε το 50% στις διεθνείς πόλεις όπως η Νέα Υόρκη, το Λος Άντζελες, το Σενζέν και η Σαγκάη, σύμφωνα με τα στατιστικά στοιχεία που δημοσίευσε η China Index Academy

(http://industry.fang.com/). Στις περισσότερες περιπτώσεις, η αγορά ενοικίασης λειτουργεί με βάση ιδιωτικές συναλλαγές, προκαλώντας κρυφές νομικές διαφορές και προβλήματα ανισότητας (Wu, 2016, Zheng et al., 2018). Σε απάντηση, για την παρακολούθηση των τιμών ενοικίασης κατοικιών (Housing Rental Prices - HRPs) έχει δοθεί προτεραιότητα στις κυβερνητικές ατζέντες για τη δημιουργία μιας υγιούς αγοράς ενοικίασης κατοικιών. Επιπλέον, οι τιμές ενοικίασης κατοικιών παρουσιάζουν μεγάλη μεταβλητότητα ανάλογα με την τοποθεσία σε αστικές περιοχές. Ο τρόπος με τον οποίο μπορεί να συλληφθούν αποτελεσματικά και ταχέως οι τιμές ενοικίασης κατοικιών μικρής κλίμακας έχει εξελιχθεί σε επείγον ζήτημα όσον αφορά τη χρήση της γης και τις μελέτες στέγασης.

Ένας μεγάλος όγκος βιβλιογραφίας υπάρχει σχετικά με τις τιμές ενοικίασης κατοικιών σε πολλές πόλεις σε όλο τον κόσμο, όπως το Βερολίνο, το Τορόντο, το Σίδνεϊ, το Χονγκ Κονγκ και το Chongqing (August and Walks, 2018, Waltl, 2018). Εντούτοις, η παρακολούθηση των τιμών ενοικίασης κατοικιών μικρής κλίμακας παραμένει μια μεγάλη πρόκληση. Παραδοσιακά, συγκεντρωτικά δεδομένα για τιμές ενοικίασης κατοικιών λαμβάνονται μέσω επίσημων στατιστικών στοιχείων (Rondinelli και Veronese, 2011), πληθυσμιακές απογραφές (Yi and Huang, 2014) και έρευνες ερωτηματολογίων (Li et al., 2017). Είναι αρκετά καταναγκαστικό και χρονοβόρο να συλλέγονται τέτοια δεδομένα. Επομένως, τα δεδομένα που συλλέγονται παρουσιάζουν χονδρική ανάλυση χρόνου και χώρου. Συνήθως, τα επίσημα δεδομένα για τις τιμές ενοικίασης κατοικιών απελευθερώνονται ετησίως με έναν αριθμό για μια πόλη στην Κίνα. Χάρη στη δημοτικότητα των κοινωνικών μέσων ενημέρωσης, η ανταλλαγή πληροφοριών και η προσφορά έγιναν δυνατές για επιχειρήσεις από επιχείρηση σε

επιχείρηση (peer-to-peer) και σε επιχειρηματίες από ομότιμους (Park and Nicolau, 2015). Οι διαδικτυακές ιστοσελίδες ενοικίασης κατοικιών (On-Line Housing Rental Websites - OHRW) έχουν αναδειχθεί ως δημοφιλή πλατφόρμες κοινωνικών μέσων στις μελέτες κατοικιών. Οι διαδικτυακές ιστοσελίδες ενοικίασης κατοικιών έχουν τρία βασικά χαρακτηριστικά όπως τα κοινωνικά μέσα, συμπεριλαμβανομένων των διαδραστικών εφαρμογών που βασίζονται στο Διαδίκτυο Web 2.0, το περιεχόμενο που δημιουργείται από τους χρήστες και διευκολύνουν την ανάπτυξη διαδικτυακών κοινωνικών δικτύων συνδέοντας το προφίλ ενός χρήστη με εκείνα άλλων ατόμων ή ομάδων. Η διαφορά μεταξύ των διαδικτυακών ιστοσελίδων ενοικίασης κατοικιών και των γενικών κοινωνικών μέσων είναι ότι ο χρήστης είναι ιδιοκτήτης. Αξιόλογα παραδείγματα με διαδικτυακές ιστοσελίδες ενοικίασης κατοικιών περιλαμβάνουν την Airbnb στις ΗΠΑ (Wang και Nicolau, 2017), την Craigslist στον Καναδά (Hogan and Berry, 2011), την Rightmove στην Αγγλία (Rae και Sener, 2016), την Door Chintai στην Ιαπωνία (Sadayuki, 2018) και την Anjuke στην Κίνα (Chen κ.ά., 2016). Μέσω αυτών των διαδικτυακών ιστοσελίδων ενοικίασης κατοικιών, ο καθένας μπορεί να δημοσιεύσει πληροφορίες ενοικίασης σπιτιών για να παρέχει δεδομένα τιμών ενοικίασης κατοικιών σε πραγματικό χρόνο και εύκολα προσβάσιμα. Επιπλέον, τα δεδομένα αυτά περιλαμβάνουν πλούσιες γεωγραφικές πληροφορίες. Επομένως, η υιοθέτηση και η χρήση διαδικτυακών ιστοσελίδων ενοικίασης κατοικιών θα πρέπει να συμβάλλουν στην κατανόηση της δυναμικής των τιμών ενοικιαζόμενων κατοικιών σε λεπτή κλίμακα.

Παρόλο που οι διαδικτυακές ιστοσελίδες ενοικίασης κατοικιών μπορούν να λειτουργήσουν ως αξιόπιστες πηγές δεδομένων για τιμές ενοικιαζόμενων κατοικιών, δεν μπορούν ωστόσο να καλύψουν όλα τα κτίρια και τις κοινότητες. Η κάλυψη αυτή απαιτεί ισχυρές μεθοδολογίες και εργαλεία για την πρόβλεψη των τιμών για περιοχές χωρίς εγγραφές. Οι περισσότερες μελέτες συνδυάζουν γραμμική παλινδρόμηση με το ηδονικό μοντέλο για την αντιμετώπιση του παραπάνω ζητήματος (Wen and Tao, 2015, Wen et al., 2017, 2014). Το ηδονικό μοντέλο υποθέτει ότι η ακίνητη περιουσία πρέπει να αφορά ετερογενή αγαθά και η τιμή της εξαρτάται από τα χαρακτηριστικά της δομής, της γειτονιάς και της τοποθεσίας (Lancaster, 1976, Rosen, 1974). Αρκετές εφαρμογές έχουν αποδείξει την ικανότητα του ηδονικού μοντέλου ως θεωρητικής βάσης να εξηγήσει τους καθοριστικούς παράγοντες τιμών ενοικιαζόμενων κατοικιών (Gluszak and Zygmunt, 2018, Waltl, 2018, Zhang and Yi, 2017). Ωστόσο, το ηδονικό μοντέλο απλώς μετασχηματίζεται στη μορφή γραμμικής παλινδρόμησης, η οποία περιορίζεται στην ικανότητά του να αποκαλύπτει τις πολύπλοκες και μη γραμμικές σχέσεις μεταξύ των τιμών ενοικιαζόμενων κατοικιών και των καθοριστικών παραγόντων. Έτσι, θα πρέπει να αναζητηθούν εναλλακτικά εύρωστα εργαλεία για την πρόβλεψη των HRP. Οι αλγόριθμοι μηχανικής μάθησης (Machine Learning Algorithms - MLA) εμπίπτουν σε αυτό το συγκεκριμένο πεδίο εφαρμογής επειδή παρουσιάζουν πολλά υποσχόμενα πλεονεκτήματα σε τρεις πτυχές: (1) δυνατότητα διερεύνησης αλληλεπιδράσεων πολλαπλών επιπέδων και μη γραμμικών συσχετίσεων, (2) δεν υπάρχει αυστηρή απαίτηση για παραδοχές, και (3) ικανότητα επεξεργασίας δεδομένων σε διάφορες δομές (Zhang et αϊ., 2017). Οι μελετητές κατέδειξαν την ευρωστία των αλγόριθμων μηχανικής μάθησης σε πρόσφατες μελέτες στέγασης (Chen et al., 2016 · Yao et al., 2018). Δυστυχώς, οι μελέτες δεν έχουν

ενσωματώσει αλγόριθμους μηχανικής μάθησης με ηδονική μοντελοποίηση για να προβλέψουν τις τιμές ενοικιαζόμενων κατοικιών μικρής κλίμακας με βάση τις διαδικτυακές ιστοσελίδες ενοικίασης κατοικιών.

1.1. Τιμές ενοικίασης κατοικιών

Οι καθοριστικοί παράγοντες της τιμής ενοικίασης κατοικιών ποικίλλουν ανάλογα με την κλίμακα. Σε μακροοικονομική κλίμακα, όπως το επίπεδο της πόλης ή της επαρχίας, η τιμή ενοικίασης κατοικιών καθορίζεται συνήθως από οικονομικά θεμελιώδη στοιχεία (Wang et al., 2017). Ορισμένες μελέτες έχουν χρησιμοποιήσει εξωγενείς μακροοικονομικές μεταβλητές όπως το εισόδημα των κατοίκων του διοικητικού επιπέδου, το κόστος του πληθυσμού και το κόστος κατασκευής για να εξηγήσουν τις τιμές κατοικιών ή τις τιμές των ενοικίων (Depetris-Chauvin and Santos, 2018, Mirkatouli et al., 2018). Στο εσωτερικό των αστικών περιοχών, οι διαφορετικές οικιστικές κοινότητες υπόκεινται στους ίδιους μακροοικονομικούς παράγοντες. Αντ 'αυτού, οι κοινωνικές και περιβαλλοντικές ευκολίες γύρω από τις κοινότητες διαδραματίζουν καίριο ρόλο. Πολλές μελέτες έχουν διερευνήσει τους καθοριστικούς παράγοντες της εσωτερικής αστικής κατοικίας ή τις τιμές ενοικίασης με βάση το ηδονικό μοντέλο σε όλο τον κόσμο συμπεριλαμβανομένων των ΗΠΑ (Chen and Fik, 2017), Ελβετία (Waltert και Schläpfer, 2010) , Γαλλία (Choumert et al., 2014), Πολωνία (Gluszak και Zygmunt, 2018), Αυστραλία (Waltl, 2018) και Κίνα (Huang et al., 2017). Το ηδονικό μοντέλο υποθέτει ότι η αξία του σπιτιού μπορεί να εκτιμηθεί με βάση τη δομή, τα χαρακτηριστικά της γειτονιάς και της τοποθεσίας του (Lancaster, 1976, Rosen, 1974).

Δεδομένης της ικανότητας αποκάλυψης του θεωρητικού μηχανισμού, το ηδονικό μοντέλο έχει χρησιμοποιηθεί ευρέως για να καθορίσει τη σχέση μεταξύ τιμής κατοικίας ή τιμής ενοικίου και ποικίλων πιθανών καθοριστικών παραγόντων. Για παράδειγμα, οι ερευνητές έχουν δημιουργήσει μοντέλα για να εξετάσουν τον ρόλο των εκπαιδευτικών ιδρυμάτων (Wen et al., 2014), τη πολυκεντρική αστική δομή (Wen and Tao, 2015) και την δημόσια υποδομή (Geng et al., 2015) για τον προσδιορισμό της τιμής κατοικίας ή τιμή ενοικίασης. Οι Waltert και Schläpfer (2010) εξέτασαν 71 μελέτες και συνόψισαν τις επιπτώσεις των τοπικών ανέσεων στις τιμές κατοικιών. Επιπροσθέτως, οι Huang et al. (2017) ανέλυσαν τα χωρικά και ηδονικά χαρακτηριστικά των τιμών κατοικιών και τις επιπτώσεις της απόστασης στις περιοχές του κέντρου, της ποιότητας του σχολείου και των δομικών ιδιοτήτων του σπιτιού στη Σαγκάη. Ο Sadayuki (2018) μέτρησε εμπειρικά τις χωρικές επιπτώσεις των δημόσιων συγκοινωνιών στις τιμές ενοικίασης κατοικιών στο Τόκιο. Οι Alawadi et al. (2018) αξιολόγησαν την επίδραση της αστικής μορφολογίας και της πολιτικής χρήσης γης στην οικονομική προσιτότητα των κατοικιών στο Ντουμπάι. Οι Chen et al. (2016) εκτιμούν τις τιμές ενοικίασης κατοικιών του Guangzhou σε συνδυασμό με εμπορικά κτίρια, επιχειρήσεις, νοσοκομεία και σταθμούς μετρό.

1.2. Χαρτογράφηση ενοικιαζόμενων κατοικιών: δεδομένα και μέθοδοι

Η χαρτογράφηση των τιμών ενοικίασης κατοικιών σε όλο το χώρο παρέχει βασικές πληροφορίες για την προώθηση της υγιούς και δίκαιης ανάπτυξης της αγοράς κατοικιών. Τα δεδομένα σε προηγούμενες μελέτες προέρχονται συνήθως από απογραφές πληθυσμού, επίσημες στατιστικές βάσεις δεδομένων και έρευνες ερωτηματολογίων (Li et al., 2017). Για παράδειγμα, οι Feng et al. (2011) μελέτησαν τις τιμές κατοικιών στο Πεκίνο με μια έρευνα σχετικά με εμπορικές αστικές συναλλαγές. Οι Rondinelli και Veronese (2011) μετρούσαν τη δυναμική των μισθωμάτων των ενοικιαστών στην Ιταλία με έρευνες των νοικοκυριών και στοιχεία για τις τιμές ενοικίασης από ιδιωτικά πρακτορεία ακινήτων. Ο Yi και ο Huang (2014) αποκάλυψαν την ανισότητα στέγασης στις κινεζικές πόλεις σε ομάδες εκπαίδευσης και διαφορετικές επαρχίες με τα τελευταία στοιχεία απογραφής. Οι Fesselmeyer et al. (2017) μέτρησαν την επίδραση της πυκνότητας κατοικιών στις τιμές κατοικιών στη Σιγκαπούρη με τη βάση δεδομένων του συστήματος πληροφοριών ακινήτων της αρχής γης της Σιγκαπούρης. Αυτοί οι τύποι δεδομένων είναι αρκετά εντατικοί στην εργασία και συλλέγονται σε σχετικά μεγάλη κλίμακα με μακρές ενημερωμένες περιόδους. Έτσι, αντιμετωπίζονται μεγάλα εμπόδια στη χαρτογράφηση των τιμών μίσθωσης κατοικιών σε μια λεπτή κλίμακα βάσει στατιστικών δεδομένων και στοιχείων έρευνας.

Η χρησιμότητα του ηλεκτρονικού ανοίγματος δεδομένων για την καλύτερη κατανόηση των ενδο-αστικών θεμάτων τονίστηκε στο Arribas-Bel (2014). Οι μελετητές χρησιμοποίησαν διαφορετικές διαδικτυακές ιστοσελίδες ενοικίασης κατοικιών για να χαρτογραφήσουν τις τιμές ενοικίασης κατοικιών σε μια λεπτή κλίμακα τα τελευταία χρόνια. Οι Chen et al. (2016) απεικόνισαν το χωρικό πρότυπο των τιμών ενοικιαζόμενων κατοικιών χρησιμοποιώντας την ηλεκτρονική λίστα ενοικίασης που αποκτήθηκε από την Anjuke.com, μια εγχώρια πλατφόρμα ακινήτων στην Κίνα. Οι Wu et al. (2016) χρησιμοποίησαν δεδομένα ελέγχου και ενοικίασης από το Sofang.com για να εξετάσουν τις πιθανές επιπτώσεις των τιμών των κατοικιών από μια μεγάλη προοπτική δεδομένων. Οι Rae και Sener (2016) διερεύνησαν τον τρόπο με τον οποίο οι χρήστες της ιστοσελίδας ταξινομούν μια πόλη βασισμένη στα χωρικά πρότυπα αναζήτησης κατοικιών στο Λονδίνο, χρησιμοποιώντας δεδομένα που παράγονται από χρήστες της πιο δημοφιλούς πύλης ακινήτων του Ηνωμένου Βασιλείου. Οι Huang et al. (2017) συγκέντρωσαν στοιχεία από ιστότοπο ακινήτων της Σαγκάης και πραγματοποίησαν ηδονική ανάλυση των τιμών των κατοικιών. Ο Sadayuki (2018) εξέτασε τον μηχανισμό βάσει του οποίου η ομαδοποίηση των δημόσιων συγκοινωνιών επηρέασε τις τιμές ενοικίασης με τη χρήση των δεδομένων που συλλέχθηκαν από την θυγατρική εταιρεία Door Chintai, μια ιστοσελίδα ενοικίασης ακινήτων. Οι Yao et al. (2018) συγχώνευσαν από απόσταση εικόνες από τηλεπισκόπηση και την ηλεκτρονική αγορά ακινήτων για τη χαρτογράφηση των τιμών των αστικών κατοικιών.

1.3. Το σύστημα Airbnb

Η Airbnb έχει γίνει όλο και πιο δημοφιλής στους ταξιδιώτες για τη διαμονή τους σε όλο τον κόσμο. Συνεπώς, συλλέγονται μεγάλα σύνολα δεδομένων από τις λίστες Airbnb με πλούσια χαρακτηριστικά. Η έρευνα των Luo et al. (2019) στοχεύει στην πρόβλεψη της τιμής εισαγωγής Airbnb σε τρεις πόλεις - τη Νέα Υόρκη (NYC), το Παρίσι και το Βερολίνο με διάφορες προσεγγίσεις μηχανικής μάθησης. Με μία από τις καλύτερες προσεγγίσεις - το νευρωνικό δίκτυο, επιτεύχθηκαν τιμές τετραγωνικών επιπέδων 0.769 στην εκπαίδευση του αλγορίθμου και 0.741 στη δοκιμή στο σύνολο δεδομένων της Νέας Υόρκης, τιμές 0.762 στην εκπαίδευση και 0.716 στη δοκιμή στο σύνολο δεδομένων του Παρισιού, και τιμές 0.816 στην εκπαίδευση και 0.773 στη δοκιμή στο συνδυασμένο σύνολο δεδομένων. Αποδείχθηκε ότι το νευρωνικό δίκτυο που εκπαιδεύεται στο συνδυασμένο σύνολο δεδομένων της Νέας Υόρκης και του Παρισιού, αντί των μεμονωμένων συνόλων δεδομένων, είναι πιο γενικεύσιμο για να προβλέψει την τιμή σε μια διαφορετική πόλη - το Βερολίνο.

Σε αντίθεση με τα ξενοδοχεία, τα οποία έχουν το δικό τους σύστημα τιμολόγησης, οι τιμές των κατοικιών στο Airbnb καθορίζονται συνήθως από τους οικοδεσπότες εμπειρικά. Τίθενται προκλήσεις για τους νέους οικοδεσπότες, καθώς και για υπάρχοντες οικοδεσπότες με νέες λίστες, για να καθορίσουν τις τιμές ως λογικά υψηλές χωρίς να χάσουν τη δημοτικότητά τους. Από την πλευρά των καταναλωτών, αν και μπορούν να συγκρίνουν την τιμή με άλλες παρόμοιες λίστες, εξακολουθεί να είναι πολύτιμο για αυτούς να γνωρίζουν αν η τρέχουσα τιμή είναι άξια και αν είναι μια καλή στιγμή για να ενοικιάσουν τα δωμάτια.

Η νυχτερινή τιμή για την ενοικίαση Airbnb εξαρτάται από πολλούς παράγοντες και ο τύπος εισόδου διαιρείται σε 4 κατηγορίες, συμπεριλαμβανομένων των χαρακτηριστικών συνέχειας, κατηγοριοποίησης, κειμένου και ημερομηνίας. Έχουν εξαχθεί περισσότερα από 60 χαρακτηριστικά από το σύνολο δεδομένων. Εδώ αναφέρονται μόνο μερικά από αυτά τα οποία είναι αντιπροσωπευτικά και σημαντικά για την εργασία, όπως το μέγεθος του δωματίου {φιλοξενία, τα μπάνια, τα υπνοδωμάτια, τα κρεβάτια, ...}, επιπλέον χρεώσεις {ασφάλεια κατάθεση, αμοιβή καθαρισμού, επιπλέον άτομα, .. .}, βαθμολογία αξιολόγησης {βαθμολογίες αξιολόγησης, ακρίβεια βαθμολογήσεων, βαθμολογία αναθεώρησης καθαριότητα, ...}, τοποθεσία {γειτονιά, γεωγραφικό πλάτος, γεωγραφικό μήκος ...} εγκαταστάσεις {μεταφορά, παροχές, τύπος χώρου…} και τη σχετική κράτηση {διαθεσιμότητα, πολιτική ακύρωσης, επαλήθευση κεντρικού υπολογιστή, ...}. Η ετικέτα των καθαρών από θόρυβο δεδομένων (ground-truth) αφορά την πραγματική τιμή εισαγωγής και χρησιμοποιείται μια ποικιλία προσεγγίσεων παλινδρόμησης, συμπεριλαμβανομένης της γραμμικής παλινδρόμησης, της πλησιέστερης παλινδρόμησης των γειτόνων (k nearest neighbor), της παλινδρόμησης των «τυχαίων δασών» (random forest regression), του XGBoost, καθώς και του νευρωνικού δικτύου, για την πρόβλεψη της τιμής.

Η πρόβλεψη των τιμών Airbnb γίνεται δημοφιλής λόγω της διαθεσιμότητας μεγάλων συνόλων δεδομένων σε πολλές πόλεις. Το 2015, οι Li et al. χρησιμοποίησαν πολλαπλασιασμό συγγένειας πολλαπλών κλιμάκων (Multi-Scale Affinity Propagation) για σύσταση τιμών και απέδειξαν ότι βελτιώνει σε μεγάλο βαθμό την ακρίβεια της εύλογης πρόβλεψης των τιμών. Το 2017, οι Wang et al. εργάστηκαν σε σύνολα δεδομένων Airbnb από 33 πόλεις και προσδιόρισε τους 25 προσδιοριστές τιμών από ένα

δείγμα 180.533 ενοικιαζόμενων χρησιμοποιώντας τις αναλύσεις κανονικών ελάχιστων τετραγώνων και quantile παλινδρόμηση (quantile regression). Ένα παρόμοιο έργο των Teubner et al. (2017) εξάγει χαρακτηριστικά που σχετίζονται με τη φήμη και διερευνά τις επιπτώσεις της στην τιμολόγηση με γραμμική παλινδρόμηση. Το 2019, ο Kalehbasti et al. χρησιμοποίησαν πολλαπλές προσεγγίσεις μηχανικής μάθησης και ανάλυση συναισθημάτων για την πρόβλεψη της τιμής Airbnb στο σύνολο δεδομένων NYC και πέτυχαν την τιμή 0.6901 R^2 στο σύνολο δεδομένων δοκιμής. Πρόσφατα, ο Lewis (2019) προέβλεψε την τιμή Airbnb για τα ακίνητα στο Λονδίνο χρησιμοποιώντας μηχανική μάθηση και βαθειά μάθηση και δείχνει ότι το XGBoost παρέχει την καλύτερη ακρίβεια (R^2 = 0,7274), ανώτερη από άλλα μοντέλα μηχανικής μάθησης σε διαγωνισμούς Kaggle [6]. Παρά το γεγονός ότι πραγματοποιήθηκαν πολλαπλά έργα για την πρόβλεψη των τιμών εισαγωγής, κανένα από αυτά δεν εκτελέστηκε σε διάφορες πόλεις. Η έρευνα των Luo et al. (2019) εστιάζει στις ακόλουθες τρεις εργασίες: πρώτον, διερευνώνται διαφορετικά χαρακτηριστικά μέσω της εξαγωγής χαρακτηριστικών και της μηχανικής· δεύτερον, εκτελούνται πειραματισμοί και συγκρίνονται διαφορετικές τεχνικές μηχανικής μάθησης στην πρόβλεψη των τιμών· τέλος, εκπαιδεύεται ένα πιο γενικευμένο μοντέλο.

Οι τιμές των κατοικιών αυξάνονται κάθε χρόνο, επομένως υπάρχει ανάγκη για ένα σύστημα πρόβλεψης των τιμών των κατοικιών στο μέλλον. Η πρόβλεψη της τιμής του σπιτιού μπορεί να βοηθήσει τον υπεύθυνο του έργου να καθορίσει την τιμή πώλησης ενός σπιτιού και μπορεί να βοηθήσει τον πελάτη να κανονίσει την κατάλληλη στιγμή για να αγοράσει ένα σπίτι (Chouthai et al., 2019).

2. Εξαγωγή συμπερασμάτων από δεδομένα

Η εξαγωγή συμπερασμάτων από δεδομένα είναι η διαδικασία της ανεύρεσης προτύπων σε μεγάλα σύνολα δεδομένων και περιλαμβάνει μεθόδους σε κοινά στοιχεία της μηχανικής μάθησης, της στατιστικής και των συστημάτων βάσης δεδομένων, (*"Data Mining Curriculum"*. *ACM SIGKDD*,. 2006). Η εξαγωγή συμπερασμάτων από δεδομένα είναι ένας διεπιστημονικός υποτομέας της επιστήμης των υπολογιστών και της στατιστικής με γενικό στόχο την εξαγωγή πληροφοριών (με έξυπνες μεθόδους) από ένα σύνολο δεδομένων και τη μετατροπή των πληροφοριών σε μία κατανοητή δομή για περαιτέρω χρήση, (*"Data Mining Curriculum"*. *ACM SIGKDD*,. 2006). Η εξαγωγή συμπερασμάτων από δεδομένα είναι το βήμα της ανάλυσης της διαδικασίας "ανακάλυψης γνώσεων σε βάσεις δεδομένων" ή KDD, (Fayyad et al, 1996). Εκτός από το στάδιο της πρώτης ανάλυσης, περιλαμβάνει επίσης πτυχές διαχείρισης βάσεων δεδομένων και δεδομένων, προεπεξεργασία δεδομένων, εκτιμήσεις μοντέλων και συμπερασμάτων, μετρήσεις ενδιαφερόντων, εκτιμήσεις πολυπλοκότητας, μετα-επεξεργασία ανακαλυφθέντων δομών, οπτικοποίηση και ηλεκτρονική ενημέρωση, (*"Data Mining Curriculum"*. *ACM SIGKDD*,. 2006). Η διαφορά μεταξύ της ανάλυσης δεδομένων και της εξαγωγής συμπερασμάτων από δεδομένα είναι ότι η ανάλυση δεδομένων χρησιμοποιείται για τη δοκιμή μοντέλων και υποθέσεων σχετικά στο σύνολο δεδομένων, π.χ. ανάλυση της αποτελεσματικότητας μιας καμπάνιας μάρκετινγκ, ανεξάρτητα από το μέγεθος των δεδομένων. αντίθετα, η εξαγωγή συμπερασμάτων από δεδομένα χρησιμοποιεί μοντέλα μηχανικής μάθησης και στατιστικά μοντέλα για την αποκάλυψη κρυφών ή κρυμμένων μοτίβων σε μεγάλο όγκο δεδομένων, (Olson, 2007).

Ο όρος "εξαγωγή δεδομένων" είναι στην πραγματικότητα μια λανθασμένη ονομασία, επειδή ο στόχος είναι η εξαγωγή μοτίβων και γνώσεων από μεγάλες ποσότητες δεδομένων, και όχι η εξαγωγή (εξόρυξη) των ίδιων των δεδομένων, (Han et al, 2001). Είναι επίσης μια λέξη-κλειδί, (OKAIRP 2005 Fall Conference, Arizona State University Archived 2014), και εφαρμόζεται συχνά σε οποιαδήποτε μορφή επεξεργασίας δεδομένων ή πληροφοριών μεγάλης κλίμακας (συλλογή, εξαγωγή, αποθήκευση, ανάλυση και στατιστικές) καθώς και οποιαδήποτε εφαρμογή συστήματος υποστήριξης λήψης αποφάσεων στον υπολογιστή, συμπεριλαμβανομένης της τεχνητής νοημοσύνης (π.χ. μηχανική μάθηση) και της επιχειρηματικής ευφυΐας. Το βιβλίο *Data mining: Practical machine learning tools and techniques with Java,* (Witten et al, 2011), (που καλύπτει κυρίως το υλικό μηχανικής μάθησης) αρχικά ονομαζόταν *Practical machine learning* και ο όρος *data mining* προστέθηκε μόνο για λόγους μάρκετινγκ, (Witten et al, 2010). Συχνά οι πιο γενικοί όροι (μεγάλης κλίμακας) της ανάλυσης δεδομένων και της αναλυτικής- ή, όταν γίνεται αναφορά σε πραγματικές μεθόδους, της τεχνητής νοημοσύνης και της μηχανικής μάθησης - είναι πιο κατάλληλες.

Η πραγματική εργασία εξαγωγής συμπερασμάτων από δεδομένα είναι η ημιαυτόματη ή αυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένων για την εξαγωγή προηγουμένως άγνωστων ενδιαφερόντων προτύπων όπως ομάδων αρχείων δεδομένων (ανάλυση συμπλέγματος), ασυνήθιστων αρχείων (ανίχνευση ανωμαλιών) και εξαρτήσεων (εξόρυξη κανόνα συσχέτισης, διαδοχική εξόρυξη προτύπων). Αυτό συνήθως συνεπάγεται τη χρήση τεχνικών βάσης δεδομένων, όπως χωρικών δεικτών. Αυτά τα πρότυπα μπορούν στη συνέχεια να θεωρηθούν ως ένα είδος περίληψης των δεδομένων εισόδου και μπορούν να χρησιμοποιηθούν σε περαιτέρω ανάλυση ή, για παράδειγμα, στη μηχανική μάθηση και στην προγνωστική ανάλυση. Για παράδειγμα, το βήμα της εξαγωγής συμπερασμάτων από δεδομένα μπορεί να εντοπίσει πολλαπλές ομάδες στα δεδομένα, τα οποία στη συνέχεια μπορούν να χρησιμοποιηθούν για την επίτευξη ακριβέστερων αποτελεσμάτων πρόβλεψης από ένα σύστημα υποστήριξης αποφάσεων. Ούτε η συλλογή δεδομένων, η προετοιμασία των δεδομένων, ούτε η ερμηνεία και αναφορά αποτελεσμάτων αποτελούν μέρος του βήματος εξαγωγής συμπερασμάτων από δεδομένα, αλλά ανήκουν στη συνολική διαδικασία KDD ως πρόσθετα βήματα.

Οι σχετικοί όροι του «κοσκινίσματος» δεδομένων, της «αλιείας» δεδομένων και της «περιφοράς» δεδομένων αναφέρονται στη χρήση μεθόδων εξαγωγής συμπερασμάτων από δεδομένα για τη δειγματοληψία τμημάτων ενός μεγαλύτερου συνόλου δεδομένων ενός πληθυσμού που είναι (ή μπορεί να είναι) πολύ μικρά για αξιόπιστα στατιστικά συμπεράσματα σχετικά με την εγκυρότητα οποιονδήποτε μοτίβων που ανακαλύπτονται. Αυτές οι μέθοδοι μπορούν, ωστόσο, να χρησιμοποιηθούν για τη δημιουργία νέων υποθέσεων για δοκιμή έναντι των μεγαλύτερων πληθυσμών δεδομένων.

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα έχει συμβεί εδώ και αιώνες. Οι πρόωρες μέθοδοι ταυτοποίησης μοτίβων σε δεδομένα περιλαμβάνουν το θεώρημα του Bayes (1700s) και την ανάλυση παλινδρόμησης (1800s). Ο πολλαπλασιασμός, η πανταχού παρουσία και η αυξανόμενη ισχύς της τεχνολογίας των υπολογιστών έχουν αυξήσει δραματικά τη συλλογή δεδομένων, την αποθήκευση και την ικανότητα χειρισμού. Δεδομένου ότι τα σύνολα δεδομένων έχουν αυξηθεί σε μέγεθος και πολυπλοκότητα, η άμεση ανάλυση «hands-on» δεδομένων αυξάνεται ολοένα και περισσότερο με την έμμεση αυτοματοποιημένη επεξεργασία δεδομένων, με τη βοήθεια άλλων ανακαλύψεων στην επιστήμη των υπολογιστών, όπως τα νευρωνικά δίκτυα, την ανάλυση cluster, τους γενετικούς αλγόριθμους (1950s) , τα δέντρα αποφάσεων και τους κανόνες αποφάσεων (δεκαετία του 1960) και τα μηχανήματα υποστήριξης φορέα (δεκαετία του 1990). Η εξαγωγή συμπερασμάτων από δεδομένα είναι η διαδικασία της εφαρμογής αυτών των μεθόδων με σκοπό την αποκάλυψη κρυφών μοτίβων, (Kantardzic, 2003), σε μεγάλα σύνολα δεδομένων. Γεφυρώνει το χάσμα από τις εφαρμοζόμενες στατιστικές μεθόδους και την τεχνητή νοημοσύνη (που συνήθως παρέχουν το μαθηματικό υπόβαθρο) στη διαχείριση βάσεων δεδομένων εκμεταλλευόμενη τον τρόπο της αποθήκευσης και του ευρετηρίου δεδομένων σε βάσεις δεδομένων για την αποτελεσματικότερη εκτέλεση των πραγματικών αλγορίθμων μάθησης και ανακάλυψης, επιτρέποντας την εφαρμογή τέτοιων μεθόδων σε όλο και μεγαλύτερα σύνολα δεδομένων.

2.1. Διαδικασία

Η διαδικασία της ανακάλυψης γνώσης σε βάσεις δεδομένων (KDD) καθορίζεται συνήθως με τα στάδια:

1. Επιλογή

2. Προεπεξεργασία

3. Μετασχηματισμός

4. Εξαγωγή δεδομένων

5. Ερμηνεία / αξιολόγηση, (Fayyad et al, 1996)

Αυτή η διαδικασία υπάρχει, ωστόσο, σε πολλές παραλλαγές σε αυτό το θέμα, όπως τη διεπαγγελματική τυποποιημένη διαδικασία εξαγωγής συμπερασμάτων από δεδομένα (CRISP-DM) η οποία ορίζει έξι φάσεις:

1. Επιχειρηματική κατανόηση

2. Η κατανόηση των δεδομένων

3. Προετοιμασία δεδομένων

4. Μοντελοποίηση

5. Αξιολόγηση

6. Ανάπτυξη

ή σε μια απλοποιημένη διαδικασία όπως (1) Προεπεξεργασία, (2) Εξαγωγή Δεδομένων, και (3) Επικύρωση Αποτελεσμάτων.

Οι δημοσκοπήσεις που διεξήχθησαν το 2002, το 2004, το 2007 και το 2014 δείχνουν ότι η μεθοδολογία CRISP-DM είναι η βασική μεθοδολογία που χρησιμοποιείται από τους ασχολούμενους με εξαγωγή συμπερασμάτων από δεδομένα, (Shapiro, 2002). Το μόνο άλλο πρότυπο εξαγωγής συμπερασμάτων από δεδομένα που κατονομάζεται σε αυτές τις δημοσκοπήσεις ήταν το SEMMA. Ωστόσο, 3-4 φορές περισσότερα άτομα ανέφεραν ότι χρησιμοποιούν το CRISP-DM. Πολλές ομάδες ερευνητών έχουν δημοσιεύσει ανασκοπήσεις των μοντέλων διεργασιών εξαγωγής συμπερασμάτων από δεδομένα, (Marban et al, 2009), και οι Azevedo και Santos πραγματοποίησαν μία σύγκριση των μεθόδων CRISP-DM και SEMMA το 2008, (Azevedo & Santos, 2008).

2.2. Προεπεξεργασία

Πριν από τη χρήση αλγορίθμων εξαγωγής συμπερασμάτων από δεδομένα, πρέπει να δημιουργηθεί ένα σύνολο δεδομένων που θα αποτελεί το στόχο. Καθώς η εξαγωγή δεδομένων μπορεί μόνο να αποκαλύψει τα πρότυπα που υπάρχουν πραγματικά στα δεδομένα, το στοχευόμενο σύνολο δεδομένων πρέπει να είναι αρκετά μεγάλο ώστε να περιέχει αυτά τα μοτίβα, παραμένοντας όμως αρκετά μικρό ώστε να εξάγεται εντός ενός αποδεκτού χρονικού ορίου. Μία κοινή πηγή δεδομένων είναι μια βάση δεδομένων ή μια αποθήκη δεδομένων. Η προεπεξεργασία είναι απαραίτητη για την ανάλυση των πολλών μεταβλητών συνόλων δεδομένων πριν από την εξαγωγή των δεδομένων. Στη συνέχεια, το σύνολο - στόχος «καθαρίζεται». Ο «καθαρισμός» των δεδομένων αφαιρεί τις παρατηρήσεις που περιέχουν θόρυβο και εκείνες με ελλείποντα δεδομένα.

2.3. Εξαγωγή συμπερασμάτων

Η εξαγωγή συμπερασμάτων από δεδομένα περιλαμβάνει έξι κοινές κατηγορίες εργασιών: (Fayyad et al, 1996).

- Ανίχνευση ανωμαλιών (ανίχνευση απόκλισης / αλλαγής / απόκλισης) – Ο Εντοπισμός των ασυνήθιστων αρχείων δεδομένων, που μπορεί να είναι ενδιαφέροντα ή των λαθών στα δεδομένα που απαιτούν περαιτέρω διερεύνηση.

- Κανόνας μάθησης και συσχέτισης (μοντελοποίηση εξάρτησης) - Έρευνες για σχέσεις μεταξύ μεταβλητών. Για παράδειγμα, ένα σούπερ μάρκετ ενδέχεται να συγκεντρώνει δεδομένα σχετικά με τις συνήθειες αγοράς των πελατών. Χρησιμοποιώντας τους κανόνες μάθησης και συσχέτισης, το σούπερ μάρκετ μπορεί να καθορίσει ποια προϊόντα αγοράζονται συχνά μαζί και να χρησιμοποιήσει αυτές τις πληροφορίες για σκοπούς μάρκετινγκ. Αυτό μερικές φορές αναφέρεται ως ανάλυση καλαθιού αγοράς.

- Ομαδοποίηση - είναι η εργασία της ανακάλυψης ομάδων και δομών στα δεδομένα που είναι κατά κάποιον τρόπο "παρόμοια", χωρίς τη χρήση γνωστών δομών στα δεδομένα.

- Ταξινόμηση - είναι το καθήκον της γενίκευσης μίας γνωστής δομής που θα εφαρμόζεται σε νέα δεδομένα. Για παράδειγμα, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να επιχειρήσει να ταξινομήσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως «νόμιμο» ή ως «ενοχλητικό».

- Ανάλυση - προσπαθεί να βρει μια συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο σφάλμα που είναι δυνατό, για την εκτίμηση των σχέσεων μεταξύ δεδομένων ή συνόλων δεδομένων.

- Σύνοψη - παρέχει μια πιο συμπαγή αναπαράσταση του συνόλου δεδομένων, συμπεριλαμβανομένης της απεικόνισης και της δημιουργίας αναφορών.

2.4. Επικύρωση των αποτελεσμάτων

Η εξαγωγή συμπερασμάτων από δεδομένα μπορεί να χρησιμοποιηθεί με λάθος τρόπο και μπορεί συνεπώς να παράγει αποτελέσματα που φαίνεται να είναι σημαντικά. αλλά δεν προβλέπουν πραγματικά τη μελλοντική συμπεριφορά και δεν μπορούν να αναπαραχθούν σε ένα νέο δείγμα δεδομένων και επομένως δεν έχουν μεγάλη χρησιμότητα. Συχνά αυτό προκύπτει από τη διερεύνηση πάρα πολλών υποθέσεων και τη μη διεξαγωγή κατάλληλων δοκιμασιών ανάλυσης στατιστικών υποθέσεων. Μια απλή εκδοχή αυτού του προβλήματος στη μηχανική μάθηση είναι γνωστή ως υπέρ-προσαρμογή , αλλά το ίδιο πρόβλημα μπορεί να προκύψει σε διαφορετικές φάσεις της διαδικασίας και, κατά συνέπεια, η διάσπαση της δοκιμής - όταν εφαρμόζεται - μπορεί να μην επαρκεί για να αποφευχθεί αυτό, (Hawkins, 2004).

Το τελευταίο βήμα της ανακάλυψης γνώσης από τα δεδομένα είναι να επαληθεύσουμε ότι τα πρότυπα που παράγονται από τους αλγόριθμους εξαγωγής συμπερασμάτων από δεδομένα εμφανίζονται στο ευρύτερο σύνολο δεδομένων. Δεν είναι απαραίτητα έγκυρα όλα τα πρότυπα που εντοπίζονται από τους αλγόριθμους εξαγωγής συμπερασμάτων. Είναι σύνηθες για τους αλγορίθμους εξαγωγής συμπερασμάτων από δεδομένα να βρίσκουν μοτίβα στο σύνολο που εξετάζεται που δεν υπάρχουν στο γενικό σύνολο δεδομένων. Αυτό ονομάζεται υπέρ-προσαρμογή. Για να ξεπεραστεί αυτό, η αξιολόγηση χρησιμοποιεί μια σειρά δοκιμών δεδομένων τα οποία δεν εφαρμόζει ο αλγόριθμος εξαγωγής συμπερασμάτων. Τα μαθησιακά μοτίβα εφαρμόζονται σε αυτό το σετ δοκιμών και η προκύπτουσα έξοδος συγκρίνεται με την επιθυμητή έξοδο. Για παράδειγμα, ένας αλγόριθμος εξαγωγής συμπερασμάτων από δεδομένα που προσπαθεί να διακρίνει τα «ενοχλητικά» από τα "αποδεκτά" μηνύματα ηλεκτρονικού ταχυδρομείου θα δοκιμασθεί σε ένα εκπαιδευτικό σετ δειγμάτων e-mail. Μόλις δοκιμασθεί, τα μαθησιακά μοτίβα θα εφαρμοστούν σε δοκιμαστική σειρά μηνυμάτων ηλεκτρονικού ταχυδρομείου στα οποία δεν είχε δοκιμασθεί ο αλγόριθμος. Η ακρίβεια των μοτίβων μπορεί στη συνέχεια να μετρηθεί από τον αριθμό των e-mail που ταξινομούνται σωστά. Μπορούν να χρησιμοποιηθούν διάφορες στατιστικές μέθοδοι για την αξιολόγηση του αλγορίθμου, όπως οι καμπύλες ROC.

Αν τα μαθησιακά μοτίβα δεν πληρούν τα επιθυμητά πρότυπα, είναι απαραίτητο να επανεκτιμηθούν και να αλλάξουν τα βήματα προ-επεξεργασίας και εξαγωγής συμπερασμάτων. Εάν τα μαθησιακά μοτίβα πληρούν τα επιθυμητά πρότυπα, τότε το τελευταίο βήμα είναι να ερμηνεύσουμε τα μαθησιακά πρότυπα και να τα μετατρέψουμε σε γνώση.

3. Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας υποτομέας της επιστήμης των υπολογιστών που ασχολείται με την κατασκευή αλγορίθμων οι οποίοι, για να είναι χρήσιμοι, βασίζονται σε μια συλλογή παραδειγμάτων κάποιου φαινομένου. Αυτά τα παραδείγματα μπορούν να προέρχονται από τη φύση, να κατασκευάζονται χειροποίητα από ανθρώπους ή να παράγονται από έναν άλλο αλγόριθμο.

Η μηχανική μάθηση μπορεί επίσης να οριστεί ως η διαδικασία επίλυσης ενός πρακτικού προβλήματος με 1) τη συλλογή ενός συνόλου δεδομένων και 2) την αλγοριθμική δημιουργία ενός στατιστικού μοντέλου βασισμένου σε αυτό το σύνολο δεδομένων.

Αυτό το στατιστικό μοντέλο θεωρείται ότι χρησιμοποιείται με κάποιο τρόπο για την επίλυση του πρακτικού προβλήματος (Michie et al., 1994).

3.1. Είδη Μηχανικής Μάθησης

3.1.1 Μηχανική Μάθηση με Επίβλεψη

Στην εποπτευόμενη μάθηση, το σύνολο δεδομένων είναι η συλλογή επισημασμένων παραδειγμάτων $\{(x_i,y_i)\}_i^N=1$. Κάθε στοιχείο x_i μεταξύ των N ονομάζεται διάνυσμα χαρακτηριστικών. Ένα διάνυσμα χαρακτηριστικών είναι ένα διάνυσμα στο οποίο κάθε διάσταση j = 1,. . . , D περιέχει μια τιμή που περιγράφει με κάποιο τρόπο το παράδειγμα. Αυτή η τιμή ονομάζεται χαρακτηριστικό και δηλώνεται ως $x^{(j)}$. Για παράδειγμα, εάν το κάθε παράδειγμα x στη συλλογή μας αντιπροσωπεύει ένα άτομο, τότε το πρώτο χαρακτηριστικό, $x^{(1)}$, θα μπορούσε να περιέχει το ύψος σε cm, το δεύτερο χαρακτηριστικό, $x^{(2)}$, θα μπορούσε να περιέχει το βάρος σε kg, το $x^{(3)}$ περιέχει το φύλο και ούτω καθεξής. Για όλα τα παραδείγματα στο σύνολο δεδομένων, το χαρακτηριστικό στη θέση j στο διάνυσμα χαρακτηριστικών πάντα περιέχει το ίδιο είδος πληροφορίας. Αυτό σημαίνει ότι αν $x_i^{(2)}$ περιέχει βάρος σε kg σε κάποιο παράδειγμα x_i, τότε το $x_k^{(2)}$ θα περιέχει επίσης το βάρος σε kg σε κάθε παράδειγμα x_k, k = 1,. . . , N. Η ετικέτα y_i μπορεί να είναι είτε ένα στοιχείο που ανήκει σε ένα πεπερασμένο σύνολο κατηγοριών {1, 2,. . . , C}, ή ένας πραγματικός αριθμός ή μια πιο περίπλοκη δομή, όπως ένα διάνυσμα, ένας πίνακας, ένα δέντρο ή ένα γράφημα. Εκτός αν αναφέρεται διαφορετικά, τοy_i είναι είτε ένα πεπερασμένο σύνολο τάξεων είτε ένας πραγματικός αριθμός (Kotsiantis et al., 2007). Μια τάξη είναι η κατηγορία στην οποία ανήκει ένα παράδειγμα. Αν τα παραδείγματα είναι μηνύματα ηλεκτρονικού ταχυδρομείου και το πρόβλημά είναι η ανίχνευση ανεπιθύμητων μηνυμάτων, τότε θέτονται δύο κατηγορίες {spam, not_spam}.

Ο στόχος ενός αλγόριθμου εποπτευόμενης μάθησης είναι η χρήση του συνόλου δεδομένων για την παραγωγή ενός μοντέλου που παίρνει ένα διάνυσμα χαρακτηριστικών x ως είσοδο και εξάγει πληροφορία που επιτρέπει την εξαγωγή μιας ετικέτας για αυτό το διάνυσμα χαρακτηριστικών. Για παράδειγμα, το μοντέλο που δημιουργήθηκε χρησιμοποιώντας το σύνολο δεδομένων των ανθρώπων θα μπορούσε να λάβει ως είσοδο ένα διάνυσμα χαρακτηριστικών που περιγράφει ένα άτομο και να εξάγει μια πιθανότητα το άτομο να έχει καρκίνο.

3.1.2. Μη Εποπτευόμενη Μηχανική Μάθηση

Στην μη εποπτευόμενη μάθηση, το σύνολο δεδομένων είναι μια συλλογή μη χαρακτηρισμένων παραδειγμάτων $\{(x_i,y_i)\}_i^N=1$. Και πάλι, το x είναι ένα διάνυσμα χαρακτηριστικών και ο στόχος ενός αλγόριθμου μάθησης χωρίς επίβλεψη είναι να δημιουργήσει ένα μοντέλο που παίρνει ένα διάνυσμα χαρακτηριστικών x ως είσοδο και είτε το μετατρέπει σε ένα άλλο διάνυσμα ή σε μια τιμή που μπορεί να χρησιμοποιηθεί για την επίλυση ενός πρακτικού προβλήματος. Για παράδειγμα, κατά την ομαδοποίηση, το μοντέλο επιστρέφει την ταυτότητα της ομάδας για κάθε διάνυσμα χαρακτηριστικών στο σύνολο δεδομένων (Kassambara, 2017).

Στη μείωση των διαστάσεων, η έξοδος του μοντέλου είναι ένα διάνυσμα χαρακτηριστικών που έχει λιγότερα χαρακτηριστικά από την είσοδο x, στην ανίχνευση των εξόδων (outlier detection), η έξοδος είναι ένας πραγματικός αριθμός που υποδεικνύει πώς το x είναι διαφορετικό από ένα "τυπικό" παράδειγμα στο σύνολο δεδομένων.

3.1.3. Ημι-Εποπτευόμενη Μηχανική Μάθηση

Στην ημι-εποπτευόμενη μάθηση, το σύνολο δεδομένων περιέχει τόσο επισημασμένα όσο και μη επισημασμένα παραδείγματα. Συνήθως, η ποσότητα των μη επισημασμένων παραδειγμάτων είναι πολύ μεγαλύτερη από τον αριθμό των επισημασμένων παραδειγμάτων. Ο στόχος ενός ημι-εποπτευόμενου αλγορίθμου μάθησης είναι ο ίδιος με τον στόχο του αλγόριθμου εποπτευόμενης μάθησης. Η ελπίδα εδώ είναι ότι η χρήση πολλών μη επισημασμένων παραδειγμάτων μπορεί να βοηθήσει τον αλγόριθμο μάθησης να βρει ("παράγει" ή "υπολογίσει") ένα καλύτερο μοντέλο (Zhu, 2005).

3.1.4. Ενισχυμένη Μάθηση

Η ενισχυμένη μάθηση είναι ένα υποπεδίο της μηχανικής μάθησης όπου η μηχανή "ζει" σε ένα περιβάλλον και είναι ικανή να αντιληφθεί την κατάσταση αυτού του περιβάλλοντος ως ένα διάνυσμα χαρακτηριστικών. Το μηχάνημα μπορεί να εκτελεί ενέργειες σε κάθε κατάσταση. Οι διαφορετικές ενέργειες φέρνουν διαφορετικές ανταμοιβές και μπορούν επίσης να μετακινήσουν το μηχάνημα σε άλλη κατάσταση του περιβάλλοντος. Ο στόχος ενός αλγόριθμου ενίσχυσης μάθησης είναι να μάθει μια πολιτική. Μια πολιτική είναι μια συνάρτηση f (παρόμοια με το μοντέλο στην εποπτευόμενη μάθηση) που παίρνει το διάνυσμα χαρακτηριστικών μιας κατάστασης ως είσοδο και εξάγει μια βέλτιστη ενέργεια για να εκτελεστεί σε αυτή την κατάσταση. Η δράση είναι βέλτιστη εάν μεγιστοποιεί την αναμενόμενη μέση ανταμοιβή (Lison, 2015).

Η ενισχυμένη μάθησης επιλύει ένα συγκεκριμένο είδος προβλημάτων όπου η λήψη αποφάσεων είναι διαδοχική και ο στόχος είναι μακροπρόθεσμος, όπως το παιχνίδι, η ρομποτική, η διαχείριση των πόρων ή η εφοδιαστική.

4. Αλγόριθμοι Κατηγοριοποίησης

4.1. k-Πλησιέστεροι γείτονες

Ο k-Πλησιέστεροι γείτονες (kNN) είναι ένας μη παραμετρικός αλγόριθμος μάθησης. Σε αντίθεση με άλλους αλγόριθμους μάθησης που επιτρέπουν την απόρριψη των δεδομένων εκπαίδευσης μετά την κατασκευή του μοντέλου, ο kNN κρατά όλα τα παραδείγματα εκπαίδευσης στη μνήμη. Μόλις εισέλθει ένα νέο, προηγουμένως αόρατο παράδειγμα x, ο kNN αλγόριθμος βρίσκει τα k παραδείγματα εκπαίδευσης πλησιέστερα στο x και επιστρέφει την επισήμανση της πλειοψηφίας σε περίπτωση ταξινόμησης ή τη μέση επισήμανση σε περίπτωση παλινδρόμησης.

Η εγγύτητα δύο παραδειγμάτων δίνεται από μια συνάρτηση απόστασης. Για παράδειγμα, η ευκλείδεια απόσταση χρησιμοποιείται συχνά στην πράξη. Μια άλλη δημοφιλής επιλογή της συνάρτησης απόστασης είναι η ομοιότητα αρνητικού συνημίτονου. Η ομοιότητα του συνημίτονου ορίζεται ως,

$$s(x_i, x_k) \stackrel{\text{def}}{=} \cos\left(\angle(x_i, x_k)\right) = \frac{\sum_{j=1}^{D} x_i^{(j)} x_k^{(j)}}{\sqrt{\sum_{j=1}^{D}\left(x_i^{(j)}\right)^2}\sqrt{\sum_{j=1}^{D}\left(x_k^{(j)}\right)^2}}$$

Αυτή είναι ένα μέτρο της ομοιότητας των κατευθύνσεων δύο διανυσμάτων. Εάν η γωνία μεταξύ δύο διανυσμάτων είναι 0 μοίρες, τότε δύο διανύσματα δείχνουν προς την ίδια κατεύθυνση και η ομοιότητα συνημίτονου είναι ίση με 1. Εάν τα διανύσματα είναι ορθογωνικά, η ομοιότητα συνημιτόνου είναι 0. Για διανύσματα που δείχνουν αντίθετες κατευθύνσεις είναι -1. Αν πρέπει να χρησιμοποιηθεί η ομοιότητα συνημιτόνου σαν μέτρηση απόστασης, πρέπει να πολλαπλασιαστεί με το -1.

Άλλες δημοφιλείς μετρήσεις απόστασης περιλαμβάνουν την απόσταση Chebychev, την απόσταση Mahalanobis και την απόσταση Hamming. Η επιλογή της μέτρησης απόστασης, καθώς και η τιμή για το k, είναι οι επιλογές που κάνει ο αναλυτής πριν από την εκτέλεση του αλγορίθμου. Έτσι αυτές είναι υπερπαράμετροι (hyperparameters). Η μέτρηση απόστασης θα μπορούσε επίσης να αντληθεί από τα δεδομένα.

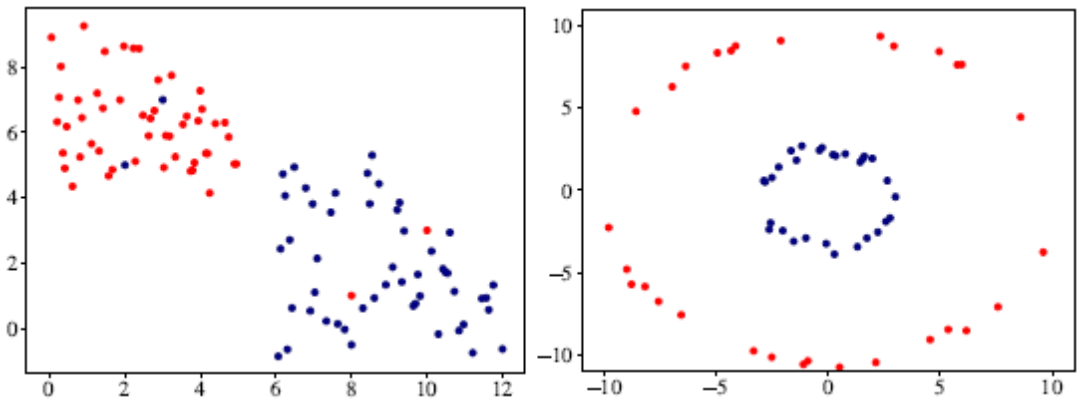
4.2. Μηχανή διανυσμάτων υποστήριξης

Πρέπει να δοθούν δύο ερωτήματα:

1. Τι γίνεται αν υπάρχει θόρυβος στα δεδομένα και κανένα υπερεπίπεδο δεν μπορεί να διαχωρίσει απόλυτα τα θετικά παραδείγματα από τα αρνητικά;

2. Τι γίνεται αν τα δεδομένα δεν μπορούν να διαχωριστούν χρησιμοποιώντας ένα επίπεδο, αλλά θα μπορούσαν να διαχωριστούν από ένα ανώτερο πολυώνυμο;

Παράδειγμα των δύο καταστάσεων απεικονίζεται στο παρακάτω σχήμα. Στην αριστερή περίπτωση, τα δεδομένα θα μπορούσαν να χωριστούν με ευθεία γραμμή, αν δεν υπήρχε θόρυβος (παραμορφώσεις ή παραδείγματα με λάθος ετικέτες). Στη δεξιά περίπτωση, το όριο απόφασης είναι ένας κύκλος και όχι μια ευθεία γραμμή.



Στο SVM, πρέπει να ικανοποιηθούν οι ακόλουθοι περιορισμοί:

$$wx_i - b \geq +1 \; \alpha\nu \; y_i = +1$$
$$wx_i - b \leq -1 \; \alpha\nu \; y_i = -1$$

(8)

Πρέπει επίσης να ελαχιστοποιηθεί το $\|w\|$ έτσι ώστε το υπερεπίπεδο (hyperplane) να είναι εξίσου απομακρυσμένο από τα πλησιέστερα παραδείγματα κάθε κλάσης. Η ελαχιστοποίηση του $\|w\|$ είναι ισοδύναμη με την ελαχιστοποίηση του 1/2 || w || ^{2}, και η χρήση αυτού του όρου καθιστά δυνατή την εκτέλεση προγραμματισμού τετραγωνικής βελτιστοποίησης. Επομένως, το πρόβλημα βελτιστοποίησης για το SVM μοιάζει με αυτό:

$$min\frac{1}{2}\|w\|^2, \tau\acute{\epsilon}\tau o\iota o \; \acute{\omega}\sigma\tau\epsilon \; y_i(x_i w - b) - 1 \geq 0, i = 1, \dots, N$$

(9)

Αντιμετώπιση του θορύβου

Για την επέκταση του SVM σε περιπτώσεις όπου τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, εισάγεται η συνάρτηση απώλειας hinge: max (0, 1 - y_i (wx_i - b)). Η συνάρτηση απώλειας hinge είναι μηδέν εάν πληρούνται οι περιορισμοί στην εξίσωση8. με άλλα λόγια, εάν το wx_i βρίσκεται στη σωστή πλευρά του ορίου λήψης αποφάσεων. Για τα δεδομένα που βρίσκονται στη λανθασμένη πλευρά του ορίου απόφασης, η τιμή της λειτουργίας είναι ανάλογη με την απόσταση από το όριο απόφασης.

Στη συνέχεια, ελαχιστοποιείται η ακόλουθη συνάρτηση κόστους,

$$C\|w\|^2 + \frac{1}{N}\sum_{i=1}^{N} \max\,(0,\ \ 1 - y_i(wx_i - b))$$

όπου η υπερπαράμετρος (hyperparameter) C καθορίζει τη συναλλαγή μεταξύ της αύξησης του μεγέθους του ορίου απόφασης και της διασφάλισης ότι κάθε x_i βρίσκεται στην σωστή πλευρά του ορίου απόφασης.

Η τιμή του C επιλέγεται συνήθως πειραματικά, ακριβώς όπως τα υπερπαραμετρικά στοιχεία του ID3 ε και d. Τα SVMs που βελτιστοποιούν την συνάρτηση απώλειας hinge καλούνται SVM μαλακού περιθωρίου, ενώ η αρχική σύνταξη αναφέρεται ως SVM σκληρού περιθωρίου.

4.3. Αλγόριθμοι Κατηγοριοποίησης βασισμένοι σε Δένδρα

4.3.1. Δέντρα αποφάσεων

Ένα δέντρο απόφασης είναι ένα ακυκλικό γράφημα που μπορεί να χρησιμοποιηθεί για τη λήψη αποφάσεων. Σε κάθε κόμβο διακλάδωσης του γραφήματος, εξετάζεται ένα συγκεκριμένο χαρακτηριστικό j του διανύσματος χαρακτηριστικών. Αν η τιμή του χαρακτηριστικού είναι κάτω από ένα συγκεκριμένο όριο, τότε ακολουθείται ο αριστερός κλάδος. Διαφορετικά, ακολουθείται ο δεξιός κλάδος. Καθώς φθάνεται ο τελικός κόμβος (leaf node), αποφασίζεται η τάξη στην οποία ανήκει το παράδειγμα. Ένα δέντρο απόφασης μπορεί να αντληθεί από τα δεδομένα (Quinlan, 1986).

Όπως και στο παρελθόν, υπάρχει μια συλλογή από επισημασμένα παραδείγματα. οι ετικέτες ανήκουν στο σετ $\{0, 1\}$. Πρέπει να φτιαχτεί ένα δέντρο αποφάσεων που θα επιτρέπει να προβλεφθεί η κλάση ενός παραδείγματος που δίνεται σε ένα διάνυσμα χαρακτηριστικών. Υπάρχουν διάφορες διαμορφώσεις του αλγορίθμου μάθησης δέντρων αποφάσεων. Θεωρείται μόνο ένας, που ονομάζεται ID3. Το κριτήριο βελτιστοποίησης, στην προκειμένη περίπτωση, είναι η μέση λογαριθμική πιθανότητα:

$$\frac{1}{N}\sum_{i=1}^{N} y_i ln f_{ID3}(\mathbf{x}_i) + (1 - y_i)\ln\left(1 - f_{ID3}(\mathbf{x}_i)\right)$$

όπου το f_{ID3} είναι ένα δέντρο απόφασης.

Μέχρι τώρα, μοιάζει πολύ με τον αλγόριθμο logistic regression. Ωστόσο, σε αντίθεση με τον αλγόριθμο μάθησης logistic regression που κατασκευάζει ένα παραμετρικό μοντέλο $f_{w^*,\ b^*}$ με την εξεύρεση βέλτιστης λύσης στο κριτήριο βελτιστοποίησης, ο αλγόριθμος ID3 το βελτιστοποιεί προσεγγιστικά κατασκευάζοντας ένα μη παραμετρικό μοντέλο

$$f_{ID3}(\mathbf{x}) \stackrel{\text{def}}{=} \Pr(y = 1|\mathbf{x})$$

Ο αλγόριθμος εκμάθησης ID3 λειτουργεί ως εξής. Ας είναι S ένα σύνολο επισημασμένων παραδειγμάτων. Στην αρχή, το δέντρο απόφασης έχει μόνο ένα κόμβο εκκίνησης που περιέχει όλα τα παραδείγματα: $S \stackrel{\text{def}}{=} \{(x_i, y_i)\}_{i=1}^{N}$. Αρχίζει με ένα σταθερό μοντέλο f_{ID3}^{S} :

$$f_{ID3}^{S} = \frac{1}{|S|}\sum_{(x,y)\in S} y \qquad (1)$$

Η πρόβλεψη που δίνεται από το παραπάνω μοντέλο, $f_{ID3}^{S}(x)$ θα είναι η ίδια για κάθε είσοδο x. Το αντίστοιχο δέντρο απόφασης παρουσιάζεται στο σχήμα 4α.

Στη συνέχεια, αναζητούνται όλα τα χαρακτηριστικά j = 1,. . . , D και όλα τα όρια t, και διαιρείται το σύνολο S σε δύο υποσύνολα:

$$S_{-} \stackrel{\text{def}}{=} \{(x,y)|(x,y) \in S, x^{(j)} < t\}$$

και

$$S_{+} \stackrel{\text{def}}{=} \{(x,y)|(x,y) \in S, x^{(j)} \geq t\}$$

Τα δύο νέα υποσύνολα θα πάνε σε δύο νέους κόμβους φύλλων και αξιολογείται, για όλα τα πιθανά ζεύγη (j, t), πόσο καλή είναι η διάσπαση με τα κομμάτια S_ και S +. Τέλος, επιλέγονται οι καλύτερες τέτοιες τιμές (j, t), διαιρείται το S σε S + και S_, σχηματίζονται δύο νέοι κόμβοι φύλλων και συνεχίζεται η διαδικασία αναδρομικώς με τα S + και S_ (ή τερματίζεται εάν ένας διαχωρισμός δεν δημιουργεί ένα μοντέλο επαρκώς καλύτερο από το τρέχον). Ένα δέντρο απόφασης μετά από μία διάσπαση απεικονίζεται στο σχ. 4b. Στον αλγόριθμο ID3, η αξία ενός διαχωρισμού εκτιμάται χρησιμοποιώντας το κριτήριο που ονομάζεται εντροπία. Η εντροπία είναι ένα μέτρο αβεβαιότητας για μια τυχαία μεταβλητή. Θα φτάσει στο μέγιστο όταν όλες οι τιμές των τυχαίων μεταβλητών είναι ισοπίθανες. Η εντροπία φθάνει στο ελάχιστο της όταν η τυχαία μεταβλητή μπορεί να έχει μόνο μία τιμή. Η εντροπία ενός συνόλου παραδειγμάτων S δίνεται από:

$$H(S) = -f_{ID3}^{S} ln f_{ID3}^{S} - (1 - f_{ID3}^{S}) \ln (1 - f_{ID3}^{S})$$

Όταν διαιρείται ένα σύνολο παραδειγμάτων με ένα συγκεκριμένο χαρακτηριστικό j και ένα όριο t, η εντροπία ενός διαχωρισμού, H (S_, S +), είναι απλώς ένα σταθμισμένο άθροισμα δύο εντροπιών:

$$H(S_-, S_+) = \frac{|S_-|}{|S|} H(S_-) + \frac{|S_+|}{|S|} H(S_+)$$

Έτσι, στον αλγόριθμο ID3, σε κάθε βήμα, σε κάθε κόμβο φύλλων, βρίσκεται ένας διαχωρισμός που ελαχιστοποιεί την εντροπία που δίνεται από την ανωτέρω εξίσωση ή γίνεται τερματισμός σε αυτόν τον κόμβο φύλλων.

Ο αλγόριθμος σταματά σε έναν κόμβο φύλλων σε οποιαδήποτε από τις παρακάτω περιπτώσεις:

• Όλα τα παραδείγματα στον κόμβο των φύλλων ταξινομούνται σωστά από το μοντέλο ενός τμήματος (εξίσωση 1).

• Δεν μπορεί να βρεθεί ένα χαρακτηριστικό για περεταίρω διαχωρισμό.

• Η διάσπαση μειώνει την εντροπία λιγότερο από κάποια τιμή "(η τιμή αυτή πρέπει να βρεθεί πειραματικά).

• Το δέντρο φτάνει σε κάποιο μέγιστο βάθος d (πρέπει επίσης να βρεθεί πειραματικά).

Επειδή στο ID3, η απόφαση να χωριστεί το σύνολο δεδομένων σε κάθε επανάληψη είναι τοπική (δεν εξαρτάται από μελλοντικούς διαχωρισμούς), ο αλγόριθμος δεν εγγυάται τη βέλτιστη λύση. Το μοντέλο μπορεί να βελτιωθεί με τη χρήση τεχνικών όπως η αναζήτηση προς τα πίσω – (back tracking) κατά την αναζήτηση του βέλτιστου δέντρου αποφάσεων, με το κόστος να χρειαστεί περισσότερο χρόνο για την κατασκευή ενός μοντέλου.

4.3.2. Τυχαία Δάση (Random Forests)

Υπάρχουν δύο παραδείγματα εκμάθησης του συνόλου: bagging και boosting. Το bagging αποτελείται από τη δημιουργία πολλών "αντιγράφων" των δεδομένων εκπαίδευσης (κάθε αντίγραφο είναι ελαφρώς διαφορετικό από το άλλο) και στη συνέχεια εφαρμόζεται «weak learner» συνάρτηση σε κάθε αντίγραφο για να αποκτηθούν πολλαπλά «weak» μοντέλα όπου στη συνέχεια συνδυάζονται. Το παράδειγμα bagging βρίσκεται πίσω από τον αλγόριθμο μάθησης random forest (Breiman, 2001).

Ο bagging αλγόριθμος "vanilla" λειτουργεί ως εξής. Με βάση ένα σετ εκπαίδευσης, δημιουργούνταιΒτυχαία δείγματα S_b (για κάθε b = 1, ..., B) του σετ εκπαίδευσης και χτίζεται ένα μοντέλο δέντρων αποφάσεων f_b χρησιμοποιώντας κάθε δείγμα S_b ως το σετ εκπαίδευσης. Για να παρθούν τα S_b για μερικά b, γίνεται δειγματοληψία με αντικατάσταση. Αυτό σημαίνει ότι γίνεται η εκκίνηση με ένα άδειο σετ και στη συνέχεια επιλέγεται τυχαία ένα παράδειγμα από το σετ εκπαίδευσης και τοποθετείται το ακριβές αντίγραφό του στο S_b διατηρώντας το αρχικό παράδειγμα στο αρχικό σετ εκπαίδευσης. Λαμβάνονται τυχαία τα παραδείγματα μέχρι το | Sb | = N.

Μετά την εκπαίδευση, έχουμε Β δέντρα αποφάσεων. Η πρόβλεψη για ένα νέο παράδειγμα x λαμβάνεται ως ο μέσος όρος των Β προβλέψεων:

$$y \leftarrow \hat{f}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{B}\sum_{b=1}^{B} f_b(\mathbf{x})$$

σε περίπτωση παλινδρόμησης, ή με τη λήψη της πλειοψηφίας στην περίπτωση ταξινόμησης.

Ο αλγόριθμος random forest είναι διαφορετικός από τον αλγόριθμο vanilla bagging με έναν μόνο τρόπο. Χρησιμοποιεί έναν τροποποιημένο αλγόριθμο εκμάθησης δέντρων που επιθεωρεί, σε κάθε διαχωρισμό στη διαδικασία εκμάθησης, ένα τυχαίο υποσύνολο των χαρακτηριστικών. Ο λόγος για να γίνει αυτό είναι να αποφευχθεί η συσχέτιση των δένδρων: εάν ένα ή λίγα χαρακτηριστικά είναι πολύ ισχυροί παράγοντες πρόβλεψης για τον στόχο, αυτά τα χαρακτηριστικά θα επιλεγούν για να χωρίσουν τα παραδείγματα σε πολλά δέντρα. Αυτό θα είχε ως αποτέλεσμα πολλά συσχετισμένα δέντρα στο «δάσος». Οι συσχετισμένοι προγνωστικοί παράγοντες δεν μπορούν να βοηθήσουν στη βελτίωση της ακρίβειας της πρόβλεψης. Ο κύριος λόγος για την καλύτερη απόδοση του μοντέλου είναι ότι τα μοντέλα που είναι καλά θα συμφωνήσουν πιθανώς στην ίδια πρόβλεψη, ενώ τα κακά μοντέλα πιθανότατα διαφωνούν σε διαφορετικές προβλέψεις. Η συσχέτιση θα κάνει τα κακά μοντέλα πιο πιθανό να συμφωνήσουν, γεγονός που θα παρεμποδίσει την πλειοψηφία ή τον μέσο όρο.

Οι πιο σημαντικοί υπερπαράμετροι για να ρυθμιστούν είναι ο αριθμός των δέντρων, B, και το μέγεθος του τυχαίου υποσυστήματος των χαρακτηριστικών που πρέπει να ληφθούν υπόψη σε κάθε διάσπαση.

Ο αλγόριθμος random forest είναι ένας από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους μάθησης του συνόλου. Είναι πολύ αποτελεσματικός γιατί με τη χρήση πολλαπλών δειγμάτων του αρχικού συνόλου δεδομένων, μειώνεται η διακύμανση του τελικού μοντέλου, λαμβάνοντας ότι η χαμηλή διακύμανση σημαίνει χαμηλή υπερπροσαρμογή.

Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο προσπαθεί να εξηγήσει μικρές παραλλαγές στο σύνολο δεδομένων, επειδή το σύνολο δεδομένων είναι ένα μικρό δείγμα του πληθυσμού όλων των πιθανών παραδειγμάτων του φαινομένου που γίνεται προσπάθεια να μοντελοποιηθεί. Αν ο τρόπος δειγματοληψίας του εκπαιδευτικού σετ δεν ήταν σωστός, τότε θα μπορούσε το δείγμα να περιέχει κάποια ανεπιθύμητα (αλλά αναπόφευκτα) αντικείμενα: θόρυβο, υπερβολικές τιμές και υπερβολικά ή υποεκπροσωπούμενα παραδείγματα. Δημιουργώντας πολλαπλά τυχαία δείγματα με αντικατάσταση του εκπαιδευτικού σετ, μειώνεται η επίδραση αυτών των αντικειμένων.

4.4. Νευρωνικά Δίκτυα

4.4.1. Η Έννοια του Νευρωνικού Δικτύου

Ένα νευρωνικό δίκτυο (NN), ακριβώς όπως μια παλινδρόμηση ή ένα μοντέλο SVM, είναι μια μαθηματική συνάρτηση:

$$y = f_{NN}(\mathbf{x})$$

Η συνάρτηση f_{NN} έχει μια συγκεκριμένη μορφή: είναι μια σύνθετη συνάρτηση. Έτσι, για ένα 3-στρωμάτωννευρωνικό δίκτυο που επιστρέφει ένα βαθμωτό μέγεθος, η f_{NN} έχει την μορφή:

$$y = f_{NN}(\mathbf{x}) = f_3(\boldsymbol{f_2}(\boldsymbol{f_1}(\mathbf{x})))$$

Στην παραπάνω εξίσωση, οι f_2, f_3 είναι διανυσματικές συναρτήσεις της ακόλουθης μορφής:

$$\boldsymbol{f_l}(\boldsymbol{z}) \stackrel{\text{def}}{=} \boldsymbol{g_l}(\boldsymbol{W_l z} + \boldsymbol{b_l})$$

όπου l ονομάζεται δείκτης επιπέδων και μπορεί να εκτείνεται από 1 σε οποιοδήποτε αριθμό επιπέδων. Η συνάρτηση g_l ονομάζεται συνάρτηση ενεργοποίησης. Είναι μια σταθερή, συνήθως μη γραμμική συνάρτηση που επιλέγεται από τον αναλυτή δεδομένων πριν αρχίσει η λειτουργία εκμάθησης. Οι παράμετροι W_l (πίνακας) και b_l (διάνυσμα) για κάθε επίπεδο μαθαίνονται χρησιμοποιώντας την γνωστή μέθοδο gradient descent βελτιστοποιώντας, ανάλογα με την εργασία, μια συγκεκριμένη λειτουργία κόστους (όπως την MSE). Αν συγκριθεί η ανωτέρω εξίσωση με την εξίσωση του logistic regression, μπορεί να αντικατασταθεί το g_l από τη sigmoid συνάρτηση, και δεν θα διαπιστωθεί καμία

διαφορά. Η συνάρτηση f_1 είναι μια βαθμωτή συνάρτηση για την εργασία παλινδρόμησης, αλλά μπορεί επίσης να είναι μια διανυσματική συνάρτηση ανάλογα με το πρόβλημά (Haykin, 1994).

Ο λόγος που χρησιμοποιείται ένας πίνακας W_l και όχι ένα διάνυσμα w_l είναι ότι g_l είναι μια διανυσματική συνάρτηση. Κάθε σειρά $w_{l,u}$ (u για μονάδα) του πίνακα W_1 είναι ένα διάνυσμα της ίδιας διάστασης με το z. Έστω $a_{l,u} = w_{l,u}z + b_{l,u}$. Η έξοδος του f_l (z) είναι ένα διάνυσμα $[g_l (a_l, 1), g_l (a_l, 2),. . . , g_l (a_l,$ μέγεθος $l)]$, όπου το g_l είναι κάποια βαθμωτή συνάρτηση, και μέγεθος l είναι ο αριθμός των τμημάτων στο στρώμα l. Για να γίνει πιο σαφές, θεωρείται μια αρχιτεκτονική των νευρωνικών δικτύων που ονομάζεται multilayer perceptron.

Μια πιο προσεκτική ματιά σε μια συγκεκριμένη διαμόρφωση των νευρωνικών δικτύων που ονομάζονται feed-forward νευρωνικά δίκτυα (FFNN), είναι συγκεκριμένα η αρχιτεκτονική που ονομάζεται multilayer perceptron (MLP). Για παράδειγμα, θεωρείται ένα MLP με τρία στρώματα. Το δίκτυό του παραδείγματος λαμβάνει διδιάστατο διάνυσμα χαρακτηριστικών ως είσοδο και εξάγει έναν αριθμό. Αυτό το FFNN μπορεί να είναι μια παλινδρόμηση ή ένα μοντέλο ταξινόμησης, ανάλογα με τη λειτουργία ενεργοποίησης που χρησιμοποιείται στο τρίτο επίπεδο εξόδου.

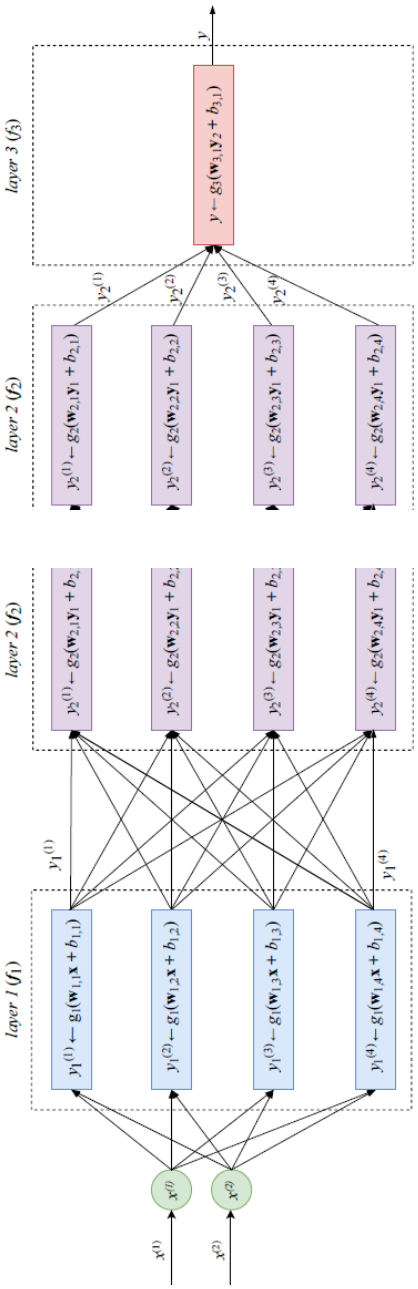
Το MLP του παραδείγματος απεικονίζεται στο σχ. 1. Το νευρωνικό δίκτυο αναπαριστάται γραφικά ως συνδεδεμένος συνδυασμός μονάδων λογικά οργανωμένων σε ένα ή περισσότερα στρώματα. Κάθε μονάδα αντιπροσωπεύεται είτε από έναν κύκλο είτε από ένα ορθογώνιο. Το εισερχόμενο βέλος αντιπροσωπεύει μια είσοδο μιας μονάδας και υποδεικνύει από πού προέρχεται αυτή η είσοδος.

Η έξοδος κάθε μονάδας είναι το αποτέλεσμα της μαθηματικής ενέργειας που είναι γραμμένη μέσα στον κύκλο ή στο ορθογώνιο. Οι μονάδες κύκλου δεν κάνουν τίποτα με την είσοδο. απλώς στέλνουν τις εισροές τους απευθείας στην έξοδο.

Τα ακόλουθα συμβαίνουν σε κάθε μονάδα ορθογωνίου. Πρώτον, όλες οι είσοδοι της μονάδας ενώνονται μεταξύ τους για να σχηματίσουν ένα διάνυσμα εισόδου. Στη συνέχεια, η μονάδα εφαρμόζει γραμμικό μετασχηματισμό στο διάνυσμα εισόδου, ακριβώς όπως το μοντέλο γραμμικής παλινδρόμησης κάνει με το διάνυσμα χαρακτηριστικών εισόδου. Τέλος, η μονάδα εφαρμόζει μια συνάρτηση ενεργοποίησης g στο αποτέλεσμα του γραμμικού μετασχηματισμού και αποκτά την τιμή εξόδου, έναν πραγματικό αριθμό. Σε ένα FFNN του vanillaμοντέλου, η τιμή εξόδου μιας μονάδας κάποιου στρώματος γίνεται μια τιμή εισόδου για κάθε μια από τις μονάδες της επόμενης στρώσης.

Στο σχ. 1, η συνάρτηση ενεργοποίησης g_l έχει έναν ευρετήριο: l, τον δείκτη του στρώματος που ανήκει η μονάδα. Συνήθως, όλες οι μονάδες ενός στρώματος χρησιμοποιούν την ίδια συνάρτηση ενεργοποίησης, αλλά αυτό δεν είναι απολύτως απαραίτητο. Κάθε στρώμα μπορεί να έχει διαφορετικό αριθμό μονάδων. Κάθε μονάδα έχει τις δικές της παραμέτρους $w_{l,\,u}$ και $b_{l,\,u}$, όπου u είναι ο δείκτης της μονάδας και l είναι ο δείκτης του στρώματος. Το διάνυσμα y_{l-1} σε κάθε μονάδα ορίζεται ως $\left[y_{l-1}^{(1)}, y_{l-1}^{(2)}, y_{l-1}^{(3)}, y_{l-1}^{(4)}\right]$. Το διάνυσμα x στην πρώτη στρώση ορίζεται ως

$\left[x^{(1)}, \ldots, x^{(D)}\right]$.



igure 1: A multilayer perceptron with two-dimensional input, two layers with fo
nit.
two layers with four units and one output layer with one

Σχήμα 1: Ένα πολυεπίπεδο perceptron με διδιάστατες εισόδους. Δύο στρώσεις με τέσσερις μονάδες και μία έξοδος με μία μονάδα.

Όπως μπορεί να φανεί στο σχ. 1, σε πολυστρωματικό perceptron όλες οι έξοδοι ενός στρώματος συνδέονται σε κάθε είσοδο του επόμενου στρώματος. Αυτή η αρχιτεκτονική καλείται πλήρως συνδεδεμένη. Ένα νευρωνικό δίκτυο μπορεί να περιέχει πλήρως συνδεδεμένα στρώματα. Αυτά είναι τα στρώματα των οποίων οι μονάδες λαμβάνουν ως είσοδο τις εξόδους εκάστης των μονάδων της προηγούμενης στρώσης.

4.4.2. Αρχιτεκτονική Νευρωνικού Δικτύου

Για την λύση ενός προβλήματος παλινδρόμησης ή ταξινόμησης, το τελευταίο (το πιο δεξιό) επίπεδο ενός νευρικού δικτύου συνήθως περιέχει μόνο μία μονάδα. Αν η λειτουργία ενεργοποίησης της τελευταίας μονάδας είναι γραμμική, τότε το νευρωνικό δίκτυο είναι ένα μοντέλο παλινδρόμησης. Εάν αυτή η συνάρτηση είναι μια λογική συνάρτηση, το νευρωνικό δίκτυο είναι ένα δυαδικό μοντέλο ταξινόμησης.

Ο αναλυτής δεδομένων είναι ελεύθερος να επιλέξει οποιαδήποτε μαθηματική συνάρτηση ως $g_{l,u}$, με την προϋπόθεση ότι είναι διαφορίσιμη. Η τελευταία ιδιότητα είναι απαραίτητη για τον αλγόριθμο gradientdescent, που χρησιμοποιείται για την εύρεση των τιμών των παραμέτρων $w_{l,u}$ και $b_{l,u}$ για όλα τα l και u. Ο πρωταρχικός σκοπός της ύπαρξης μη γραμμικών συνιστωσών στη συνάρτηση f_{NN} είναι να επιτρέψει στο νευρωνικό δίκτυο να προσεγγίσει μη γραμμικές συναρτήσεις. Χωρίς τις μη γραμμικότητες, το f_{NN} θα είναι γραμμικό, ανεξάρτητα από το πόσα στρώματα έχει. Ο λόγος είναι ότι το $W_l z + b_l$ είναι μια γραμμική συνάρτηση και μια γραμμική συνάρτηση μιας γραμμικής συνάρτησης είναι επίσης μια γραμμική συνάρτηση.

Οι δημοφιλείς επιλογές των συναρτήσεων ενεργοποίησης είναι η logisticσυνάρτηση , καθώς και η TanH και η ReLU. Η πρώτη είναι η συνάρτηση υπερβολικής εφαπτόμενης, παρόμοια με τη συνάρτηση της εφοδιαστικής αλυσίδας, αλλά κυμαίνεται από -1 έως 1 (χωρίς να τα φτάνει). Η τελευταία είναι η συνάρτηση διορθωμένης γραμμικής μονάδας, η οποία ισούται με το μηδέν όταν η είσοδος της z είναι αρνητική και ίση με z διαφορετικά:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$relu(z) = \{^{0, z<0}_{z, z\geq 0}$$

Το W_l στην έκφραση $W_l z + b_l$, είναι ένας πίνακας, ενώ το b_l είναι ένα διάνυσμα. Αυτό φαίνεται διαφορετικό από το wz + b της γραμμικής παλινδρόμησης. Στον πίνακα W_l, κάθε σειρά u αντιστοιχεί σε ένα διάνυσμα των παραμέτρων $\mathbf{w}_{l,\,u}$. Η διαστατικότητα του διανύσματος $\mathbf{w}_{l,\,u}$ ισούται με τον αριθμό των μονάδων στο στρώμα $l \neq 1$. Η λειτουργία $W_l\mathbf{z}$ έχει ως αποτέλεσμα ένα διάνυσμα $a_l \stackrel{\text{def}}{=} \left[w_{l,1}z, w_{l,2}z, \ldots, w_{l,sizel}z\right]$

Στη συνέχεια, το άθροισμα $a_l + b_l$ δίνει ένα διάνυσμα c_l διάστασης size l. Τέλος, η συνάρτηση $\boldsymbol{g}_{l(}\mathbf{c}_{l)}$ παράγει το διάνυσμα $y_l \stackrel{\text{def}}{=} \left[y_l^{(1)}, y_l^{(2)}, \ldots, y_l^{(sizel)}\right]$ ως έξοδο.

4.4.3. Βαθιά μάθηση (Deep Learning)

Η βαθιά εκμάθηση αφορά την κατάρτιση νευρωνικών δικτύων με περισσότερα από δύο επίπεδα μη εξόδου. Στο παρελθόν έγινε πιο δύσκολο να γίνει η διαδικασία μάθησης σε τέτοια δίκτυα καθώς ο αριθμός των στρώσεων αυξήθηκε. Οι δύο μεγαλύτερες προκλήσεις αναφέρθηκαν ως προβλήματα έκρηξης κλίσης (exploding gradient)και εξαφάνισης της κλίσης (vanishing gradient), καθώς η μέθοδος καθόδου κλίσης χρησιμοποιήθηκε για την εκπαίδευση των παραμέτρων του δικτύου.

Ενώ το πρόβλημα της έκρηξης της κλίσης ήταν ευκολότερο να αντιμετωπιστεί εφαρμόζοντας απλές τεχνικές όπως η κλιμάκωση της κλίσης και η κανονικοποίηση L1 ή L2, το πρόβλημα της εξαφάνισης της κλίσης παρέμεινε άλυτο για δεκαετίες.

Τι είναι η εξαφάνιση κλίσης και γιατί προκύπτει; Για την ενημέρωση των τιμών των παραμέτρων στα νευρωνικά δίκτυα χρησιμοποιείται συνήθως ο αλγόριθμος που ονομάζεται backpropagation. Ο αλγόριθμος backpropagation είναι ένας αποτελεσματικός αλγόριθμος για τον υπολογισμό των κλίσεων στα νευρωνικά δίκτυα χρησιμοποιώντας τον κανόνα της αλυσίδας.

Κατά τη διάρκεια της καθόδου κλίσης, οι παράμετροι του νευρωνικού δικτύου λαμβάνουν μια ενημέρωση ανάλογη με την μερική παράγωγο της συνάρτησης κόστους ως προς την τρέχουσα παράμετρο σε κάθε επανάληψη της εκπαίδευσης. Το πρόβλημα είναι ότι σε ορισμένες περιπτώσεις, η κλίση θα είναι πολύ μικρή, αποτρέποντας αποτελεσματικά ορισμένες παραμέτρους από την αλλαγή της τιμής τους.

Στη χειρότερη περίπτωση, αυτό μπορεί να σταματήσει τελείως το νευρωνικό δίκτυο από περαιτέρω εκπαίδευση. Οι παραδοσιακές συναρτήσεις ενεργοποίησης, όπως η συνάρτηση υπερβολικής εφαπτομένης, έχουν κλίσεις στην περιοχή (0, 1), και ο αλγόριθμος back propagation υπολογίζει κλίσεις από τον κανόνα της αλυσίδας.

Αυτό έχει ως αποτέλεσμα τον πολλαπλασιασμό n από αυτούς τους μικρούς αριθμούς για τον υπολογισμό των κλίσεων των παλαιότερων (αριστερά) στρώσεων σε ένα δίκτυο n-στρώσεων, που σημαίνει ότι η κλίση μειώνεται εκθετικά με το n. Αυτό έχει ως αποτέλεσμα το ότι τα προηγούμενα στρώματα μεταβάλλονται πολύ αργά, αν όχι καθόλου.

Ωστόσο, οι σύγχρονες εφαρμογές αλγορίθμων μάθησης νευρωνικών δικτύων επιτρέπουν να εκπαιδεύονται αποτελεσματικά πολύ βαθιά νευρωνικά δίκτυα (έως και εκατοντάδες επίπεδα). Αυτό οφείλεται σε αρκετές βελτιώσεις που συνδυάζονται μεταξύ τους, όπως το ReLU, το LSTM (και άλλες μονάδες με κλείδωμα), καθώς και τεχνικές όπως οι συνδέσεις παράκαμψης που χρησιμοποιούνται στα υπολειπόμενα νευρωνικά δίκτυα, καθώς και προχωρημένες τροποποιήσεις του αλγόριθμου καθόδου κλίσης.

Ως εκ τούτου, σήμερα, καθώς τα προβλήματα της εξαφάνισης και της έκρηξης κλίσης επιλύονται ως επί το πλείστων (ή μειώνεται η επίδρασή τους) σε μεγάλο βαθμό, ο όρος "βαθιά μάθηση" αναφέρεται στην κατάρτιση νευρωνικών δικτύων χρησιμοποιώντας τα σύγχρονα αλγοριθμικά και μαθηματικά εργαλεία ανεξάρτητα από το πόσο βαθύ είναι το νευρωνικό δίκτυο. Στην πράξη, πολλά επιχειρηματικά προβλήματα μπορούν να λυθούν με νευρωνικά δίκτυα που έχουν 2-3 στρώματα μεταξύ των επιπέδων εισόδου και εξόδου. Τα στρώματα που δεν εισάγονται ούτε εξάγονται συχνά αποκαλούνται κρυφά στρώματα.

5. Κατάλογος βάσεων δεδομένων για έρευνα μηχανικής μάθησης

5.1. Kaggle

Η Kaggle, θυγατρική της Google LLC, είναι μια διαδικτυακή κοινότητα επιστημόνων μαζικών δεδομένων (data scientists) και μηχανικών μηχανικής μάθησης (machine learning practitioners). Το Kaggle επιτρέπει στους χρήστες να βρίσκουν και να δημοσιεύουν σύνολα δεδομένων, να εξερευνούν και να δημιουργούν μοντέλα σε ένα διαδικτυακό περιβάλλον επιστημών μαζικών δεδομένων, να συνεργάζονται με άλλους επιστήμονες μαζικών δεδομένων και μηχανικούς μηχανικής μάθησης και να συμμετέχουν σε διαγωνισμούς για την επίλυση προκλήσεων επιστήμης δεδομένων.

Η Kaggle ξεκίνησε προσφέροντας διαγωνισμούς μηχανικής μάθησης και τώρα προσφέρει επίσης μια πλατφόρμα δημόσιων δεδομένων, έναν «πάγκο εργασίας» που βασίζεται σε υπολογιστικό νέφος (cloud) για την επιστήμη των δεδομένων και την εκπαίδευση της Τεχνητής Νοημοσύνης. Στις 8 Μαρτίου 2017, η Google ανακοίνωσε ότι εξαγόρασε την Kaggle.

Τον Ιούνιο του 2017, η Kaggle ανακοίνωσε ότι ξεπέρασε τους 1.000.000 εγγεγραμμένους χρήστες, ή Kagglers. Η κοινότητα εκτείνεται σε 194 χώρες. Είναι η μεγαλύτερη και πιο διαφοροποιημένη κοινότητα δεδομένων στον κόσμο, που οι χρήστες της κυμαίνονται από εκείνους που μόλις ξεκινούν ως πολλούς από τους πιο γνωστούς ερευνητές του κόσμου.

Οι διαγωνισμοί Kaggle προσελκύουν τακτικά πάνω από χίλιες ομάδες και μεμονωμένα άτομα. Η κοινότητα της Kaggle έχει χιλιάδες δημόσια σύνολα δεδομένων και αποσπάσματα κώδικα (που ονομάζονται «πυρήνες» (“kernels”) στο Kaggle). Πολλοί από αυτούς τους ερευνητές δημοσιεύουν άρθρα σε επιστημονικά περιοδικά με βάση την απόδοσή τους σε διαγωνισμούς Kaggle.

Μέχρι το Μάρτιο του 2017, το Ταμείο Two Sigma Investments διοργάνωνε διαγωνισμό στο Kaggle για να κωδικοποιήσει έναν αλγόριθμο διαπραγμάτευσης.

Πώς δουλεύουν οι διαγωνισμοί Kaggle

1. Ο διοργανωτής του αγώνα προετοιμάζει τα δεδομένα και μια περιγραφή του προβλήματος.

2. Οι συμμετέχοντες πειραματίζονται με διαφορετικές τεχνικές και ανταγωνίζονται μεταξύ τους για να παράγουν τα καλύτερα μοντέλα. Οι εργασίες μοιράζονται δημόσια μέσω των Kaggle Kernels για να επιτύχουν ένα καλύτερο σημείο αναφοράς και να εμπνεύσουν νέες ιδέες. Οι υποβολές μπορούν να γίνουν μέσω των πυρήνων Kaggle, μέσω της μη αυτόματης μεταφόρτωσης ή μέσω του API Kaggle. Για τους περισσότερους διαγωνισμούς, οι υποβολές βαθμολογούνται αμέσως (με βάση την ακρίβειά τους ως προς την πρόβλεψη σε σχέση με ένα αρχείο κρυφών λύσεων) και συνοψίζονται σε έναν πίνακα βαθμολογιών σε απευθείας αναμετάδοση.

3. Μετά την πάροδο της προθεσμίας, ο διοργανωτής του διαγωνισμού καταβάλλει το χρηματικό έπαθλο σε αντάλλαγμα με «μια παγκόσμια, αέναη, αμετάκλητη και απεριόριστη άδεια [...] να χρησιμοποιήσει τη νικήτρια είσοδο», δηλαδή τον αλγόριθμο, το λογισμικό και τη σχετική πνευματική ιδιοκτησία που είναι «μη αποκλειστική, εκτός εάν ορίζεται διαφορετικά».

Παράλληλα με τους δημόσιους διαγωνισμούς, η Kaggle προσφέρει επίσης ιδιωτικούς διαγωνισμούς που περιορίζονται στους κορυφαίους συμμετέχοντες της Kaggle. Η Kaggle προσφέρει ένα δωρεάν εργαλείο για τους καθηγητές των επιστημών δεδομένων για να τρέχουν ακαδημαϊκούς διαγωνισμούς μάθησης μηχανών, το Kaggle In Class [9]. Η Kaggle φιλοξενεί επίσης διαγωνισμούς πρόσληψης στους οποίους οι επιστήμονες δεδομένων ανταγωνίζονται για την ευκαιρία να κάνουν συνέντευξη σε κορυφαίες εταιρείες επιστήμης δεδομένων όπως το Facebook, η Winton Capital και η Walmart.

Η Kaggle διοργάνωσε εκατοντάδες διαγωνισμούς μηχανικής μάθησης από τότε που ιδρύθηκε η εταιρεία. Οι διαγωνισμοί κυμαίνονται από τη βελτίωση της αναγνώρισης χειρονομίας για τη Microsoft Kinect ως τη βελτίωση της αναζήτησης του μποζόνιου Higgs στο CERN.

Οι διαγωνισμοί έχουν οδηγήσει σε πολλά επιτυχημένα έργα, συμπεριλαμβανομένης της προώθησης της τεχνολογίας στον τομέα της έρευνας για τον ιό HIV, τις βαθμολογίες σκακιού και την πρόβλεψη της κυκλοφορίας. Οι πιο διάσημοι, ο Geoffrey Hinton και ο George Dahl χρησιμοποίησαν βαθιά νευρωνικά δίκτυα (deep neural networks) για να κερδίσουν έναν διαγωνισμό που φιλοξένησε η Merck. Και ο Vlad Mnih (ένας από τους σπουδαστές του Hinton) χρησιμοποίησε βαθιά νευρωνικά δίκτυα για να κερδίσει έναν διαγωνισμό που φιλοξένησε ο Adzuna. Αυτό βοήθησε να αναδειχθεί η δύναμη των βαθιών νευρωνικών δικτύων και είχε ως αποτέλεσμα η τεχνική να απορροφηθεί κι από άλλους στην κοινότητα Kaggle. Ο Tianqi Chen από το Πανεπιστήμιο της Ουάσιγκτον χρησιμοποίησε επίσης το Kaggle για να δείξει τη δύναμη του XGBoost, το οποίο από τότε έχει αναλάβει το Random Forest ως μία από τις κύριες μεθόδους που χρησιμοποιούνται για να κερδίσουν τους αγώνες Kaggle.

Αρκετές επιστημονικές εργασίες έχουν δημοσιευθεί με βάση τα πορίσματα των διαγωνισμών της Kaggle. Ένα κλειδί για αυτό είναι το αποτέλεσμα του πίνακα βαθμολογιών σε απευθείας αναμετάδοση, ο οποίος ενθαρρύνει τους συμμετέχοντες να συνεχίσουν να καινοτομούν πέρα από τις υπάρχουσες βέλτιστες πρακτικές (Athanasopoulos & Hyndman, 2011).

5.2. UC Irvine Machine Learning Repository

Το UCI Machine Learning Repository είναι μια συλλογή από βάσεις δεδομένων, θεωρίες τομέα και γεννήτριες δεδομένων που χρησιμοποιούνται από την κοινότητα μηχανών μάθησης για την εμπειρική ανάλυση αλγορίθμων μηχανικής μάθησης. Το αρχείο δημιουργήθηκε ως αρχείο ftp το 1987 από τον David Aha και τους συναδέλφους μεταπτυχιακούς φοιτητές του UC Irvine. Από τότε, έχει χρησιμοποιηθεί ευρέως από τους σπουδαστές, τους εκπαιδευτικούς και τους ερευνητές σε όλο τον κόσμο ως κύρια πηγή των συνόλων δεδομένων μηχανικής μάθησης. Ως ένδειξη του αντίκτυπου του αρχείου, έχει γίνει παραπομπή σε αυτό πάνω από 1000 φορές, καθιστώντας το ένα από τα 100 πιο παραπεμπόμενα "έγγραφα" σε όλες τις επιστήμες υπολογιστών. Η τρέχουσα έκδοση του ιστότοπου σχεδιάστηκε το 2007 από τους Arthur Asuncion και David Newman και το έργο αυτό είναι σε συνεργασία με το Rexa.info στο Πανεπιστήμιο της Μασαχουσέτης Amherst, με την ευγενή χορηγία του Εθνικού Ιδρύματος Επιστημών.

Το MLDB είναι μια βάση δεδομένων ανοιχτού κώδικα σχεδιασμένη για μηχανική μάθηση. Μπορεί να το εγκατασταθεί οποτεδήποτε και να δεχθεί εντολές μέσω ενός API RESTful, να αποθηκεύσει δεδομένα τα οποία μετέπειτα θα εξερευνηθούν από τον χρήστη χρησιμοποιώντας SQL, στη συνέχεια, να εκπαιδευτούν μοντέλα μηχανικής μάθησης και να εκτεθούν ως API.

Τα εξέχοντα εργαλεία για την μηχανική μάθηση και τη βαθειά εκμάθηση είναι τα MLDB, Keras, Edward, Lime, Apache Singa και Shogun. Οι λεπτομέρειες για κάθε μία μπορούν να βρεθούν στις αντίστοιχες ιστοσελίδες τους.

Η Βάση Δεδομένων Μηχανικής Μάθησης (Machine Learning Database - MLDB) είναι ένα ισχυρό και υψηλής απόδοσης σύστημα βάσεων δεδομένων ειδικά σχεδιασμένο για μηχανική μάθηση, ανακάλυψη γνώσεων και προγνωστική ανάλυση. Το MLDB είναι ένα λογισμικό ανοικτού κώδικα (Free and Open – Source Software – FOSS) και είναι συμβατό με διάφορες πλατφόρμες. Χρησιμοποιεί το RESTful API για την αποθήκευση δεδομένων, διερευνά τα δεδομένα χρησιμοποιώντας Δομημένη Γλώσσα Ερωτημάτων (Structured Query Language - SQL) και, τέλος, εκπαιδεύει μοντέλα μηχανικής μάθησης.

Τα παρακάτω είναι τα βασικά χαρακτηριστικά του MLDB που ταιριάζουν σε μια σειρά εφαρμογών:

- *Ταχύτητα:* Η διαδικασία κατάρτισης, μοντελοποίησης και ανακάλυψης στο MLDB έχει μεγάλη επίδοση. Έχει τεράστια δύναμη επεξεργασίας σε σύγκριση με το H2O, το Scikit-Learn ή το Spark MLib, που είναι σημαντικές βιβλιοθήκες μηχανικής μάθησης.
- *Εξομοιωσιμότητα*: Το MLDB υποστηρίζει την κατακόρυφη κλιμάκωση με υψηλότερη απόδοση, έτσι ώστε όλες οι μονάδες μνήμης καθώς και οι πυρήνες να μπορούν να χρησιμοποιηθούν ταυτόχρονα χωρίς προβλήματα καθυστέρησης ή απόδοσης.

- *Ελεύθερη και ανοιχτή πηγή*: Η κοινοτική έκδοση του MLDB διατίθεται και διανέμεται σε ένα ισχυρό αποθετήριο και φιλοξενία του GitHub (https://github.com/mldbai/mldb).
- *Υποστήριξη SQL:* Αυτό καθιστά το MLDB πολύ φιλικό προς το χρήστη μαζί με την υποστήριξη για την επεξεργασία μεγάλων δεδομένων. Το MLDB μπορεί να επεξεργαστεί, να εκπαιδεύσει και να κάνει προβλέψεις χρησιμοποιώντας πίνακες βάσεων δεδομένων που έχουν εκατομμύρια στήλες, με ταυτόχρονη επεξεργασία και χωρίς συμβιβασμούς στην ακεραιότητα.
- *Μηχανική μάθηση*: Το MLDB αναπτύσσεται για εφαρμογές και μοντέλα μηχανικής εκμάθησης υψηλής απόδοσης. Έχει υποστήριξη για βαθιά μάθηση με τα γραφήματα του TensorFlow που το κάνουν ανώτερο στην ανακάλυψη της γνώσης.
- *Ευκολία εφαρμογής:* Υπάρχουν πακέτα εγκατάστασης για πολλαπλές πλατφόρμες και περιβάλλοντα προγραμματισμού όπως το Jupyter, το Docker, το JSON, το Cloud, το Hadoop και πολλά άλλα.
- *Συμβατότητα και ενσωμάτωση:* Το MLDB παρέχει υψηλότερο βαθμό συμβατότητας με διαφορετικές διεπαφές προγραμματισμού εφαρμογών (Application Programming Interfaces - API) και μονάδες, συμπεριλαμβανομένων των JSON, REST και Python.
- *Ανάπτυξη:* Το MLDB μπορεί να αναπτυχθεί εύκολα σε ένα τελικό σημείο HTTP που παρέχει εύκολη διασύνδεση και γρήγορη ανάπτυξη.

- *Υποστήριξη MongoDB και NoSQL:* Η γέφυρα ή η διασύνδεση του MongoDB και του MLDB μπορούν να δημιουργηθούν για να υποστηρίξουν τα SQL ερωτήματα του MLDB. Αυτά τα ερωτήματα SQL μπορούν να εκτελεστούν σε συλλογές MongoDB, οι οποίες δίνουν στο MLDB περισσότερες δυνατότητες αλληλεπίδρασης με βάσεις δεδομένων NoSQL για αδόμητα και ετερογενή σύνολα δεδομένων.

6. Πρακτικό Μέρος - Υλοποίηση

6.1. Αλγόριθμοι ταξινόμησης

Σε αυτή την εργασία δοκιμάσαμε τους αλγορίθμους ταξινόμησης K-Nearest-Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF). Παρακάτω δίνεται μια θεωρητική περιγραφή των αλγορίθμων.

K-Κοντινότεροι Γείτονες (K-Nearest-Neighbors)

Αυτός είναι ένας πολύ απλός και αρκετά διαδεδομένος αλγόριθμος ταξινόμησης. Για κάθε δείγμα των δεδομένων δοκιμής, βρίσκει τους Κ κοντινότερους γείτονες από τα δεδομένα εκπαίδευσης. Η εύρεση των κοντινότερων γειτόνων πραγματοποιείται με κάποια μετρική, όπως η Ευκλείδεια απόσταση. Στη συνέχεια βρίσκεται ποια κατηγορία από τους Κ γείτονες έχει την πλειοψηφία και επιστρέφεται ως έξοδος.

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM)

Στον αλγόριθμο SVM η κατηγοριοποίηση των δεδομένων στηρίζεται στην εύρεση μιας βέλτιστης γραμμής (για δεδομένα δύο διαστάσεων) ή ενός βέλτιστου υπερεπιπέδου (για παραπάνω διαστάσεις) που διαχωρίζει τα δεδομένα δημιουργώντας το μέγιστο περιθώριο (Cortes, C., & Vapnik, 1995). Η ικανότητα γενίκευσης της χρήσης των SVM σε μη γραμμικά δεδομένα στηρίζεται στο τέχνασμα του πυρήνα (kernel trick). Στην περίπτωση που ο γραμμικός διαχωρισμός δεν είναι δυνατός, γίνεται χρήση κατάλληλων απεικονίσεων που μεταφέρουν το σύνολο των δεδομένων σε μεγαλύτερη διάσταση ώστε να επιτευχθεί τελικά ο διαχωρισμός τους. Το SVM είναι ένας δυαδικός ταξινομητής, έχει δηλαδή τη δυνατότητα κατηγοριοποίησης σε δύο κλάσεις.

Τυχαία Δάση (Random Forests - RF) είναι

Τα Τυχαία Δάση είναι μια μέθοδος ταξινόμησης, η οποία χρησιμοποιεί ένα μεγάλο αριθμό από ταξινομητικά και παλινδρομικά δένδρα (Classification and Regression Trees - CART) με σκοπό να παρέχει υψηλότερη ακρίβεια σε σχέση με ένα μεμονωμένο δένδρο αποφάσεων (Breiman, 2001). Άλλα πλεονεκτήματα είναι ότι τα δέντρα επιτρέπουν εύκολα την ερμηνεία των δεδομένων μέσω της σημαντικότητας των μεταβλητών. Επίσης, μπορούν να χειριστούν αποτελεσματικά μεγάλα σύνολα δεδομένων.

Ένα δέντρο είναι χτισμένο με τη λογική από την κορυφή προς τα κάτω (top-down) με αναδρομικό διαχωρισμό του χώρου των μεταβλητών. Σε κάθε βήμα, προσδιορίζεται ένα διαχωριστικό σημείο για κάθε μεταβλητή, έτσι ώστε να ελαχιστοποιείται ένα συγκεκριμένο κριτήριο στα δύο νέα υποσύνολα και η μεταβλητή που έχει ως αποτέλεσμα την ελάχιστη τιμή επιλέγεται να διαχωριστεί σε αυτόν τον κόμβο. Αυτή η διαδικασία ολοκληρώνεται όταν ένας κόμβος περιέχει έναν ελάχιστο αριθμό δειγμάτων ή όταν το κριτήριο δεν μπορεί να ελαχιστοποιηθεί παραπάνω από ένα όριο. Το προεπιλεγμένο κριτήριο τερματισμού είναι ο δείκτης Gini, ο οποίος δείχνει κατά πόσο τα δείγματα που καταλήγουν σε έναν κόμβο είναι «καθαρά», δηλαδή ανήκουν σε μία μόνο κλάση. Στη συνέχεια, το δέντρο μπορεί να κλαδευτεί για απλοποίηση και για να αποφευχθεί η υπερ-εκπαίδευση, βελτιώνοντας την ακρίβεια της δοκιμής, ωστόσο στο RF δεν πραγματοποιείται α αυτή η λειτουργία.

Τα RF δημιουργούν ένα μεγάλο αριθμό δένδρων χωρίς κλάδεμα, τα οποία διαφέρουν αρκετά μεταξύ τους λόγω της τυχαίας κατασκευής τους. Έτσι, τα δένδρα δεν συσχετίζονται μεταξύ τους, οπότε τα RF μπορούν να αποφύγουν την υπερ-προσαρμογή (overfit) στα δεδομένα εκπαίδευσης και μπορούν να επιτύχουν μεγαλύτερη ακρίβεια από ένα μεμονωμένο δέντρο. Η τυχαιότητα χρησιμοποιείται σε δύο σημεία: α) κάθε δέντρο εκπαιδεύεται σε ένα διαφορετικό υποσύνολο των δεδομένων (bagging), με ομοιόμορφη δειγματοληψία από τα αρχικά δείγματα, με αποτέλεσμα να περιέχει κατά μέσο όρο το 63% όλων των δειγμάτων και β) οι μεταβλητές που εξετάζονται για την καλύτερη διαίρεση σε κάθε κόμβο επιλέγονται τυχαία (παράμετρος mtry). Τέλος, η κλάση στην έξοδο καθορίζεται από την πλειοψηφία της εξόδου των μεμονωμένων δέντρων. Οι προεπιλεγμένες τιμές των παραμέτρων είναι για το mtry η ρίζα του αριθμού των

μεταβλητών, ο ελάχιστος αριθμός δειγμάτων είναι 2 και το κριτήριο τερματισμού είναι 10^{-7}.

Επιπλέον, τα RF παρέχουν μερικές ενσωματωμένες ιδιότητες, όπως η μετρική σημαντικότητα μεταβλητής (variable importance). Αυτή η μετρική υπολογίζεται σε κάθε κόμβο μέσω του δείκτη Gini και δείχνει ποσοτικά πόσο σημαντική είναι η παρουσία αυτής της μεταβλητής στο μοντέλο για τη σωστή ταξινόμηση.

6.2. Αξιολόγηση ταξινόμησης

Αρχικά, υπολογίζουμε τον πίνακα σύγχυσης, όπως φαίνεται παρακάτω:

		Πρόβλεψη κατηγορίας	
		Αρνητική	Θετική
Πραγματική κατηγορία	Αρνητική	True Negative (TN)	False Positive (FP)
	Θετική	False Negative (FN)	True Positive (TP)

Για την αξιολόγηση, χρησιμοποιήθηκαν οι παρακάτω μετρικές:

Accuracy = (TP + TN) / (TP + FP + FN +TN)

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN +FP)

6.3. Προεπεξεργασία Δεδομένων

Κατεβάσαμε το σύνολο δεδομένων Berlin Airbnb από το αποθετήριο Kaggle (https://www.kaggle.com/brittabettendorf/berlin-airbnb-data). Στη συνέχεια κάναμε τα παρακάτω βήματα για να καταλήξουμε στο τελικό σύνολο δεδομένων:

Επιλέξαμε μόνο τις πιο σχετικές στήλες, οι οποίες έχουν αριθμητικές τιμές ή μπορούν να μετατραπούν σε αριθμητικές:

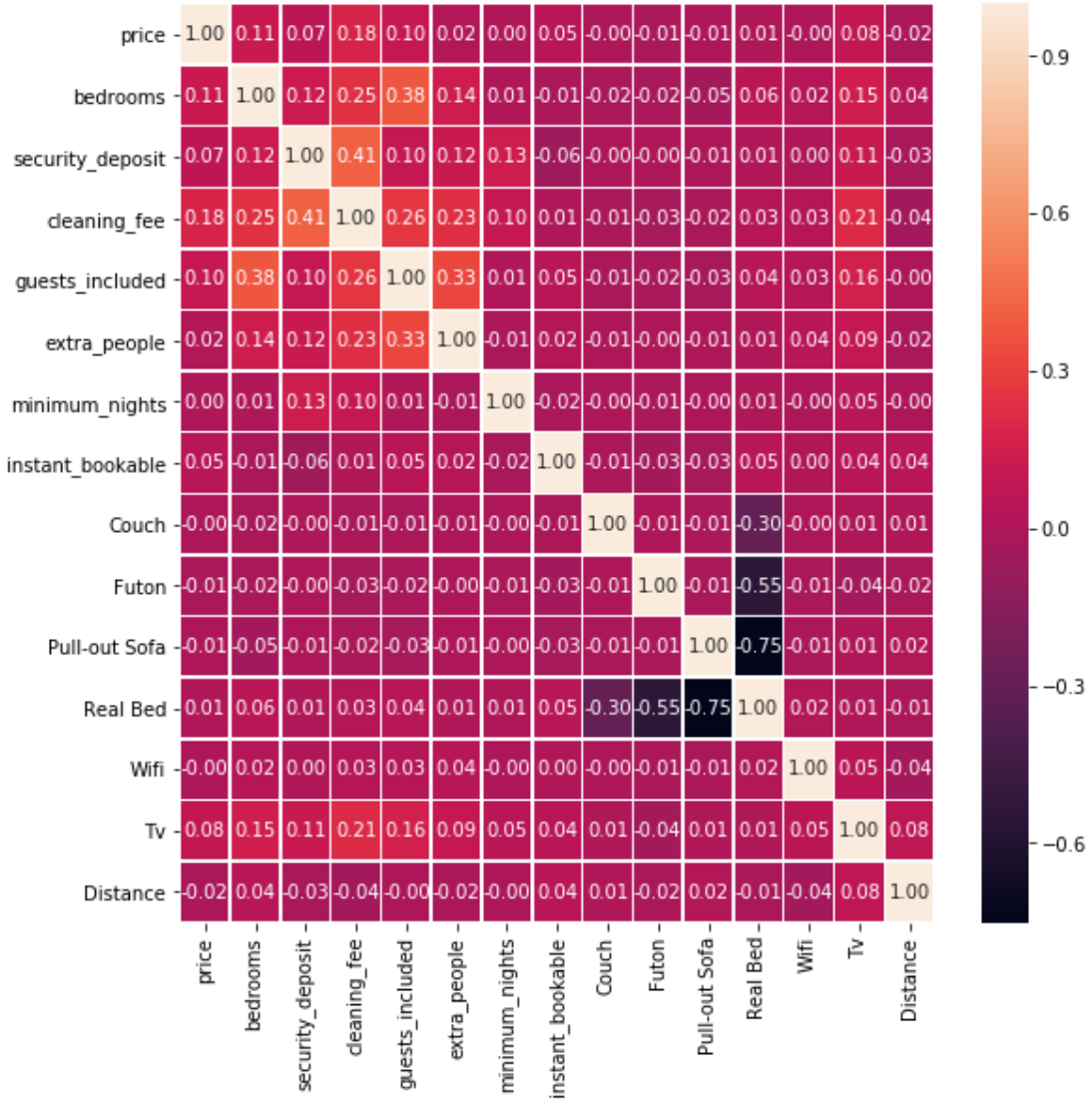
"price", "bedrooms", "bed_type", "amenities", "security_deposit", "cleaning_fee", "guests_included", "extra_people", "minimum_nights", "instant_bookable","latitude", "longitude"

Η πρώτη μεταβλητή, η τιμή, είναι αυτή που θέλουμε να προβλέψουμε. Στη συνέχεια δημιουργήσαμε επιπλέον στήλες από το bed_type (4 τύποι: Couch, Futon, Pull-out Sofa, Real Bed) και το amenities (επιλέξαμε μόνο αν υπάρχει wifi ή tv), οι οποίες έχουν τιμή 0 ή 1 αν υπάρχει το αντίστοιχο αντικείμενο στην αγγελία. Από τα πεδία τιμών, αφαιρέσαμε το σύμβολο του $ και το διαχωριστικό των χιλιάδων «,», ώστε να μείνει μόνο το αριθμητικό τμήμα. Από τις στήλες security_deposit και cleaning_fee αντικαταστήσαμε τα NaN με 0. Από τη στήλη bedrooms, όπου υπήρχε NaN αντικαταστάθηκε με 1. Έτσι καταλήξαμε σε ένα σύνολο δεδομένων μόνο με αριθμητικές τιμές, χωρίς NaN, το οποίο αποτελείται από 22552 δείγματα (γραμμές) και 16 μεταβλητές (στήλες).

Οι τελικές μεταβλητές του συνόλου δεδομένων είναι:

- price: η τιμή ενοικίασης (πραγματικός)
- bedrooms: αριθμός υπνοδωματίων (ακέραιος)
- security_deposit: ποσό εγγύησης (πραγματικός)
- cleaning_fee: κόστος καθαρισμού (πραγματικός)
- guests_included: η χωρητικότητα σε άτομα (ακέραιος)
- extra_people: κόστος φιλοξενίας επιπλέον ατόμων (πραγματικός)
- minimum_nights: ελάχιστος αριθμός διανυκτερεύσεων (ακέραιος)
- instant_bookable: αν είναι άμεσα διαθέσιμο (λογική, τιμή 0 ή 1)
- distance: απόσταση από κέντρο πόλης (πραγματικός σε χμλ) (υπολογίστηκε από το γεωγραφικό μήκος και πλάτος)
- Couch: αν έχει καναπέ (βρέθηκε από το bed type) (λογική, τιμή 0 ή 1)
- Futon: αν έχει στρώμα (βρέθηκε από το bed type) (λογική, τιμή 0 ή 1)
- Pull-out Sofa: αν έχει καναπέ-κρεβάτι (βρέθηκε από το bed type) (λογική, τιμή 0 ή 1)
- Real Bed: αν έχει κρεβάτι (βρέθηκε από το bed type) (λογική, τιμή 0 ή 1)
- Wifi: αν έχει wifi (βρέθηκε από το amenities) (λογική, τιμή 0 ή 1)
- Tv: αν έχει τηλεόραση (βρέθηκε από το amenities) (λογική, τιμή 0 ή 1)

Στην Εικόνα 1 φαίνεται η συσχέτιση μεταξύ των μεταβλητών των δεδομένων. Ενδιαφέρον παρουσιάζει κυρίως η συσχέτιση της τιμής (price) με τις υπόλοιπες μεταβλητές. Η μεγαλύτερη συσχέτιση είναι με τη μεταβλητή cleaning_fee (0.18) και στη συνέχεια με τη bedrooms (0.11) και guests_included (0.10).



Εικόνα 1. Οι συσχετίσεις μεταξύ των μεταβλητών του συνόλου δεδομένων.

6.4. Ταξινόμηση

Ταξινόμηση 2 κλάσεων

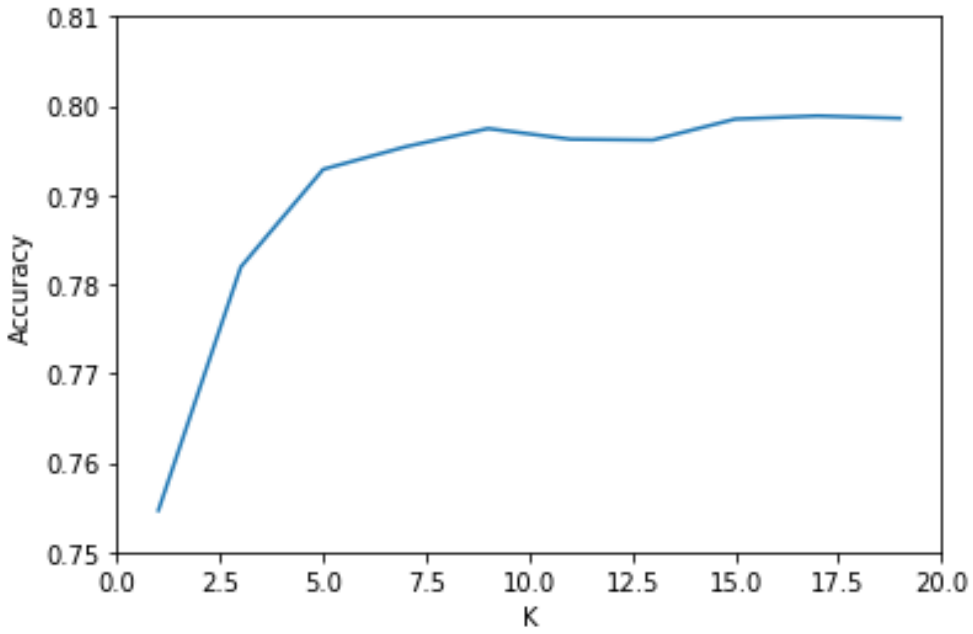
Για ταξινόμηση, απλά χωρίσαμε τις τιμές με βάση το μέσο όρο σε «φτηνές» (μικρότερες από το μέσο όρο) και «ακριβές» (ίσες ή μεγαλύτερες από το μέσο όρο). Οπότε καταλήξαμε σε 16385 τιμές ως φτηνές και 6167 τιμές ως ακριβές. Για την εύρεση της ακρίβειας ταξινόμησης υλοποιήσαμε cross-validation χωρίζοντας τα δεδομένα σε 10 τμήματα, όπου σε κάθε επανάληψη ένα τμήμα χρησιμοποιήθηκε για πρόβλεψη και τα υπόλοιπα εννέα για εκπαίδευση.

KNN

Αρχικά, δοκιμάσαμε τον αλγόριθμο KNN. Ο KNN δέχεται μια παράμετρο, τον αριθμό των γειτόνων K. Δοκιμάσαμε διάφορες τιμές του K από 1 ως 21 με βήμα 2 και υπολογίσαμε την ακρίβεια ταξινόμησης. Η ακρίβεια για τις διάφορες τιμές του K φαίνεται στην Εικόνα 2. Η υψηλότερη ακρίβεια του KNN είναι 79.70 % για K=19. Για αυτή την τιμή του K υπολογίσαμε τον πίνακα σύγχυσης (Πίνακας 1).

Πίνακας 1. Πίνακας Σύγχυσης για τον αλγόριθμο KNN

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	15086	1299
	Ακριβό	3279	2888



Εικόνα 2. Ακρίβεια του αλγορίθμου KNN για διάφορες τιμές του K.

SVM

Στη συνέχεια δοκιμάσαμε τον αλγόριθμο SVM με πυρήνα radial basis function (RBF). Για τον RBF, η παράμετρος gamma προσδιορίστηκε αυτόματα για τη βέλτιστη απόδοση. Ο SVM είχε ακρίβεια RBF 78.98%, ενώ ο πίνακας σύγχυσης παρουσιάζεται στον Πίνακα 2.

Πίνακας 2. Πίνακας Σύγχυσης για τον αλγόριθμο SVM

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	15469	916
	Ακριβό	3825	2342

RF

Τέλος, δοκιμάσαμε τον αλγόριθμο RF, όπου θέσαμε τον αριθμό των δένδρων σε 100, ενώ για την παράμετρο mtry δοκιμάσαμε τις τιμές $0.5 * \sqrt{n}, \sqrt{n}, 2 * \sqrt{n}$, όπου n ο αριθμός των μεταβλητών (εδώ έχουμε n=15, οπότε δοκιμάσαμε mtry = 2, 4, 8). Η υψηλότερη ακρίβεια είναι 81.80% για mtry=4 (ενώ λάβαμε ακρίβεια 81.7% για mtry=2 και 81.4% για mtry=8) και ο πίνακας σύγχυσης φαίνεται στον Πίνακα 3.

Πίνακας 3. Πίνακας Σύγχυσης για τον αλγόριθμο RF

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	14833	1552
	Ακριβό	2554	3613

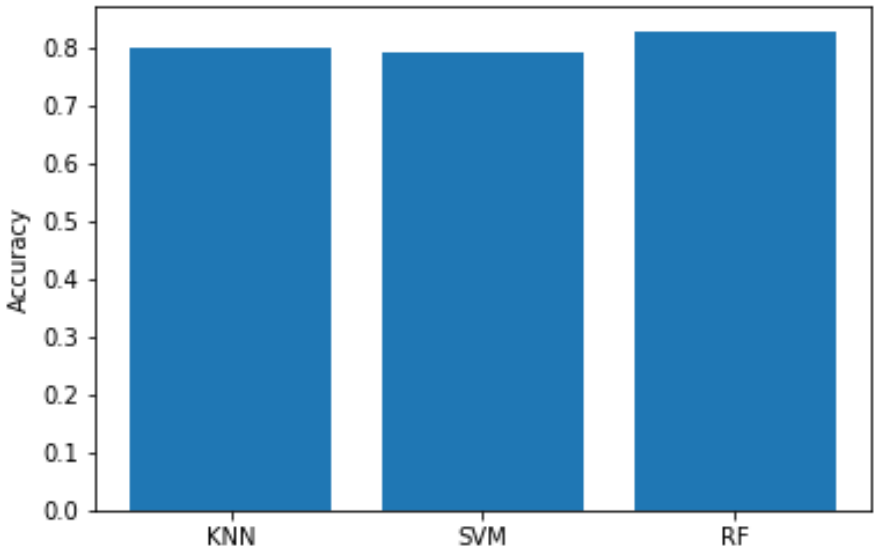
Η ακρίβεια όλων των αλγορίθμων φαίνεται στην Εικόνα 3. Παρατηρούμε ότι την υψηλότερη απόδοση είχε ο RF, ενώ τη χαμηλότερη ο SVM, αλλά με μικρή διαφορά από τον KNN. Σημειώνεται ότι ο RF είναι ένας πιο πολύπλοκος αλγόριθμος, ο οποίος συμπεριλαμβάνει πολλά δέντρα για να εξάγει την πρόβλεψη, οπότε δικαιολογείται η υψηλότερη ακρίβεια πρόβλεψης.

Με βάση όλους τους πίνακες σύγχυσης υπολογίσαμε την ευαισθησία (sensitivity) και την ειδικότητα (specificity) κάθε αλγορίθμου, οι οποίες φαίνονται μαζί με την ακρίβεια στον Πίνακα 4. Παρατηρούμε ότι ο RF έχει την καλύτερη ακρίβεια και ευαισθησία, ωστόσο στην ειδικότητα ο SVM είναι καλύτερος, αλλά με μικρή διαφορά.

Πίνακας 4. Ακρίβεια, ευαισθησία, ειδικότητα ταξινόμησης

Αλγόριθμος	**Ακρίβεια**	**Ευαισθησία**	**Ειδικότητα**
KNN	0.7970	0.4683	0.9207
SVM	0.7898	0.3798	0.9441
RF	0.8179	0.5859	0.9053

Σε πρόβλημα ταξινόμησης δύο κατηγοριών, η τυχαία πρόβλεψη έχει ακρίβεια 50%, οπότε το γεγονός ότι όλοι οι αλγόριθμοι πέτυχαν ακρίβεια κοντά στο 80% είναι αρκετά ικανοποιητικό αποτέλεσμα. Επίσης πρέπει να λάβουμε υπόψη ότι ο αριθμός των δειγμάτων (22000) είναι σχετικά μεγάλος, οπότε το πρόβλημα γίνεται πιο δύσκολο, ενώ το σύνολο δεδομένων προέρχεται από πραγματικά δεδομένα, τα οποία μπορεί να περιέχουν διάφορες ακραίες τιμές.



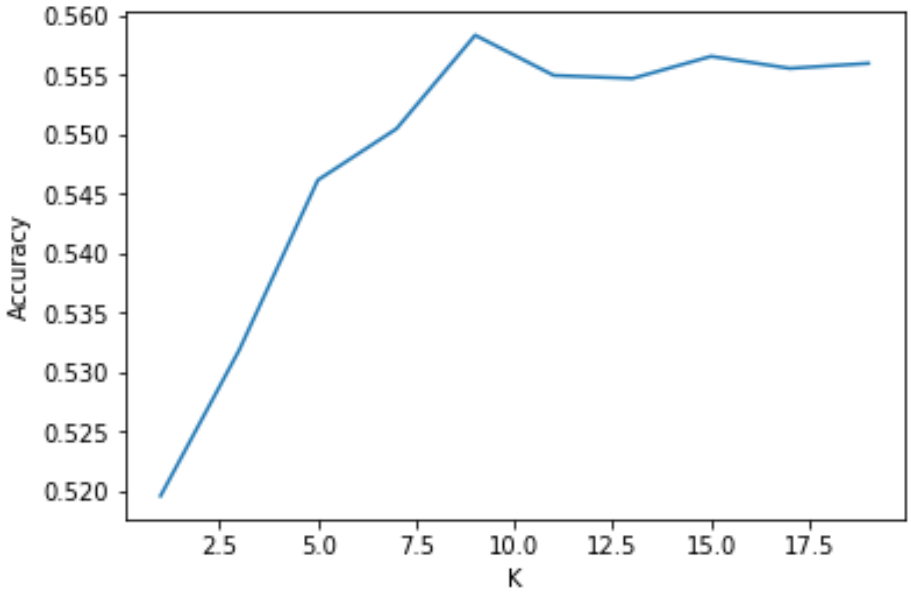
Εικόνα 3. Σύγκριση της ακρίβειας των διαφορετικών αλγορίθμων.

Ταξινόμηση 2 κλάσεων με υποδειγματοληψία (undersampling)

Οι δύο κατηγορίες έχουν μεγάλη διαφορά στον αριθμό δειγμάτων που περιέχουν, οπότε αυτό μπορεί να επηρεάζει το αποτέλεσμα. Για παράδειγμα, ένας αλγόριθμος που θα κατέτασσε όλα τα δείγματα ως φτηνά, θα πετύχαινε 69,7% ακρίβεια. Αυτό το πρόβλημα γίνεται φανερό από τη μεγάλη ειδικότητα (ανίχνευση κατηγορίας «φτηνό») και τη μικρή ευαισθησία (ανίχνευση κατηγορίας «ακριβό»). Οπότε για να αποφύγουμε αυτό το πρόβλημα, δοκιμάσαμε υποδειγματοληψία. Δηλαδή, πήραμε τυχαία 6167 δείγματα από την κατηγορία «φτηνές τιμές», ώστε να έχουμε ακριβώς τον ίδιο αριθμό δειγμάτων με την μικρότερη κατηγορία, οπότε έχουμε συνολικά 12334 δείγματα.

Επιπλέον, δοκιμάσαμε τον Naïve Bayes (NB) Classifier με Gaussian κατανομή καθώς και ένα δέντρο αποφάσεων (Decision Tree - DT). Παρακάτω δίνονται στους Πίνακες 5-9 οι πίνακες σύγχυσης για τους αλγόριθμος KNN, SVM και RF αντίστοιχα.

Στον Πίνακα 10 φαίνονται συγκεντρωτικά οι μετρικές ακρίβεια, ευαισθησία, ειδικότητα. Σημειώνουμε ότι στον KNN το καλύτερο αποτέλεσμα είναι για K=13, όπως φαίνεται στην Εικόνα 4, ενώ για τους υπόλοιπους αλγορίθμους ισχύουν όσα αναφέρθηκαν στην προηγούμενη ενότητα.



Εικόνα 4. Ακρίβεια του αλγορίθμου KNN για διάφορες τιμές του K.

Πίνακας 5. Πίνακας Σύγχυσης για τον αλγόριθμο KNN

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	4679	1488
	Ακριβό	1773	4394

Πίνακας 6. Πίνακας Σύγχυσης για τον αλγόριθμο SVM

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	4461	1706
	Ακριβό	1545	4622

Πίνακας 7. Πίνακας Σύγχυσης για τον αλγόριθμο RF

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	4717	1450
	Ακριβό	1444	4723

Πίνακας 8. Πίνακας Σύγχυσης για τον αλγόριθμο NB

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	5226	941
	Ακριβό	2416	3751

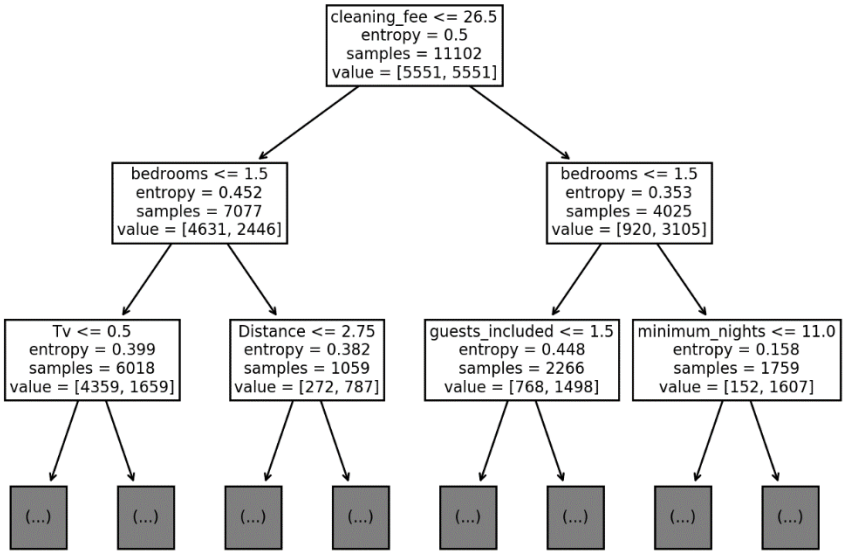
Πίνακας 9. Πίνακας Σύγχυσης για τον αλγόριθμο Decision Tree

		Πρόβλεψη κατηγορίας	
		Φτηνό	Ακριβό
Πραγματική κατηγορία	Φτηνό	4410	1757
	Ακριβό	1845	4322

Πίνακας 10. Ακρίβεια, ευαισθησία, ειδικότητα ταξινόμησης

Αλγόριθμος	**Ακρίβεια**	**Ευαισθησία**	**Ειδικότητα**
KNN	0.7356	0.7125	0.7587
SVM	0.7364	0.7495	0.7233
RF	0.7654	0.7659	0.7649
NB	0.7278	0.6082	0.8474
DT	0.7080	0.7008	0.7151

Παρατηρούμε ότι τώρα που οι κλάσεις είναι ισορροπημένες αυξήθηκε η ευαισθησία, ωστόσο μειώθηκε η ακρίβεια και η ειδικότητα, αλλά όλες οι μετρικές κυμαίνονται στα ίδια επίπεδα (71-77%). Σε αυτή την περίπτωση, ο RF εξακολουθεί να είναι ο καλύτερος αλγόριθμος σε ακρίβεια κατά 3% από τις άλλες δύο μεθόδους, όπως και στις μετρικές ευαισθησία και ειδικότητα. Ο NB αλγόριθμος έχει την υψηλότερη ειδικότητα, επειδή προβλέπει αρκετά καλά την φτηνή κατηγορία. Όμως έχει αρκετά χαμηλή ευαισθησία και συνολικά χαμηλότερη ακρίβεια από τους υπόλοιπους αλγορίθμους. Το δέντρο αποφάσεων έχει τη χαμηλότερη ακρίβεια, ωστόσο μας δίνει μια οπτική απεικόνιση του πώς οι μεταβλητές χρησιμοποιούνται για να εξαχθεί το αποτέλεσμα. Στην Εικόνα 5 παρατηρούμε ότι η πρώτη μεταβλητή είναι το cleaning_fee, μετά στο επόμενο επίπεδο και στους δύο κόμβους ρόλο παίζει η bedrooms, ενώ στο τρίτο επίπεδο έχουμε διάφορες μεταβλητές: tv, guests_included, distance, minimum_nights.



Εικόνα 5. Τα τρία πρώτα επίπεδα του δέντρου αποφάσεων.

Ταξινόμηση 3 κλάσεων

Στη συνέχεια δοκιμάσαμε να χωρίσουμε τις τιμές σε τρεις κατηγορίες. Σε αυτή την περίπτωση ορίζουμε ως «φτηνές τιμές» τις μικρότερες από το 80% της μέσης τιμής, ως «μεσαίες τιμές» όσες είναι 20% φτηνότερες έως 20% ακριβότερες από τη μέση τιμή και ως «ακριβές τιμές» τις μεγαλύτερες από το 120% της μέσης τιμής. Επομένως προκύπτουν τρεις κατηγορίες με 13549, 5064, 3939 δείγματα. Στη συνέχεια πραγματοποιούμε υποδειγματοληψία, οπότε παίρνουμε 3939 δείγματα από κάθε κατηγορία (συνολικά 11817 δείγματα). Ομοίως με πριν παραθέτουμε στους Πίνακες 11-15 τους πίνακες σύγχυσης για τους αλγορίθμους KNN, SVM, RF, NB, DT αντίστοιχα και στον Πίνακα 16 τις μετρικές ακρίβεια, ευαισθησία, ειδικότητα. Στις τρεις κλάσεις, υπολογίζουμε κάθε μετρική ξεχωριστά για κάθε κλάση (θεωρώντας τις άλλες δύο ως αρνητικές) και μετά παίρνουμε τον μέσο όρο. Σημειώνεται ότι το καλύτερο αποτέλεσμα για τον KNN είναι για K=9.

Πίνακας 11. Πίνακας Σύγχυσης για τον αλγόριθμο KNN

		Πρόβλεψη κατηγορίας		
		Φτηνό	Μεσαίο	Ακριβό
Πραγματική κατηγορία	Φτηνό	2774	833	332
	Μεσαίο	1421	1642	876
	Ακριβό	653	1104	2182

Πίνακας 12. Πίνακας Σύγχυσης για τον αλγόριθμο SVM

		Πρόβλεψη κατηγορίας		
		Φτηνό	Μεσαίο	Ακριβό
Πραγματική κατηγορία	Φτηνό	2570	778	591
	Μεσαίο	1184	1500	1255
	Ακριβό	492	871	2576

Πίνακας 13. Πίνακας Σύγχυσης για τον αλγόριθμο RF

		Πρόβλεψη κατηγορίας		
		Φτηνό	Μεσαίο	Ακριβό
Πραγματική κατηγορία	Φτηνό	2603	1007	329
	Μεσαίο	1049	1882	1008
	Ακριβό	373	976	2590

Πίνακας 14. Πίνακας Σύγχυσης για τον αλγόριθμο NB

		Πρόβλεψη κατηγορίας		
		Φτηνό	Μεσαίο	Ακριβό
Πραγματική κατηγορία	Φτηνό	2625	1104	210
	Μεσαίο	1133	2225	581
	Ακριβό	484	1682	1773

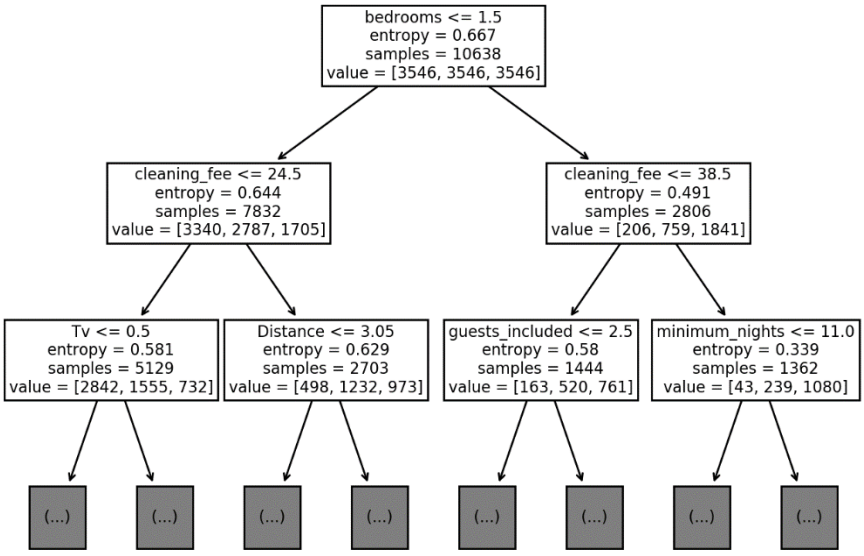
Πίνακας 15. Πίνακας Σύγχυσης για τον αλγόριθμο Decision Tree

		Πρόβλεψη κατηγορίας		
		Φτηνό	Μεσαίο	Ακριβό
Πραγματική κατηγορία	Φτηνό	2603	1007	329
	Μεσαίο	1049	1882	1008
	Ακριβό	373	976	2590

Πίνακας 16. Ακρίβεια, ευαισθησία, ειδικότητα ταξινόμησης

Αλγόριθμος	**Κλάση**	**Ακρίβεια**	**Ευαισθησία**	**Ειδικότητα**
KNN	1		0.7042	0.7367
	2		0.4169	0.7541
	3		0.5539	0.8467
	MO	**0.5583**	**0.55834**	**0.7792**
SVM	1		0.6524	0.7873
	2		0.3808	0.7906
	3		0.6540	0.7657
	MO	**0.5624**	**0.5624**	**0.7812**
RF	1		0.6608	0.8185
	2		0.4778	0.7483
	3		0.6575	0.8303
	MO	**0.5987**	**0.5987**	**0.7994**
NB	1		0.6664	0.7947
	2		0.5649	0.6464
	3		0. 4501	0.8996
	MO	**0.5604**	**0.5604**	**0.7802**
DT	1		0.5958	0.7862
	2		0.4265	0.7210
	3		0.5897	0.7988
	MO	**0.5374**	**0.5374**	**0.7687**

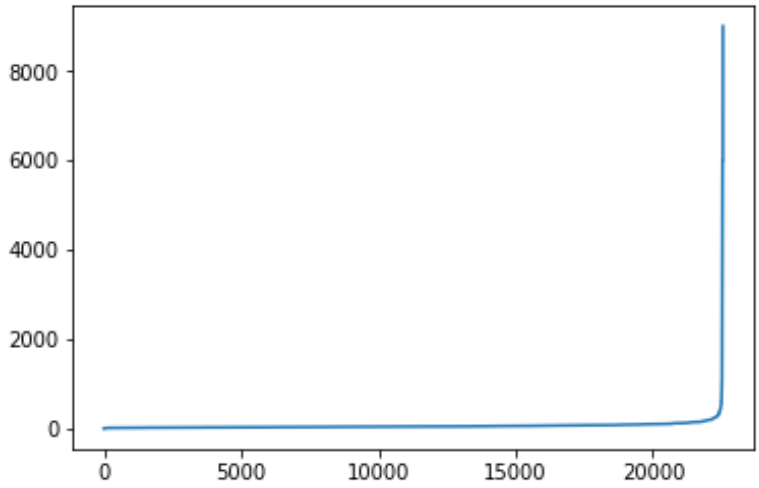
Από τα αποτελέσματα, παρατηρούμε ότι ο RF παραμένει ο καλύτερος αλγόριθμος σε ακρίβεια, ευαισθησία και ειδικότητα. Επίσης, από τους πίνακες σύγχυσης φαίνεται ότι είναι πιο δύσκολο να προβλεφθεί η μεσαία κατηγορία, το οποίο είναι λογικό, δηλαδή αναμένουμε αρκετά διαμερίσματα που είναι στα «όρια» μεταξύ των κατηγοριών να μοιάζουν μεταξύ τους. Τέλος, αναφέρουμε ότι με τρεις κατηγορίες, η τυχαία ταξινόμηση έχει ακρίβεια 33%, οπότε η ακρίβεια που επιτυγχάνουμε και με τους τρεις αλγορίθμους είναι ικανοποιητική. Στην Εικόνα 6 φαίνεται το δέντρο αποφάσεων για τις τρεις κατηγορίες. Σε σύγκριση με την Εικόνα 5, στην ρίζα είναι η μεταβλητή bedrooms και μετά η cleaning_fee, ενώ στο τρίτο επίπεδο βρίσκουμε τις ίδιες τέσσερις μεταβλητές με πριν. Επομένως για ταξινόμηση είτε σε δύο είτε σε τρεις κατηγορίες η δομή του δέντρου, δηλαδή οι μεταβλητές σε κάθε κόμβο, είναι παρόμοια, ενώ αλλάζουν επίσης λίγο τα όρια των αποφάσεων.



Εικόνα 6. Το δέντρο αποφάσεων για τις τρεις κλάσεις.

Παλινδρόμηση

Για να αποφύγουμε το χωρισμό σε κατηγορίες, θα αντιμετωπίζουμε την τιμή σαν συνεχή μεταβλητή και θα πραγματοποιήσουμε παλινδρόμηση. Θα χρησιμοποιήσουμε γραμμική παλινδρόμηση (linear regression) καθώς και ταξινόμηση με Random Forest (RF). Ως μετρικές αξιολόγησης παρέχουμε το μέσο τετραγωνικό σφάλμα (mean squared error – MSE) και το R^2. Αρχικά απεικονίζουμε την κατανομή των τιμών. Επειδή υπάρχουν μερικές ακραίες τιμές, θα κρατήσουμε μόνο τις τιμές κάτω από 500$, οπότε η κατανομή θα είναι πιο «γραμμική».



Εικόνα 5. Η κατανομή των τιμών.

Ομοίως με την ταξινόμηση, πραγματοποιήσαμε 10-fold cross-validation και υπολογίσαμε τις μετρικές MSE και R^2 από τα αποτελέσματα της πρόβλεψης στα δεδομένα δοκιμής. Τα αποτελέσματα των δύο αλγορίθμων πρόβλεψης φαίνονται στον Πίνακα 13.

Πίνακας 13.

Αλγόριθμος	**MSE**	**R^2**
Γραμμική Παλινδρόμηση	2747977.58	0.3566
RF	2240593.80	0.4734

Παρατηρούμε ότι ο αλγόριθμος RF έχει μικρότερο MSE καθώς και R^2 σε σχέση με τη γραμμική παλινδρόμηση. Αυτό είναι αναμενόμενο, καθώς ο αλγόριθμος RF είναι πιο πολύπλοκος και κατάλληλος για μη γραμμικά δεδομένα.

6.5. Συμπεράσματα

Σε αυτή την εργασία πραγματοποιήθηκε ταξινόμηση ενός πραγματικού συνόλου δεδομένων με μεγάλο αριθμό δειγμάτων, προερχόμενο από στοιχεία της ιστοσελίδας Airbnb για το Βερολίνο. Αρχικά, έγινε προεπεξεργασία των δεδομένων για να κρατήσουμε τις μεταβλητές με αριθμητικά δεδομένα και να μετατρέψουμε κάποιες κατηγορικές μεταβλητές σε αριθμητικά δεδομένα, ενώ επίσης απορρίψαμε δείγματα με ελλιπείς τιμές. Στη συνέχεια πραγματοποιήσαμε ταξινόμηση, χωρίζοντας τα δεδομένα σε δύο κατηγορίες (φτηνές και ακριβές τιμές με βάση το μέσο όρο). Επειδή παρατηρήσαμε ότι η μια κατηγορία έχει πολύ περισσότερα παραδείγματα, προχωρήσαμε σε υποδειγματοληψία για να έχουμε ισορροπημένες τις δύο κατηγορίες.

Στη συνέχεια δοκιμάσαμε διάφορους αλγορίθμους που χρησιμοποιούνται ευρέως για ταξινόμηση, τους KNN, SVM, RF, NB και δέντρα αποφάσεων (DT). Όλοι πέτυχαν αρκετά υψηλή απόδοση, με τον RF να επιτυγχάνει την υψηλότερη ακρίβεια και συγκεκριμένα η ακρίβεια ήταν 76.54%, ενώ οι υπόλοιποι αλγόριθμοι κατά σειρά ακρίβειας ήταν ο SMV (73.64%), ο KNN (73.56%), ο NB (72.8%) και ο DT (70.8%). Παρατηρούμε ότι όλοι έχουν σχετικά παρόμοια απόδοση, στο διάστημα 70%-76.5%. Παρά τη χαμηλή ακρίβεια, το δέντρο αποφάσεων μπορεί να μας δώσει μια οπτική απεικόνιση για τη σημαντικότητα των μεταβλητών, όπου παρατηρούμε ότι πιο σημαντική μεταβλητή για την ταξινόμηση της τιμής ενοικίασης είναι το κόστος καθαρισμού, μετά ο αριθμός υπνοδωματίων και στη συνέχεια μεταβλητές όπως η ύπαρξη τηλεόρασης, η απόσταση από το κέντρο, η χωρητικότητα σε άτομα και ο ελάχιστος αριθμός διανυκτερεύσεων.

Στο επόμενο βήμα, προχωρήσαμε σε ταξινόμηση τριών επιπέδων (φτηνό, μεσαίο, ακριβό). Πάλι ο καλύτερος αλγόριθμος ήταν ο RF (59.9%), ενώ στη συνέχεια με σειρά ακρίβειας ακολούθησαν ο SVM (56.2%), ο NB (56%), ο KNN (55.8%) και ο DT (53.7%). Παρατηρούμε ότι η ακρίβεια είναι σχετικά χαμηλή, αλλά υπενθυμίζουμε ότι σε τρία επίπεδα η τυχαία πρόβλεψη έχει ακρίβεια 33.3%. Ομοίως, εξετάζοντας τις πιο σημαντικές μεταβλητές με βάση την απεικόνιση του DT, πρώτα παίζει ρόλο ο αριθμός των υπνοδωματίων, στη συνέχεια το κόστος καθαρισμού και μετά οι τέσσερις μεταβλητές που αναφέραμε προηγουμένως. Και στις δύο περιπτώσεις ο RF είχε υψηλότερη απόδοση, το οποίο οφείλεται ότι είναι πιο πολύπλοκος (εκπαιδεύει πολλά DT), ενώ ο SVM ήταν ο δεύτερος καλύτερος και έπειτα ακολουθούν οι NB και KNN, ενώ ο DT ήταν ο τελευταίος. Πάντως, οι υπόλοιποι αλγόριθμοι, αν και είχαν μικρότερη απόδοση, δεν είχαν πολύ μεγάλη διαφορά από τη βέλτιστη τιμή.

Τέλος, θεωρώντας την τιμή ενοικίασης ως μια συνεχή μεταβλητή, πραγματοποιήσαμε παλινδρόμηση για πρόβλεψη της τιμής. Σε αυτή την περίπτωση χρησιμοποιήσαμε δύο μεθόδους, γραμμική παλινδρόμηση και RF παλινδρόμηση. Η γραμμική παλινδρόμηση οδήγησε σε R^2 = 35.66%, ενώ το RF σε R^2 = 47.34%. Η καλύτερη απόδοση του RF είναι αναμενόμενη, καθώς αυτός ο αλγόριθμος είναι πιο πολύπλοκος, ενώ είναι κατάλληλος και για μη γραμμικά δεδομένα.

Επομένως, δείξαμε ότι είναι δυνατή η πρόβλεψη της τιμής ενός διαμερίσματος από χαρακτηριστικά που προέρχονται από την περιγραφή σε αγγελίες. Αυτή η υλοποίηση μπορεί να χρησιμοποιηθεί για να διαπιστώσουμε αν η πραγματική τιμή είναι φτηνή ή ακριβή σε σχέση με παρόμοια διαμερίσματα ή για να προβλέψει την προτεινόμενη τιμή σε ένα διαμέρισμα με συγκεκριμένα χαρακτηριστικά.

ΠΑΡΑΡΤΗΜΑ:

Κώδικας Berlin code.pdf

Βιβλιογραφία

Alawadi, K., Khanal, A., & Almulla, A. (2018). Land, urban form, and politics: A study on Dubai's housing landscape and rental affordability. *Cities*, *81*, 115-130.

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *applied Geography*, *49*, 45-53.

Athanasopoulos, G., & Hyndman, R. J. (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting*, *27*(3), 845-849.

August, M., & Walks, A. (2018). Gentrification, suburban decline, and the financialization of multi-family rental housing: The case of Toronto. *Geoforum*, *89*, 124-136.

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKAâˆ’Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, *11*(Sep), 2533-2541.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., ... & Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). *Intensive Working Group of ACM SIGKDD Curriculum Committee*, *140*.

Chen, Y. H., & Fik, T. (2017). Housing-market bubble adjustment in coastal communities-A spatial and temporal analysis of housing prices in Midwest Pinellas County, Florida. *Applied Geography*, *80*, 48-63.

Chen, Y., Liu, X., Li, X., Liu, Y., & Xu, X. (2016). Mapping the fine-scale spatial pattern of housing rent in the metropolitan area by using online rental listings and ensemble learning. *Applied geography*, *75*, 200-212.

Choumert, J., Stage, J., & Uwera, C. (2014). Access to water as determinant of rental values: A housing hedonic analysis in Rwanda. *Journal of Housing Economics*, *26*, 48-54.

Chouthai, A., Rangila, M. A., Amate, S., Adhikari, P., & Kukre, V. (2019). HOUSE PRICE PREDICITION USING MACHINE LEARNING.

Clifton, C. (2010). Encyclopædia britannica: definition of data mining. *Retrieved on*, *9*(12), 2010.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Depetris-Chauvin, E., & Santos, R. J. (2018). Unexpected guests: The impact of internal displacement inflows on rental prices in Colombian host cities. *Journal of Development Economics*, *134*, 289-309.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Feng, C., Li, W. X., & Zhao, F. F. (2011). Influence of rail transit on nearby commodity housing prices: A case study of Beijing Subway Line Five. *Acta Geographica Sinica*, *66*(8), 1055-1062.

Fesselmeyer, E., & Seah, K. Y. S. (2018). The effect of localized density on housing prices in Singapore. *Regional Science and Urban Economics*, *68*, 304-315.

Geng, B., Bao, H., & Liang, Y. (2015). A study of the effect of a high-speed rail station on spatial variations in housing price based on the hedonic model. *Habitat International*, *49*, 333-339.

Gilbert, A. (2016). Rental housing: The international experience. *Habitat International*, *54*, 173-181.

Gluszak, M., & Zygmunt, R. (2018). Development density, administrative decisions, and land values: An empirical investigation. *Land Use Policy*, *70*, 153-161.

Han, J., & Kamber, M. (2001). Data mining: concepts and techniques. Copyright by Morgan Kaufmann Publishers.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1-12.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hogan, B., & Berry, B. (2011). Racial and ethnic biases in rental housing: An audit study of online apartment listings. *City & Community*, *10*(4), 351-372.

Höppner, F. (2005). Association rules. In *Data Mining and Knowledge Discovery Handbook* (pp. 353-376). Springer, Boston, MA.

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land use policy*, *82*, 657-673.

Huang, Z., Chen, R., Xu, D., & Zhou, W. (2017). Spatial and hedonic analysis of housing prices in Shanghai. *Habitat International*, *67*, 69-78.

Hui, E. C., Zhong, J., & Yu, K. (2016). Land use, housing preferences and income poverty: In the context of a fast rising market. *Land use policy*, *58*, 289-301.

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *arXiv preprint arXiv:1907.12665*.

Kantardzic, M. (2011). Data Mining Concepts, Models, Methods, and Algorithms. A John Wiley & Sons. *Inc. Hoboken, New Jersey*.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). STHDA.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of political economy*, *74*(2), 132-157.

Li, J., Wang, C. C., & Sun, J. (2017). Empirical analysis of tenants' intention to exit public rental housing units based on the Theory of Planned Behavior–The case of Wuhan, China. *Habitat International*, *69*, 27-36.

Li, Y., Pan, Q., Yang, T., & Guo, L. (2016, July). Reasonable price recommendation on Airbnb using Multi-Scale clustering. In *2016 35th Chinese Control Conference (CCC)* (pp. 7038-7041). IEEE.

Lison, P. (2015). An introduction to machine learning. *Language Technology Group (LTG), 1*, *35*.

Luo, Y., Zhou, X., & Zhou, Y. (2019). Predicting Airbnb Listing Price Across Different Cities.

Mao, R. (2001). *Adaptive-FP: an efficient and effective method for multi-level multi-dimensional frequent pattern mining* (Doctoral dissertation, Simon Fraser University).

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. In *Data mining and knowledge discovery in real life applications*. IntechOpen.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning. *Neural and Statistical Classification*, *13*.

Mirkatouli, J., Samadi, R., & Hosseini, A. (2018). Evaluating and analysis of socio-economic variables on land and housing prices in Mashhad, Iran. *Sustainable cities and society*, *41*, 695-705.

Olson, D. L. (2006). Data mining in business services. Service Business, 1 (3), 181–193.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research*, *50*, 67-83.

Piatetsky-Shapiro, G. (2002). KDnuggets methodology poll. *Gregory Piatetsky-Shapiro (2004) KDnuggets Methodology Poll, Gregory Piatetsky-Shapiro (2007) KDnuggets Methodology Poll, Gregory Piatetsky-Shapiro (2014) KDnuggets Methodology Poll*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Rae, A., & Sener, E. (2016). How website users segment a city: The geography of housing search in London. *Cities*, *52*, 140-147.

Rondinelli, C., & Veronese, G. (2011). Housing rent dynamics in Italy. *Economic Modelling*, *28*(1-2), 540-548.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, *82*(1), 34-55.

Sadayuki, T. (2018). Measuring the spatial effect of multiple sites: An application to housing rent and public transportation in Tokyo, Japan. *Regional Science and Urban Economics*, *70*, 155-173.

Teubner, T., Hawlitschek, F., & Dann, D. (2017). Price determinants on AirBnB: How reputation pays off in the sharing economy. *Journal of Self-Governance & Management Economics*, *5*(4).

Waltert, F., & Schläpfer, F. (2010). Landscape amenities and local development: A review of migration, regional economic and hedonic pricing studies. *Ecological Economics*, *70*(2), 141-152.

Waltl, S. R. (2018). Estimating quantile-specific rental yields for residential housing in Sydney. *Regional Science and Urban Economics*, *68*, 204-225.

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com. *International Journal of Hospitality Management*, *62*, 120-131.

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb. com. *International Journal of Hospitality Management*, *62*, 120-131.

Wang, Y., Wang, S., Li, G., Zhang, H., Jin, L., Su, Y., & Wu, K. (2017). Identifying the determinants of housing prices in China using spatial regression and the geographical detector technique. *Applied Geography*, *79*, 26-36.

Wen, H., & Tao, Y. (2015). Polycentric urban structure and housing price in the transitional China: Evidence from Hangzhou. *Habitat International*, *46*, 138-146.

Wen, H., Xiao, Y., & Zhang, L. (2017). Spatial effect of river landscape on housing price: An empirical study on the Grand Canal in Hangzhou, China. *Habitat International*, *63*, 34-44.

Wen, H., Zhang, Y., & Zhang, L. (2014). Do educational facilities affect housing price? An empirical study in Hangzhou, China. *Habitat International*, *42*, 155-163.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, C., Ye, X., Ren, F., Wan, Y., Ning, P., & Du, Q. (2016). Spatial and social media data analytics of housing prices in Shenzhen, China. *PloS one*, *11*(10).

Wu, F. (2016). Housing in Chinese urban villages: the dwellers, conditions and tenancy informality. *Housing Studies*, *31*(7), 852-870.

Yao, Y., Zhang, J., Hong, Y., Liang, H., & He, J. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS*, *22*(2), 561-581.

Yi, C., & Huang, Y. (2014). Housing consumption and housing inequality in Chinese cities during the first decade of the twenty-first century. *Housing Studies*, *29*(2), 291-311.

Zhang, C., Jia, S., & Yang, R. (2016). Housing affordability and housing vacancy in China: The role of income inequality. *Journal of housing Economics*, *33*, 4-14.

Zhang, L., & Yi, Y. (2017). Quantile house price indices in Beijing. *Regional Science and Urban Economics*, *63*, 85-96.

Zhang, Q., Gao, W., Su, S., Weng, M., & Cai, Z. (2017). Biophysical and socioeconomic determinants of tea expansion: Apportioning their relative importance for sustainable land use policy. *Land Use Policy*, *68*, 438-447.

Zheng, X., Xia, Y., Hui, E. C., & Zheng, L. (2018). Urban housing demand, permanent income and uncertainty: Microdata analysis of Hong Kong's rental market. *Habitat International*, *74*, 9-17.

Zhu, X. J. (2005). *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.