ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΤΜΗΜΑ
ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

“Μέθοδοι μηχανικής εκμάθησης για ταξινόμηση και παλινδρόμηση”

Υπο τους φοιτητές:
Παπαδόπουλο Δημητριο 2537
Παπαϊωάννου Ιωάννη 2427

Επιβέπων Αντωνόπουλος Χρήστος, Επίκουρος καθηγητής
Πάτρα 2021

Περιεχόμενα

Πρόλογος

Η κύρια ιδέα της συγκεκριμένης εργασίας είναι η επιθυμία στο να μάθουμε κάποιους μεθόδους ταξινόμησης και παλινδρόμησης με κάποιους αλγόριθμους.Ετσι λοιπόν ασχοληθήκαμε με την μηχανική μάθηση, όπου θα διαλέξουμε ένα μεγάλο σύνολο δεδομένων κατάλληλο για μεθόδους μηχανικής μάθησης, όπως το Berlin Airbnb dataset, το οποίο είναι διαθέσιμο στο αποθετήριο Kaggle: https://www.kaggle.com/brittabettendorf/berlin-airbnb-data . Στη συνέχεια θα ακολουθήσουμε δύο είδη ανάλυσης: ταξινόμηση και παλινδρόμηση για πρόβλεψη της τιμής ενός διαμερίσματος. Στην ταξινόμηση θα χωρίσουμε τις τιμές σε κατηγορίες και θα προβλέπουμε αν ένα διαμέρισμα είναι ακριβό ή φτηνό. Θα δοκιμάσουμε διάφορες τιμές των παραμέτρων τους για να βρούμε τη βέλτιστη απόδοση. Στην παλινδρόμηση, θα χειριστούμε την τιμή σαν μια συνεχή μεταβλητή και θα δοκιμάσουμε να την προβλέψουμε. Θα δοκιμάσουμε επίσης διάφορους αλγορίθμους, όπως γραμμική παλινδρόμηση και τυχαίο δάσος.

Περίληψη εργασίας πρακτικού μέρους

Στην παρούσα εργασία θα παραρουσιαστούν και θα αναλυθούν οι τεχνικές και οι διαδικασίες πρόβλεψης κάποιων συγκεκριμένων τιμών. Πιο συγκερκιμένα θα εργαστούμε πάνω στο AirBnB, όπου ως βασικός στόχος είναι η πρόβλεψη τιμών των διαμερισμάτων σε μελλοντικό χρόνο με βάση τις τιμές που έχουμε ως δεδομένα απο το παρελθόν.Χρησιμοποιώντας τις βάσεις δεδομένων που διαθέτει η Kaggle θα εργαστούμε όπως αναφέρθηκε και παραπάνω στην πρόβληψη τίμων των διαμερισμάτων του Βερολίνου. Η εργασία θα υλοποιηθεί με γλώσσα python, όπου το πρώτο βήμα θα είναι η συλλογή και η ταξινόμηση των δεδομένων ώστε να μπορούμε να κάνουμε plot στα διαμερίσματα και να βλέπουμε μέσα απο διαγράμματα πως κυμένονται οι τιμές στο χρόνο. Στη συνέχεια χρησιμοποιώντας κάποιους αλγορίθμους θα προσπαθήσουμε να προβλέψουμε τις πιθανές τιμές που θα έχουν τα διαμερίσματα στο μέλλον και κατά πόσο σωστές θα είναι οι προβλέψεις που θα παίρνουμε ως αποτέλεσμα.

1. Εισαγωγή

Τα προβλήματα ταξινόμησης κειμένου έχουν μελετηθεί ευρέως και αντιμετωπίζονται σε πολλές πραγματικές εφαρμογές (Jiang et al., 2018; Kowsari et al., 2017) τις τελευταίες δεκαετίες. Ειδικά με τις πρόσφατες ανακαλύψεις στην Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing-NLP) και την εξαγωγή κειμένου, πολλοί ερευνητές ενδιαφέρονται για την ανάπτυξη εφαρμογών που αξιοποιούν μεθόδους ταξινόμησης κειμένου. Τα περισσότερα συστήματα ταξινόμησης κειμένων και κατηγοριοποίησης εγγράφων μπορούν να αποδομηθούν στις ακόλουθες τέσσερις φάσεις: Εξαγωγή χαρακτηριστικών, μείωση διαστάσεων, επιλογή ταξινομητή και αξιολογήσεις. Στην εργασία των (Kowsari et al., 2019), συζητήθηκαν η δομή και οι τεχνικές υλοποιήσεις των συστημάτων ταξινόμησης κειμένου[1].

Η αρχική είσοδος του αγωγού αποτελείται από κάποιο σύνολο δεδομένων ακατέργαστου κειμένου. Γενικά, τα σύνολα δεδομένων κειμένου περιέχουν ακολουθίες κειμένου σε έγγραφα ως $D = \{X_1, X_2, \ldots, X_N\}$ όπου το X_i αναφέρεται σε ένα σημείο δεδομένων (π.χ. έγγραφο, τμήμα κειμένου) με s τον αριθμό των προτάσεων έτσι ώστε κάθε πρόταση να περιλαμβάνει λέξεις w_s με l_w γράμματα. Κάθε σημείο επισημαίνεται με μια τιμή τάξης από ένα σύνολο k διαφορετικών δεικτών διακριτής τιμής.

Στη συνέχεια, πρέπει να δημιουργηθεί ένα δομημένο σύνολο για τους εκπαιδευτικούς σκοπούς και η λειτουργία αυτή ονομάζεται «Εξαγωγή χαρακτηριστικών». Το βήμα μείωσης διαστάσεων είναι ένα προαιρετικό μέρος του αγωγού που θα μπορούσε να είναι τμήμα του συστήματος ταξινόμησης (π.χ., εάν χρησιμοποιείται Συχνότητα όρων-Συχνότητα ανάστροφων εγγράφων (Term Frequency-Inverse Document Frequency-TF-IDF) ως εξαγωγή χαρακτηριστικών και ένα σύνολο με 200k μοναδικές λέξεις, ο υπολογιστικός χρόνος είναι πολύ ακριβός, οπότε θα έπρεπε να μειωθεί αυτήν η επιλογή φέρνοντας χώρο χαρακτηριστικών σε άλλο χώρο άλλων διαστάσεων). Το πιο σημαντικό βήμα στην κατηγοριοποίηση εγγράφων είναι η επιλογή του καλύτερου αλγορίθμου ταξινόμησης. Το άλλο μέρος του αγωγού είναι το στάδιο αξιολόγησης που χωρίζεται σε δύο μέρη (πρόβλεψη του συνόλου δοκιμών και αξιολόγηση του μοντέλου). Γενικά, το σύστημα ταξινόμησης

[1] Ο πηγαίος κώδικας και τα αποτελέσματα κοινοποιούνται στη διεύθυνση: https://github.com/kk7nc/Text_Classification

κειμένου περιέχει τέσσερα διαφορετικά επίπεδα εμβέλειας που μπορούν να εφαρμοστούν:

- **Επίπεδο εγγράφου:** Στο επίπεδο εγγράφου, ο αλγόριθμος λαμβάνει τις σχετικές κατηγορίες ενός πλήρους εγγράφου.
- **Επίπεδο παραγράφου:** Στο επίπεδο παραγράφου, ο αλγόριθμος αποκτά τις σχετικές κατηγορίες μιας παραγράφου (τμήμα ενός εγγράφου).
- **Επίπεδο πρότασης:** Στο επίπεδο των προτάσεων, ο αλγόριθμος αποκτά τις σχετικές κατηγορίες μιας μεμονωμένης πρότασης (ένα τμήμα μιας παραγράφου).
- **Επίπεδο υπο-πρότασης:** Στο επίπεδο υπο-πρότασης, ο αλγόριθμος αποκτά τις σχετικές κατηγορίες υπο-εκφράσεων μέσα σε μια πρόταση (ένα τμήμα μιας πρότασης).

(I) **Εξαγωγή χαρακτηριστικών**: Σε γενικές γραμμές, τα κείμενα και τα έγγραφα είναι σύνολα μη δομημένων δεδομένων. Ωστόσο, αυτές οι μη δομημένες ακολουθίες κειμένου πρέπει να μετατραπούν σε χώρο δομημένων χαρακτηριστικών κατά τη χρήση μαθηματικών μοντέλων ως μέρος ενός ταξινομητή. Κατ 'αρχάς, τα δεδομένα πρέπει να καθαριστούν ώστε να παραληφθούν τυχόν περιττοί χαρακτήρες και λέξεις. Μετά τον καθαρισμό των δεδομένων, μπορούν να εφαρμοστούν επίσημες μέθοδοι εξαγωγής χαρακτηριστικών. Οι κοινές τεχνικές εξαγωγής χαρακτηριστικών είναι η Συχνότητα όρων-Συχνότητα ανάστροφων εγγράφων (Term Frequency-Inverse Document Frequency,TF-IDF), η συχνότητα όρου (Term Frequency-TF) (Salton & Buckley, 1988), η Word2Vec (Goldberg & Levy, 2014) και η Global Vectors for Word Representation (GloVe) (Pennington et al., 2014).

(II) **Μείωση διαστάσεων**: Καθώς τα σύνολα δεδομένων κειμένου ή εγγράφων περιέχουν συχνά πολλές μοναδικές λέξεις, τα βήματα προεπεξεργασίας δεδομένων μπορεί να καθυστερήσουν λόγω του υψηλού χρόνου και της πολυπλοκότητας της μνήμης. Μια κοινή λύση σε αυτό το πρόβλημα είναι απλά η χρήση φθηνών αλγορίθμων. Ωστόσο, σε ορισμένα σύνολα δεδομένων, αυτά τα είδη φθηνών αλγορίθμων δεν αποδίδουν σύμφωνα με το αναμενόμενο. Προκειμένου να αποφευχθεί η μείωση της απόδοσης, πολλοί ερευνητές προτιμούν να χρησιμοποιούν μείωση διαστάσεων για να μειώσουν τον χρόνο και την πολυπλοκότητα της μνήμης για τις εφαρμογές τους. Η χρήση της μείωσης διαστάσεων για την προεπεξεργασία θα

μπορούσε να είναι πιο αποτελεσματική σε σχέση με την ανάπτυξη φθηνών ταξινομητών.

Οι πιο κοινές τεχνικές μείωσης διαστάσεων, είναι η Ανάλυση Κύριων Συστατικών (Principal Component Analysis-PCA), η Ανάλυση Γραμμικών Διακρίσεων (Linear Discriminant Analysis-LDA) και η μη αρνητική παραγοντοποίηση μήτρας (non-negative matrix factorization-NMF). Επίσης υπάρχουν νέες τεχνικές για τη μείωση της διαστατικότητας εξαγωγής χαρακτηριστικών, όπως η τυχαία προβολή, οι αυτοκωδικοποιητές και η ενσωματωμένη στοχαστική ενσωμάτωση t-SNE (t-distributed stochastic neighbor embedding, t-SNE).

(III) **Τεχνικές ταξινόμησης**: Το πιο σημαντικό βήμα του αγωγού ταξινόμησης κειμένου είναι η επιλογή του καλύτερου ταξινομητή. Χωρίς πλήρη εννοιολογική κατανόηση κάθε αλγορίθμου, δεν είναι δυνατός ο κατάλληλος προσδιορισμός του πιο αποτελεσματικού μοντέλου για μια εφαρμογή ταξινόμησης κειμένου. Αρχικά καλύπτονται οι παραδοσιακές μεθόδουι ταξινόμησης κειμένου, όπως η ταξινόμηση Rocchio. Στη συνέχεια, παρουσιάζονται οι τεχνικές μάθησης που βασίζονται σε σύνολο όπως το boosting και το bagging, οι οποίες έχουν χρησιμοποιηθεί κυρίως για στρατηγικές εκμάθησης ερωτημάτων και ανάλυση κειμένου (Kim et al., 2000; Schapire & Singer, 2000). Ένας από τους απλούστερους αλγόριθμους ταξινόμησης είναι η λογιστική παλινδρόμηση (logistic regression-LR) που έχει χρησιμοποιηθεί στους περισσότερους τομείς εξαγωγής δεδομένων (Dou et al., 2018; Chen et al., 2017). Η πρώτη δημοφιλής εφικτή εφαρμογή στην ιστορία της ανάκτησης πληροφοριών ήταν, η The Na Theve Bayes Classifier (NBC). Παρουσιάζεται μία σύντομη επισκόπηση του Naïve Bayes Classifier που είναι υπολογιστικά φθηνή και απαιτεί πολύ χαμηλή ποσότητα μνήμης (Larson, 2010).

Οι μη παραμετρικές τεχνικές έχουν μελετηθεί και χρησιμοποιηθεί ως εργασίες ταξινόμησης, όπως k-πλησιέστερος γείτονας (KNN) (Li et al., 2001). Η Μηχανή υποστήριξης φορέα (Support Vector Machine-SVM) (Manevitz & Yousef, 2001; Han & Karypis, 2000) απποτελεί μια άλλη δημοφιλή τεχνική που χρησιμοποιεί έναν διακριτικό ταξινομητή για την κατηγοριοποίηση εγγράφων. Αυτή η τεχνική μπορεί επίσης να χρησιμοποιηθεί σε όλους τους τομείς της εξαγωγής δεδομένων, όπως βιοπληροφορική, εικόνα, βίντεο, ταξινόμηση ανθρώπινων δραστηριοτήτων, ασφάλεια και προστασία κ.λπ. Αυτό το μοντέλο χρησιμοποιείται από πολλούς ερευνητές ως βάση

σύγκρισης με τα δικά τους έργα προκειμένου να επισημάνουν καινοτομία και συνεισφορές.

Ταξινομητές με βάση τα δέντρα όπως το δέντρο αποφάσεων και το Random Forest έχουν επίσης μελετηθεί σε σχέση με την κατηγοριοποίηση εγγράφων (Xu et al., 2012). Τα τελευταία χρόνια, οι γραφικές ταξινομήσεις θεωρήθηκαν (Lafferty et al., 2001) ως εργασία ταξινόμησης όπως υπό συνθήκη τυχαία πεδία (conditional random fields-CRFs). Ωστόσο, αυτές οι τεχνικές χρησιμοποιούνται ως επί το πλείστον για την περίληψη εγγράφων (Shen et al., 2007) και την αυτόματη εξαγωγή λέξεων-κλειδιών (Zhang, 2008).

Πρόσφατα, οι προσεγγίσεις βαθιάς μάθησης έχουν επιτύχει εξαιρετικά αποτελέσματα σε σύγκριση με προηγούμενους αλγόριθμους μηχανικής μάθησης σε εργασίες όπως ταξινόμηση εικόνας, επεξεργασία φυσικής γλώσσας, αναγνώριση προσώπου κ.λπ. Η επιτυχία των αλγορίθμων βαθιάς μάθησης βασίζεται στην ικανότητά τους να μοντελοποιούν σύνθετες και μη γραμμικές σχέσεις μεταξύ δεδομένων (LeCun et al., 2015).

(IV) **Αξιολόγηση**: Το τελευταίο μέρος του αγωγού ταξινόμησης κειμένου είναι η αξιολόγηση. Η κατανόηση της απόδοσης ενός μοντέλου είναι απαραίτητη για τη χρήση και την ανάπτυξη μεθόδων ταξινόμησης κειμένου. Υπάρχουν πολλές διαθέσιμες μέθοδοι για την αξιολόγηση εποπτευόμενων τεχνικών. Ο υπολογισμός ακρίβειας είναι η απλούστερη μέθοδος αξιολόγησης αλλά δεν λειτουργεί για μη ισορροπημένα σύνολα δεδομένων (Huang & Ling, 2005). Ρεριγράφονται οι ακόλουθες μεθόδοι αξιολόγησης για αλγόριθμους ταξινόμησης κειμένου: Βαθμολογία Fβ (Lock, 2002), Συντελεστής συσχέτισης Matthews (Matthews Correlation Coefficient-MCC) (Matthews, 1975), χαρακτηριστικά λειτουργίας δέκτη (receiver operating characteristics-ROC)(Hanley & McNeil, 1982) και περιοχή κάτω από την καμπύλη ROC (area under the ROC curve-AUC) (Pencina et al., 2008).

1.1. Προεπεξεργασία Κειμένου

Η εξαγωγή χαρακτηριστικών και η προεπεξεργασία είναι κρίσιμα βήματα για εφαρμογές ταξινόμησης κειμένου. Παρουσιάζονται μέθοδοι για τον καθαρισμό συνόλων δεδομένων κειμένου, απομακρύνοντας έτσι τον σιωπηρό θόρυβο και επιτρέποντας την ενημερωτική βελτίωση. Επιπλέον, αναπτύσσονται δύο κοινές

μέθοδοι εξαγωγής χαρακτηριστικών κειμένου: Τεχνικές σταθμισμένων λέξεων και τεχνικές ενσωμάτωσης λέξεων.

1.1.1. Καθαρισμός και Προεπεξεργασία Κειμένου

Τα περισσότερα σύνολα δεδομένων κειμένου και εγγράφων περιέχουν πολλές περιττές λέξεις, όπως λέξεις-κλειδιά, ορθογραφικά λάθη, αργκό κ.λπ. Σε πολλούς αλγόριθμους, ειδικά στατιστικούς και πιθανοτικούς αλγόριθμους μάθησης, ο θόρυβος και οι περιττές δυνατότητες μπορεί να έχουν δυσμενείς επιπτώσεις στην απόδοση του συστήματος. Αναλύονται εν συντομία μερικές τεχνικές και μέθοδοι για τον καθαρισμό κειμένων και την προεπεξεργασία συνόλων δεδομένων κειμένου.

1.1.2. Συμβολισμός

Ο Συμβολισμός είναι μια μέθοδος προεπεξεργασίας που διασπά μια ροή κειμένου σε λέξεις, φράσεις, σύμβολα ή άλλα σημαντικά στοιχεία που ονομάζονται **σύμβολα** (Gupta & Malhotra, 2015). Ο κύριος στόχος αυτού του βήματος είναι η διερεύνηση των λέξεων σε μια πρόταση (Verma et al., 2014). Τόσο η ταξινόμηση όσο και η εξαγωγή κειμένου απαιτούν ένα πρόγραμμα ανάλυσης που επεξεργάζεται την κωδικοποίηση των εγγράφων, για παράδειγμα: πρόταση (Aggarwal, 2018):

After sleeping for four hours, he decided to sleep for another four.

Σε αυτήν την περίπτωση τα σύμβολα είναι όπως παρακάτω:

{"After" "sleeping" "for" "four" "hours" "he" "decided" "to" "sleep" "for" "another" "four"}.

1.1.3. Διακοπή λέξεων

Η ταξινόμηση κειμένου και εγγράφων περιλαμβάνει πολλές λέξεις που δεν είναι τόσο σημαντικές ώστε να χρησιμοποιούνται σε αλγόριθμους ταξινόμησης, όπως {"a", "about", "above", "cross", "after", "after after", "again",}. Η πιο συνηθισμένη τεχνική αντιμετώπισης αυτών των λέξεων είναι η αφαίρεσή τους από τα κείμενα και τα έγγραφα (Saif et al., 2014).

1.1.4. Κεφαλοποίηση

Τα σημεία δεδομένων κειμένου και εγγράφων έχουν ποικιλία κεφαλαίων για να σχηματίσουν μια πρόταση. Δεδομένου ότι τα έγγραφα αποτελούνται από πολλές προτάσεις, η διαφορετική χρήση κεφαλαίων μπορεί να είναι εξαιρετικά προβληματική

κατά την ταξινόμηση μεγάλων εγγράφων. Η πιο συνηθισμένη προσέγγιση για την αντιμετώπιση της ασυνεπούς κεφαλαιοποίησης είναι η μετατροπή κάθε γράμματος σε πεζό. Αυτή η τεχνική προβάλλει όλες τις λέξεις σε κείμενο και έγγραφο στον ίδιο χώρο χαρακτηριστικών, προκαλεί όμως ένα σημαντικό πρόβλημα στην ερμηνεία ορισμένων λέξεων π.χ. "US" (Ηνωμένες Πολιτείες της Αμερικής) σε "us" (αντωνυμία) (Gupta & Lehal, 2009). Οι μετατροπείς αργκό και συντομογραφίας μπορούν να βοηθήσουν να ληφθούν υπόψη αυτές οι εξαιρέσεις (Dalal & Zaveri, 2011).

1.1.5. Συντομογραφίες

Η αργκό και η συντομογραφία είναι άλλες μορφές ανωμαλιών κειμένου που αντιμετωπίζονται στο στάδιο προεπεξεργασίας. Ωσ συντομογραφία (Whitney & Evans, 2010) ορίζεται μια συντομευμένη μορφή μιας λέξης ή φράσης που περιέχει κυρίως τα πρώτα γράμματα από τις λέξεις, όπως το SVM που σημαίνει Support Vector Machine.

Η αργκό είναι ένα υποσύνολο της γλώσσας που χρησιμοποιείται σε ανεπίσημη ομιλία ή κείμενο και μπορεί έχει διαφορετική έννοια από την επίσημη γλώσσα (Helm, 2003). Μια κοινή μέθοδος αντιμετώπισης αυτών των λέξεων είναι η μετατροπή τους στην επίσημη γλώσσα (Dhuliawala et al., 2016)

1.1.6. Αφαίρεση θορύβου

Τα περισσότερα σύνολα δεδομένων κειμένου και εγγράφων περιέχουν πολλούς περιττούς χαρακτήρες όπως σημεία στίξης και ειδικούς χαρακτήρες. Τα κρίσιμα σημεία στίξης και οι ειδικοί χαρακτήρες είναι σημαντικά για την κατανόηση των εγγράφων από τον άνθρωπο, αλλά μπορεί να είναι επιζήμια για τους αλγόριθμους ταξινόμησης (Pahwa et al., 2018).

1.1.7. Διόρθωση Ορθογραφίας

Η ορθογραφία είναι ένα προαιρετικό βήμα προεπεξεργασίας. Τα τυπογραφικά λάθη είναι συνήθως παρόντα σε κείμενα και έγγραφα, ειδικά σε σύνολα δεδομένων κειμένου κοινωνικών μέσων (π.χ. Twitter). Πολλοί αλγόριθμοι, τεχνικές και μέθοδοι έχουν αντιμετωπίσει αυτό το πρόβλημα στο NLP (Mawardi et al., 2018). Πολλές τεχνικές και μέθοδοι είναι διαθέσιμες για τους ερευνητές, συμπεριλαμβανομένων των τεχνικών διόρθωσης ορθογραφίας που βασίζονται σε κατακερματισμούς και συμφραζόμενα (Dziadek et al., 2017), καθώς και διόρθωση ορθογραφίας χρησιμοποιώντας Trie και Damerau – Levenshtein απόσταση bigram.

1.1.8. Εύρεση του Θέματος μίας Λέξης (Stemming)

Στο NLP, μια λέξη θα μπορούσε να εμφανιστεί σε διαφορετικές μορφές (π.χ., ενικό και πληθυντικό) ενώ η σημασιολογική έννοια κάθε μορφής είναι η ίδια (Spirovski et al., 2018). Το stemming αποτελεί μια μέθοδο για την ενοποίηση διαφορετικών μορφών μιας λέξης στον ίδιο χώρο χαρακτηριστικών. Το κείμενο που προέρχεται τροποποιεί τις λέξεις για να αποκτήσει διάφορες μορφές λέξεων χρησιμοποιώντας διαφορετικές γλωσσικές διαδικασίες, όπως επίθεση (προσθήκη επιθεμάτων)(Singh & Gupta, 2016; Sampson, 2005). Για παράδειγμα, το θέμα (stem) της λέξης «studying» είναι «study».

1.1.9. Δημιουργία Λήμματος

Πρόκειται για μια διαδικασία NLP που αντικαθιστά το επίθεμα μιας λέξης με μια διαφορετική ή αφαιρεί το επίθεμα μιας λέξης εντελώς για να πάρει τη βασική μορφή λέξης (λήμμα) (Sampson, 2005).

1.2. Τεχνικές Εξαγωγής Χαρακτηριστικών Κειμένου

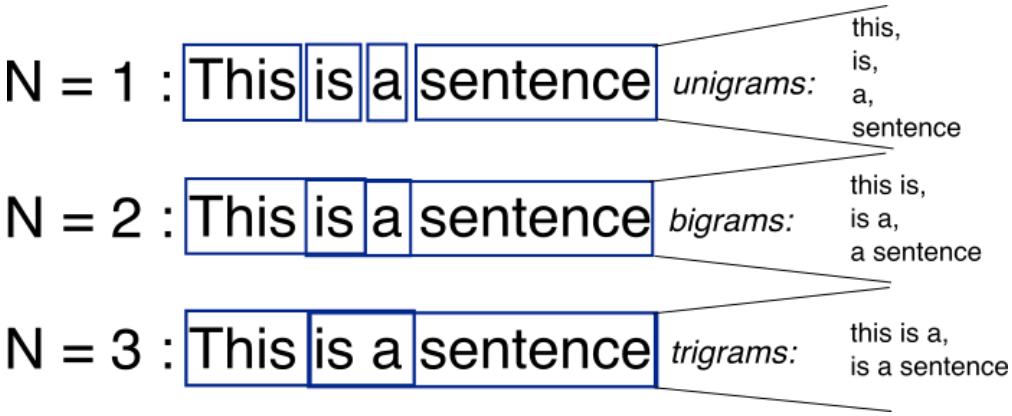
1.2.1. Συντακτική Αναπαράσταση Λέξεων

Πολλοί ερευνητές έχουν εργαστεί σε αυτήν την τεχνική εξαγωγής χαρακτηριστικών κειμένου για την επίλυση της χαμένης συντακτικής και σημασιολογικής σχέσης μεταξύ των λέξεων. Αναπτύχθηκαν νέες τεχνικές για την επίλυση αυτού του προβλήματος, αλλά πολλές από αυτές εξακολουθούν να έχουν περιορισμούς. Οι (Caropreso & Matwin, 2006) εισήγαγαν ένα μοντέλο στο οποίο η χρησιμότητα της συντακτικής και σημασιολογικής γνώσης στην αναπαράσταση κειμένου για την επιλογή των προτάσεων προέρχεται από τεχνικά γονιδιωματικά κείμενα. Η άλλη λύση για συντακτικό πρόβλημα είναι η χρήση της τεχνικής n-gram για εξαγωγή χαρακτηριστικών.

1.2.2. N-Gram

Η τεχνική n-gram είναι ένα σύνολο n-λέξεων που εμφανίζεται «με αυτή τη σειρά» σε ένα σύνολο κειμένων. Δεν πρόκειται για αναπαράσταση ενός κειμένου, αλλά θα μπορούσε να χρησιμοποιηθεί ως χαρακτηριστικό για την αναπαράσταση ενός

κειμένου.



Το BOW είναι μια αναπαράσταση ενός κειμένου χρησιμοποιώντας τις λέξεις του (1-gram) που χάνει τη σειρά τους (συντακτική). Αυτό το μοντέλο είναι πολύ εύκολο να ληφθεί και το κείμενο μπορεί να αναπαρασταθεί μέσω ενός διανύσματος, γενικά ενός διαχειρίσιμου μεγέθους του κειμένου. Από την άλλη πλευρά, το n-gram είναι ένα χαρακτηριστικό του BOW για την αναπαράσταση ενός κειμένου που χρησιμοποιεί 1-gram. Είναι πολύ συνηθισμένο να χρησιμοποιούνται 2-gram ή και 3-gram. Με αυτόν τον τρόπο, η εξαγόμενη δυνατότητα κειμένου μπορεί να ανιχνεύσει περισσότερες πληροφορίες σε σύγκριση με το 1-gram.

Η πολυωνυμική (ή πολύτιμη) υλικοτεχνική ταξινόμηση (Krishnapuram et al., 2005) χρησιμοποιεί την πιθανότητα του x να ανήκει στην τάξη i (όπως ορίζεται στην ακόλουθη εξίσωση:

$$p\left(y^{(i)} = 1 \mid x, \theta\right) = \frac{\exp\left(\theta^{(i)^T} x\right)}{\sum_{j=1}^{m} \exp\left(\theta^{(i)^T} x\right)}$$

όπου θ (i) είναι ο φορέας βάρους που αντιστοιχεί στην τάξη i.

Για δυαδική ταξινόμηση (m = 2) η οποία είναι γνωστή ως βασική LR, αλλά για πολυωνυμική λογιστική παλινδρόμηση (m> 2) συνήθως χρησιμοποιείται η συνάρτηση softmax.

Η συνάρτηση κανονικοποίησης είναι:

$$\sum_{i=1}^{m} p\left(y^{(i)} = 1 \mid x, \theta\right) = 1$$

Σε μια εργασία ταξινόμησης ως εποπτευόμενο πλαίσιο μάθησης, το στοιχείο θ υπολογίζεται από το υποσύνολο των δεδομένων εκπαίδευσης D που ανήκει στην τάξη i όπου $i \in \{1, \dots, n\}$

Η ταξινόμηση κειμένου Naïve Bayes χρησιμοποιείται ευρέως για εργασίες κατηγοριοποίησης εγγράφων από τη δεκαετία του 1950 (Porter, 1980). Η συγκεκριμένη μέθοδος ταξινόμησης βασίζεται στο θεώρημα Bayes, το οποίο διατυπώθηκε από τον Thomas Bayes μεταξύ 1701–1761 (Hill, 1968). Πρόσφατες μελέτες έχουν χρησιμοποιήσει ευρέως αυτήν την τεχνική στην ανάκτηση πληροφοριών (Qu et al., 2018). Πρόκειται για ένα γενετικό μοντέλο και αποτελεί την πιο παραδοσιακή μέθοδο κατηγοριοποίησης κειμένου. Ακολουθείται η πιο βασική έκδοση του NBC που αναπτύχθηκε χρησιμοποιώντας TF (bag-of-word), μια τεχνική εξαγωγής χαρακτηριστικών που μετρά τον αριθμό των λέξεων στα έγγραφα.

1.2.3. Περιγραφή Υψηλού Επιπέδου του Ταξινομητή Naïve Bayes

Εάν (n) ο αριθμός των εγγράφων που ταιριάζει σε k κατηγορίες όπου k ∈ {c_1, c_2, ..., c_k}, η προβλεπόμενη τάξη ως έξοδος είναι c ∈ C. Ο αλγόριθμος Naïve Bayes μπορεί να περιγραφεί ως εξής:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Όπου d είναι το έγγραφο και c οι τάξεις.

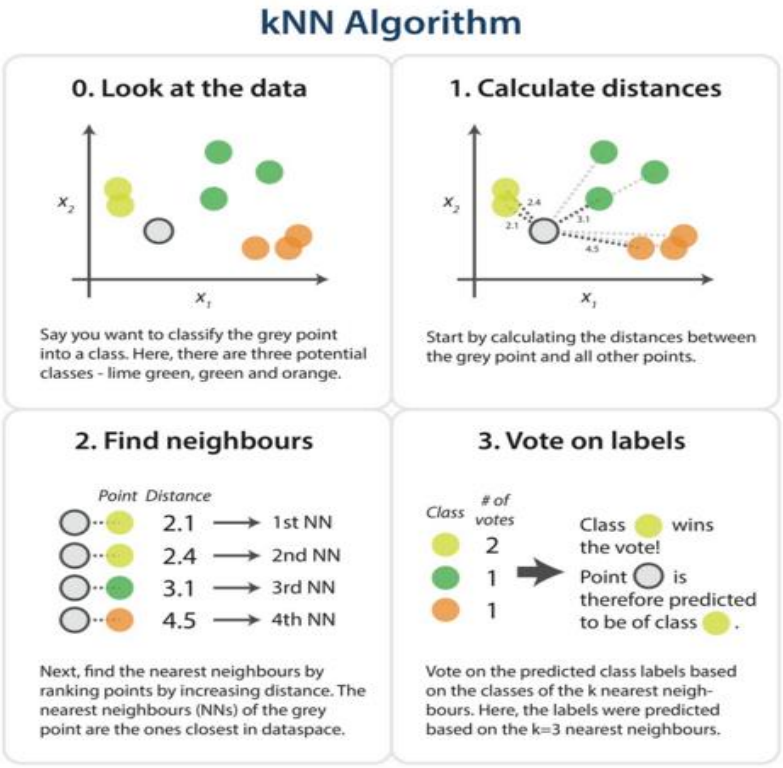
$$\begin{aligned} C_{MAP} &= \arg \max_{c \in C} P(d \mid c)P(c) \\ &= \arg \max_{c \in C} P(x_1, x_2, \dots, x_n \mid c)p(c) \end{aligned}$$

Αυτό το μοντέλο χρησιμοποιείται ως βασική γραμμή πολλών εγγράφων που είναι επίπεδο λέξης του ταξινομητή Naïve Bayes (Kim et al., 2006) ως εξής:

$$P(c_j \mid d_i; \hat{\theta}) = \frac{P(c_j \mid \hat{\theta}) P(d_i \mid c_j; \hat{\theta}_j)}{P(d_i \mid \hat{\theta})}$$

1.2.4. k-nearest Neighbors

Ο αλγόριθμος k-nearest Neighbors (KNN) είναι μια μη παραμετρική τεχνική που χρησιμοποιείται για την ταξινόμηση. Αυτή η μέθοδος χρησιμοποιείται για εφαρμογές ταξινόμησης κειμένου σε πολλούς ερευνητικούς τομείς (Jiang et al., 2012) τις τελευταίες δεκαετίες.



Δεδομένου ενός δοκιμαστικού εγγράφου x, ο αλγόριθμος KNN βρίσκει τους k πλησιέστερους γείτονες του x μεταξύ όλων των εγγράφων στο προς επεξεργασία σύνολο και βαθμολογεί τους υποψηφίους κατηγορίας με βάση την τάξη των k-γειτόνων. Η ομοιότητα του x και του εγγράφου κάθε γείτονα θα μπορούσε να είναι η βαθμολογία της κατηγορίας των γειτονικών εγγράφων. Πολλά έγγραφα KNN

ενδέχεται να ανήκουν στην ίδια κατηγορία. Σε αυτήν την περίπτωση, η άθροιση αυτών των βαθμολογιών θα είναι η βαθμολογία ομοιότητας της τάξης k σε σχέση με το έγγραφο δοκιμής x. Μετά την ταξινόμηση των τιμών βαθμολογίας, ο αλγόριθμος εκχωρεί τον υποψήφιο στην τάξη με την υψηλότερη βαθμολογία από το δοκιμαστικό έγγραφο x (Jiang et al., 2012). Ο κανόνας απόφασης του KNN είναι:

$$f(x) = \arg\max_j S(x, C_j) = \sum_{d_i \in KNN} \text{sim}\,(x, d_i) y(d_i, C_j)$$

Όπου το S αναφέρεται στην τιμή βαθμολογίας σε σχέση με το S (x, C_j), η τιμή βαθμολογίας του υποψηφίου i στην κλάση j, και η έξοδος του f (x) είναι μια ετικέτα στο έγγραφο δοκιμαστικού συνόλου.

Η αρχική έκδοση της SVM αναπτύχθηκε από τους (Vapnik & Chervonenkis, 1964). Οι (Boser et al., 1992) προσάρμοσαν αυτήν την έκδοση σε μια μη γραμμική διατύπωση στις αρχές της δεκαετίας του 1990. Η SVM σχεδιάστηκε αρχικά για εργασίες δυαδικής ταξινόμησης. Ωστόσο, πολλοί ερευνητές εργάζονται σε προβλήματα πολλαπλών τάξεων χρησιμοποιώντας αυτήν την κυρίαρχη τεχνική (Bo & Xianwu, 2006).

1.2.5. Δυαδική τάξη SVM

Στο πλαίσιο της ταξινόμησης κειμένου, έστω x_1, x_2, ..., x_l είναι παραδείγματα κατάρτισης που ανήκουν σε μια κατηγορία X, όπου το X είναι ένα συμπαγές υποσύνολο του R^N (Manevitz & Yousef, 2001). Ένας δυαδικός ταξινομητής μπορεί να διατυπωθεί ως εξής:

$$min \frac{1}{2} \|w\|^2 + \frac{1}{vl} \sum_{i=1}^{l} \xi_i - p$$

Που υπακούει σε:

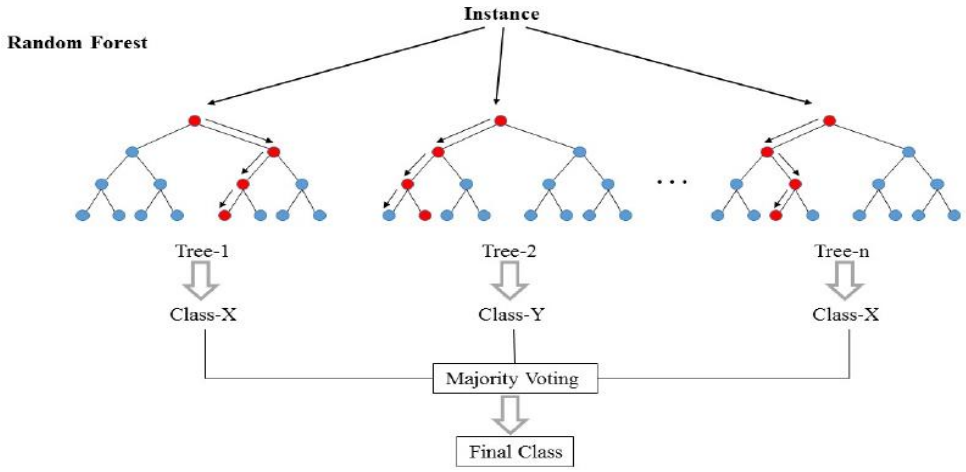
$\left(w.\Phi(x_i)\right) \geq p - \xi_i$ όπου i=1,2,…l και ξ≥0

Εάν w και p επιλύουν αυτό το πρόβλημα, τότε η συνάρτηση απόφασης δίνεται από:

$$f(x) = sign(\big(w.\Phi(x)\big) - p)$$

1.2.6. Random Forest

Το Random Forest ή η τεχνική αποφάσεων Random Forest αποτελεί μέθοδο μάθησης για την ταξινόμηση κειμένου. Αυτή η μέθοδος, η οποία χρησιμοποίησε το δέντρο t ως παράλληλο, εισήχθη από τον (Ho, 1995) το 1995. Η κύρια ιδέα του RF (Random forests-RF) είναι η δημιουργία τυχαίων δέντρων απόφασης.



Αυτή η τεχνική αναπτύχθηκε περαιτέρω το 1999 από τον (Breiman, 1999), ο οποίος διαπίστωσε τη σύγκλιση για το RF ως μετρήσεις περιθωρίου (mg (X, Y)) ως εξής:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j)$$

όπου I (.) αναφέρεται στη συνάρτηση δείκτη.

Μετά την εκπαίδευση όλων των δέντρων ως δάσος, οι προβλέψεις εκχωρούνται ως εξής (Wu et al., 2004):

$$\delta_V = \arg \max_i \sum_{j:j \neq j} I_{\{r_{ij} > r_{ji}\}}$$

Με:

$$r_{ij} + r_{ji} = 1$$

1.2.7. Επαναλαμβανόμενο νευρωνικό δίκτυο (RNN)

Μια άλλη αρχιτεκτονική νευρωνικών δικτύων που έχει χρησιμοποιηθεί για την εξαγωγή κειμένων και την ταξινόμηση είναι το επαναλαμβανόμενο νευρωνικό δίκτυο (RNN) (Sutskever et al., 2011). Το RNN εκχωρεί περισσότερα βάρη στα προηγούμενα σημεία δεδομένων μιας ακολουθίας. Επομένως, αυτή η τεχνική είναι μια ισχυρή μέθοδος για την ταξινόμηση κειμένου, συμβολοσειράς και διαδοχικών δεδομένων. Ένα RNN εξετάζει τις πληροφορίες προηγούμενων κόμβων με μια πολύ εξελιγμένη μέθοδο που επιτρέπει καλύτερη σημασιολογική ανάλυση της δομής του συνόλου δεδομένων. Το RNN λειτουργεί κυρίως χρησιμοποιώντας LSTM ή GRU για ταξινόμηση κειμένου και περιέχει στρώμα εισόδου (ενσωμάτωση λέξεων), κρυμμένα στρώματα και τέλος στρώμα εξόδου. Αυτή η μέθοδος διατυπώνεται ως:

$$x_t = F(x_{t-1}, u_t, \theta)$$

Όπου x_t είναι η κατάσταση τη χρονική στιγμή t και u_t είναι η είσοδος για το χρονικό βήμα t. Πιο συγκεκριμένα, μπορούν να χρησιμοποιηθούν βάρη για να τη διαμόρφωση της παραπάνω εξίσωσης, παραμετροποιημένης από:

$$x_t = W_{rec}\sigma(x_{t-1}) + W_{in}u_t + b$$

όπου το W_{rec} αναφέρεται σε επαναλαμβανόμενο βάρος μήτρας, το W_{in} αναφέρεται σε βάρη εισόδου, το b είναι η μεροληψία και το σ σημαίνει μια συνάρτηση με στοιχεία.

Παρά τα οφέλη που περιγράφονται παραπάνω, το RNN είναι ευάλωτο στα προβλήματα της διαβάθμισης και της έκρηξης, όταν το σφάλμα του αλγόριθμου διαβάθμισης κλίσης διαδίδεται ξανά μέσω του δικτύου (Bengio et al., 1994).

1.2.8. Ημι-εποπτευόμενη μάθηση για ταξινόμηση κειμένου

Πολλοί ερευνητές έχουν αναπτύξει αποτελεσματικούς ταξινομητές τόσο για επισημασμένα όσο και για μη επισημασμένα έγγραφα. Η ημι-εποπτευόμενη μάθηση είναι ένας τύπος εποπτευόμενου μαθησιακού προβλήματος που χρησιμοποιεί δεδομένα χωρίς ετικέτα για την εκπαίδευση ενός μοντέλου. Συνήθως, ερευνητές και επιστήμονες προτιμούν να χρησιμοποιούν ημι-εποπτευόμενες τεχνικές όταν ένα μικρό μέρος του συνόλου δεδομένων περιέχει επισημασμένα σημεία δεδομένων και μια μεγάλη

ποσότητα συνόλου δεδομένων δεν περιλαμβάνει ετικέτες (Gowda et al., 2016). Οι περισσότεροι από τους ημι-εποπτευόμενους αλγόριθμους μάθησης για εργασίες ταξινόμησης χρησιμοποιούν μια τεχνική ομαδοποίησης (συνήθως χρησιμοποιείται για μη εποπτευόμενη μάθηση) (Kowsari, 2014) ως εξής: Αρχικά εφαρμόζεται μια τεχνική ομαδοποίησης στο D^T με K = K (ο αριθμός των τάξεων), καθώς η D^T έχει επισημασμένα δείγματα όλων των τάξεων K (Gowda et al., 2016). Εάν ένα τμήμα P_i έχει επισημάνει δείγματα, τότε όλα τα σημεία δεδομένων σε αυτό το σύμπλεγμα ανήκουν σε αυτήν την ετικέτα.

Ο ερευνητικός στόχος για τις τεχνικές ομαδοποίησης είναι να προσδιοριστεί εάν υπάρχουν περισσότερες από μία τάξεις επισημασμένες σε ένα σύμπλεγμα και φυσικά τί θα συμβεί εάν δεν υπάρχει σημειωμένο σημείο δεδομένων σε ένα σύμπλεγμα (Kowsari et al., 2015). Περιγράφεται εν συντομία η πιο δημοφιλής τεχνική ημι-εποπτευόμενης ταξινόμησης κειμένου και εγγράφων. Οι (Chapelle & Zien, 2005) εργάστηκαν σε ημι-εποπτευόμενη ταξινόμηση μέσω διαχωρισμού χαμηλής πυκνότητας, ο οποίος συνδυάζει τον υπολογισμό της απόστασης γραφήματος με την κατάρτιση μηχανημάτων φορέα μεταγωγής (transductive support vector machine-TSVM). Οι (Nigam et al., 2006) ανέπτυξαν μια τεχνική για ταξινόμηση κειμένου χρησιμοποιώντας μεγιστοποίηση προσδοκίας (expectation maximization-EM) και γενετικά μοντέλα για ημι-εποπτευόμενη μάθηση με επισημασμένα και μη επισημασμένα δεδομένα στον τομέα των ταξινομήσεων κειμένου. Οι (Shi et al., 2010) εισήγαγαν μια μέθοδο για τη μεταφορά γνώσης ταξινόμησης σε γλώσσες μέσω μεταφρασμένων χαρακτηριστικών. Αυτή η τεχνική χρησιμοποιεί έναν αλγόριθμο EM που φυσικά λαμβάνει υπόψη την ασάφεια που σχετίζεται με τη μετάφραση μιας λέξης. Επίσης εισήγαγαν την «Ημι-εποπτευόμενη Εκτίμηση Συχνότητας (SFE)», μια μέθοδο MNBC για ταξινόμηση κειμένου μεγάλης κλίμακας. Οι (Zhou et al., 2014) ανέπτυξαν μια νέα μέθοδο βαθιάς μάθησης που χρησιμοποιεί ασαφή DBN για ταξινόμηση ημι-εποπτευόμενων συναισθημάτων. Αυτή η μέθοδος χρησιμοποιεί μια ασαφή συνάρτηση συμμετοχής για κάθε κατηγορία κριτικών με βάση τη γνώση της αρχιτεκτονικής.

2. Βιβλιογραφική Ανασκόπηση Μεθόδων Μηχανικής Μάθησης για Ταξινόμηση και Παλινδρόμηση

2.1. Εξόρυξη Εκπαιδευτικών Δεδομένων και Μηχανική Μάθηση με Αλγορίθμους Ταξινόμησης

Πριν από πενήντα χρόνια υπήρχαν μόνο λίγα πανεπιστήμια σε όλο τον κόσμο που θα μπορούσαν να προσφέρουν εξειδικευμένα εκπαιδευτικά μαθήματα. Σήμερα τα πανεπιστήμια παράγουν όχι μόνο πτυχιούχους αλλά και τεράστιες ποσότητες δεδομένων από τα συστήματά τους. Έτσι, το ερώτημα που προκύπτει είναι πώς μπορεί ένα ανώτατο εκπαιδευτικό ίδρυμα να αξιοποιήσει τη δύναμη αυτών των διδακτικών δεδομένων για τη στρατηγική του χρήση; Η οικοδόμηση ενός Πληροφοριακού συστήματος που μπορεί να μάθει από τα δεδομένα είναι μια δύσκολη εργασία, αλλά έχει επιτευχθεί με επιτυχία χρησιμοποιώντας διάφορες προσεγγίσεις εξόρυξης δεδομένων όπως ταξινόμηση, ταξινόμηση, αλγόριθμοι πρόβλεψης κ.λπ. Ωστόσο, η χρήση αυτών των αλγορίθμων με εκπαιδευτικό σύνολο δεδομένων είναι αρκετά χαμηλή. Η εργασία των (Dutt et al., 2015) επικεντρώνεται στην ενοποίηση των διαφορετικών τύπων αλγορίθμων ταξινόμησης όπως εφαρμόζονται στο πλαίσιο "Εξόρυξης Εκπαιδευτικών Δεδομένων (Educational Data Mining – EDM)".

Σύμφωνα με τη διεθνή κοινοπραξία για την εξόρυξη εκπαιδευτικών δεδομένων, το EDM ορίζεται ως «ένας αναδυόμενος επιστημονικός κλάδος που ασχολείται με την ανάπτυξη μεθόδων για την εξερεύνηση των μοναδικών τύπων δεδομένων που προέρχονται από εκπαιδευτικά περιβάλλοντα και τη χρήση αυτών των μεθόδων για την καλύτερη κατανόηση των μαθητών και των χώρων που αυτοί μαθαίνουν».

Το EDM εστιάζει στην ανάλυση δεδομένων που δημιουργούνται σε μια εκπαιδευτική εγκατάσταση από τα διάφορα διασυνδεδεμένα ή ανομοιογενή συστήματα για την ανάπτυξη μοντέλου για τη βελτίωση της μαθησιακής εμπειρίας και της ιδρυματικής αποτελεσματικότητας. Η εξόρυξη δεδομένων, αναφέρεται επίσης μερικές φορές ως "Ανακάλυψη Γνώσεων σε Βάσεις Δεδομένων (Knowledge Discovery in Databases - KDD)", είναι ένα γνωστό πεδίο σπουδών στις βιοεπιστήμες και το εμπόριο, αλλά η εφαρμογή της εξόρυξης δεδομένων στο εκπαιδευτικό πλαίσιο είναι περιορισμένη.

Το EDM μετατρέπει τα ανεπεξέργαστα δεδομένα που προέρχονται από εκπαιδευτικά συστήματα σε χρήσιμες πληροφορίες που θα μπορούσαν ενδεχομένως να έχουν μεγαλύτερο αντίκτυπο στην εκπαιδευτική έρευνα και πρακτική». Παραδοσιακά, οι

ερευνητές έχουν εφαρμόσει μεθόδους εξόρυξης δεδομένων όπως ταξινόμηση, ταξινόμηση, εξόρυξη κανόνων συσχέτισης, εξόρυξη κειμένων σε εκπαιδευτικό πλαίσιο όπως περιγράφεται συνοπτικά. Οι (Romero et al., 2007), διεξήγαγαν μια έρευνα που παρέχει μια περιεκτική πηγή άρθρων που δημοσιεύθηκαν μεταξύ 1995 και 2005 σχετικά με την Εκπαιδευτική Εξόρυξη Δεδομένων (EDM). Ο (Zaiane, 2001) έχει προτείνει την εφαρμογή τεχνικών εξόρυξης δεδομένων για τη μελέτη διαδικτυακών μαθημάτων. Ο (Zaíane, 2002) είχε προτείνει κανόνες συσχέτισης και ταξινόμηση για την υποστήριξη συνεργατικού φιλτραρίσματος για την ανάπτυξη πιο ευαίσθητων και αποτελεσματικών συστημάτων ηλεκτρονικής μάθησης.

Σε ένα μαθησιακό περιβάλλον οι μαθησιακοί τρόποι του μαθητή είναι καθοριστικοί παράγοντες. Σε πολλές περιπτώσεις υπήρξε αναντιστοιχία μεταξύ των προσωπικών τρόπων μάθησης και των μαθησιακών απαιτήσεων διαφορετικών επιστημονικών κλάδων. Οι (Salazar et al., 2004), έχουν χρησιμοποιήσει μια προσέγγιση ανάλυσης συστάδων δύο-βημάτων που εξέτασε τα κεντροειδή του εγκεφάλου και η οποία χρησιμοποίησε την τεχνολογία ηλεκτροεγκεφαλογραφίας (electroencephalography - EEG) για τη μέτρηση του τρόπου μάθησης των συμμετεχόντων ώστε να ήταν σε θέση να τον ταξινομήσουν με επιτυχία σε 4 μοναδικές συστάδες. Οι μαθητές συνήθως σημειώνουν κείμενα κατά την ανάγνωση του βιβλίου επισημαίνοντας το πλαίσιο ενδιαφέροντος ή υπογραμμίζοντάς το ή γράφοντας σχόλια στα πλευρικά περιθώρια. Αυτή η δραστηριότητα ονομάζεται επισημείωση (annotation). Οι ερευνητές εφάρμοσαν τη μέθοδο στατιστικής ταξινόμησης όπως η ταξινόμηση K-means και η ιεραρχική ταξινόμηση στις σημειώσεις των μαθητών.

Και απέδειξαν ότι με τη χρήση αυτών των μεθόδων ταξινόμησης, η δημιουργία συστάδας με μαθητές που έχουν παρόμοιο τρόπο μάθησης βελτιώνεται και είναι ταχύτερη. Η κατανόηση γραπτού λόγου είναι μια πολύ ευρέως χρησιμοποιούμενη δραστηριότητα στην τάξη σε σχολεία και κολέγια. Αυτό βοηθά στην οικοδόμηση μιας δια βίου συνήθειας ανάγνωσης και διαδικασίας μάθησης. Αυτή η ικανότητα των συμπεριφοριστικών μαθησιακών προτύπων έχει χαρτογραφηθεί υπολογιστικά με την εφαρμογή της μεθόδου Forgy για την ταξινόμηση k-means και σε συνδυασμό με την ταξινομία του Bloom για τον προσδιορισμό θετικών και αρνητικών γνωστικών δεξιοτήτων που αναφέρονται σε δεξιότητες κατανόησης γραπτού λόγου. Ωστόσο, σε μια άλλη μελέτη, συνδυάστηκαν “Προγράμματα Διδασκαλίας Βασισμένα στο Διαδίκτυο (Web Based Instruction - WBI)” με το γνωστικό τρόπο μάθησης του

εκπαιδευόμενου για να μελετήσουν τις επιπτώσεις τους στα εκπαιδευτικά πρότυπα μάθησης. Ο αλγόριθμος ταξινόμησης K-means χρησιμοποιήθηκε για να οδηγήσει σε συστάδα μαθητών που μοιράζονταν παρόμοια πρότυπα μάθησης που καταλήγει περαιτέρω στην ταυτοποίηση του σχετικού γνωστικού τρόπου για κάθε ομάδα.

Το "Σύστημα Διαχείρισης Μάθησης (Learning Management System - LMS)" έχει γίνει αναπόσπαστο μέρος των εκπαιδευτικών ιδρυμάτων για τη διδασκαλία και τη μάθηση. Ένα τυπικό LMS καταγράφει τις περισσότερες από τις δραστηριότητες του χρήστη όπως δήλωση μαθημάτων, ανάγνωση διδακτικών ενοτήτων, δήλωση πρακτικής εξέτασης, βαθμολογία εξετάσεων, αλληλεπίδραση μαθητή-μαθητή μέσω καταγραφής συνομιλίας ή πινάκων συζήτησης, παρόμοια αλληλεπίδραση μαθητή-δασκάλου μέσω πινάκων συζήτησης καταγράφεται επίσης στο LMS. Έχουν διεξαχθεί αρκετές μελέτες ως προς αυτό.

Οι Moreno-Clari et al (2009) μελέτησαν τα στατιστικά χρήσης που παρέχει ένα LMS και εργάστηκε για την ανάλυση στατιστικών δεδομένων και τα αποτελέσματα εφαρμόστηκαν στο Πανεπιστήμιο της Βαλένθια (Ισπανία). Παρόλο που κατάφεραν να επιτύχουν στη στατιστική ανάλυση των δεδομένων χρήσης του LMS χρησιμοποιώντας το λογισμικό SPSS, αλλά η προτυποποίηση της μεθοδολογίας τους, η μετέπειτα διαδικασία αυτοματισμού δεν έχει ολοκληρωθεί και έχει παραμείνει ως μελλοντική εργασία. Η απόδοση σε εξετάσεις, τα στατιστικά χρήσης, η παλινδρόμηση, ο αριθμός επισκέψεων, οι κορυφαίοι όροι αναζήτησης, ο αριθμός λήψεων πόρων ηλεκτρονικής μάθησης (e-learning) παρουσιάζονται στην εργασία των (Valsamidis et al., 2012). Αρκετές προσεγγίσεις και τεχνικές DM (ταξινόμηση, ταξινόμηση και ανάλυση συσχέτισης) έχουν προταθεί για συνδυαστική χρήση στην εξόρυξη δεδομένων αξιολόγησης των μαθητών στο LMS. Κανόνες συσχέτισης, ταξινόμηση, ανάλυση διαδοχικών προτύπων, μοντελοποίηση εξάρτησης και πρόβλεψη έχουν χρησιμοποιηθεί για τη βελτίωση διαδικτυακών μαθησιακών περιβαλλόντων για να ενισχύσουν στη συνέχεια τον βαθμό στον οποίο ο εκπαιδευτικός μπορεί να αξιολογήσει τη μαθησιακή διαδικασία. Η ανάλυση της καταγραφής πρόσβασης χρήστη στο Moodle για τη βελτίωση της ηλεκτρονικής μάθησης και για την υποστήριξη της ανάλυσης των τάσεων παρουσιάζεται στο (Lahane et al., 2012). Η σύγκριση διαφορετικών αλγορίθμων DM γίνεται για την ταξινόμηση των μαθητών (πρόβλεψη τελικών βαθμών) με βάση τα δεδομένα χρήσης του Moodle. Η πρόβλεψη της απόδοσης των φοιτητών (τελικός βαθμός) με βάση τα χαρακτηριστικά που εξάγονται από τα καταγεγραμμένα δεδομένα

παρουσιάζεται στο (Jing & Shiying, 2010) και η απόδοση των ακαδημαϊκών φοιτητών πανεπιστημίου παρουσιάζεται στο (Stes & Petegem, 2014). Η πρόβλεψη των βαθμών διαδικτυακών μαθητών (χρησιμοποιώντας έναν ορθογώνιο αλγόριθμο εξαγωγής κανόνων βάσει αναζήτησης) παρουσιάζεται στο (Rashid et al., 2010).

Έχουν διεξαχθεί και άλλες μελέτες για την πρόβλεψη της απόδοσης του μαθητών από τις βαθμολογίες καταγραφής και δοκιμών σε διαδικτυακές διδασκαλίες (χρησιμοποιώντας παλινδρόμηση πολλαπλών μεταβλητών). Ενώ επίσης έχουν χρησιμοποιήσει ταξινόμηση, ταξινόμηση, εξόρυξη κανόνων συσχέτισης και παλινδρόμηση για την ανακάλυψη πιθανών εξαρτήσεων μεταξύ της μέσης απόδοσης του μαθητή και των χαρακτηριστικών του μαθήματος. Τα αποτελέσματά τους επιβεβαιώνουν ότι η συμπεριφορά των μαθητών σε μια διαδικτυακή πλατφόρμα μάθησης επηρεάζει την απόδοσή τους.

Σε μια άλλη μελέτη, οι ερευνητές έχουν δείξει πώς τα εκπαιδευτικά ιδρύματα μπορούν να επωφεληθούν από τα δεδομένα που συλλέγονται από το LMS. Έχουν προτείνει έναν αλγόριθμο που ονομάζεται «Αλγόριθμος ταξινόμησης μαθήματος» όταν εφαρμόζεται στο LMS (ανοιχτή πλατφόρμα e-Class) που χρησιμοποιεί το ίδρυμα για να προσδιορίσει και να δημιουργήσει την ποιότητα περιεχομένου των μαθημάτων και τις διαδικτυακές αναφορές χρήσης από τους μαθητές. Αυτές οι αναφορές αποστέλλονται στη συνέχεια στους εκπαιδευτές για σκοπούς αξιολόγησης και κινητοποίησης. Οι Sabitha & Mehrotra (2012) πρότειναν τη χρήση της ταξινόμησης k-means και αυτο-οργανούμενη αντιστοίχιση σε μαθησιακά αντικείμενα συστάδας (τα μαθησιακά αντικείμενα είναι μορφωτικοί πόροι όπως το eBook, το ερωτηματολόγιο, το ευρετήριο απαντήσεων κ.λπ.) έτσι ώστε να διευκολύνεται η ταχύτερη προσβασιμότητα τέτοιων πόρων μέσω αναζήτησης σε ένα LMS. Οι Govindarajan et al (2013) έχουν προτείνει ταξινόμηση βασισμένη στη "Βελτιστοποίηση Σμήνους Σωματιδίων (Particle Swarm Optimization - PSO)" για τη βελτίωση της ποιότητας της μάθησης ενσωματώνοντας το "Εξατομικευμένο Μαθησιακό Περιβάλλον (Personalized Learning Environment - PLE)" σε συνδυασμό με το συμβατικό Σύστημα Διαχείρισης Μάθησης (LMS).

Ωστόσο, ένα από τα σημαντικότερα προβλήματα που αντιμετωπίζουν οι ερευνητές στην εύρεση ενδιαφέρων προτύπων από το σύνολο εκπαιδευτικών δεδομένων είναι το σχετικά μικρό μέγεθος των δεδομένων.

Σε μια άλλη μελέτη έχει εφαρμοστεί ο αλγόριθμος ταξινόμησης "Μεγιστοποίησης Προσδοκίας (Expectation Maximization - EM)" για την ανακάλυψη προφίλ μαθητών από δεδομένα αξιολόγησης μαθημάτων και για την εύρεση συσχετίσεων μεταξύ θεμάτων με βάση την απόδοση των μαθητών.

Η απορρόφηση των αποφοίτων τους στην αγορά εργασίας ήταν πρωταρχικός στόχος των ιδρυμάτων τριτοβάθμιας εκπαίδευσης. Οι εργαζόμενοι της γνώσης καταφεύγουν σε βασικά εκπαιδευτικά μαθήματα χρησιμοποιώντας "Μαζικά Ανοικτά Διαδικτυακά Μαθήματα (Massive Open Online Courses - MOOCs)" που παρέχονται διαδικτυακά από ιδρύματα φημισμένα όπως το MIT, το Στάνφορντ και το Χάρβαρντ. Το έτος 2012 ήταν μάρτυρας μιας ταχείας ανάπτυξης και επέκτασης αρκετών μαζικών ανοικτών διαδικτυακών πλατφόρμων εκπαίδευσης (Massive Open Online Education Platforms - MOOEPs) όπως Canvas, ClassToGo, Coursera, edX, NPTEL, Udacity για να αναφερθούν μερικά. Ο Subbian (2013) είχε πραγματοποιήσει μια μελέτη για να διερευνήσει το πεδίο της διεπιστημονικής εκπαίδευσης μέσω MOOCs. Η εκπαίδευση απασχολησιμότητας είναι αναπόσπαστο στοιχείο της ανωτάτης εκπαίδευσης και ένας σημαντικός δρόμος με τον οποίο οι εταιρείες αποκτούν εξαιρετικούς υπαλλήλους. Είναι λοιπόν ένα βιώσιμο επιχείρημα ότι στην παρούσα κοινωνικοοικονομική ανάπτυξη, το εκπαιδευτικό περιεχόμενο που βασίζεται στην απορρόφηση από την αγορά εργασίας αποτελεί υποχρέωση.

2.2. Πρόβλεψη Κινδύνου με Μεθόδους Μηχανικής Μάθησης και Παλινδρόμησης

Η αλληλεπίδραση μεταξύ ιατρικών στατιστικών και επιδημιολογίας αφενός και "Τεχνικών Μηχανικής Μάθησης (Machine Learning Techniques - MLTs)" αφετέρου μπορεί να είναι πολύ διεγερτική (Kruppa et al., 2014). Η εκτίμηση πιθανότητας είναι το κλειδί για τον τομέα της πρόβλεψης κινδύνου, ο οποίος αυξάνεται σε σπουδαιότητα στην ιατρική, όπου η εξατομικευμένη ιατρική καθίσταται όλο και πιο δυνατή μέσω του συνδυασμού κλασικών προβλέψεων κινδύνου και βιοδεικτών.

Ένα ζήτημα προσοχής με τις MLTs είναι ότι έχουν διάφορες παραμέτρους συντονισμού. Αυτές περιλαμβάνουν τον αριθμό των γειτόνων που πρέπει να ληφθούν υπόψη στην μέθοδο "Πλησιέστερου Γείτονα (Nearest Neighbor – NN)", τις παραμέτρους ομαλοποίησης και τον τύπο του πυρήνα για την "Μηχανή Διανυσμάτων Υποστήριξης (Support Vector Machine – SVM)", και τις προδιαγραφές δέντρων για

“Τυχαία Δάση (Random Forests – RFs)”, οι οποίες ουσιαστικά χρησιμεύουν στον έλεγχο της πολυπλοκότητας του ταιριαστού μοντέλου. Παρομοίως, διάφορες στρατηγικές και προσεγγίσεις μοντελοποίησης είναι δυνατές για “Λογιστική Παλινδρόμηση (Logistic Regression – LOGREG)”.

Πρώτον, οι μοντελοποιητές προβλέψεων ιατρικών δεδομένων πρέπει να αξιολογούν τη μη γραμμικότητα των συνεχών μεταβλητών (Harrell, 2001). Η τυφλή εφαρμογή του μοντέλου λογιστικής παλινδρόμησης $y{\sim}x_1 + x_2$, δεν είναι πολύ ρεαλιστική. Το υποκείμενο μοντέλο του κύκλου απαιτεί κάποιο είδος λειτουργιών αύξησης και μείωσης για τα x_1 και x_2. Οποιοσδήποτε επιδημιολόγος θα έκανε κάποια μορφή επιθεώρησης δεδομένων και θα σημείωνε αμέσως την περισσότερο ή λιγότερο τετραγωνική σχέση με το x_i. Οι προτιμήσεις για μοντελοποίηση μη γραμμικότητας ποικίλλουν: Ο Harrell (2001) προτείνει “Περιορισμένες Κυβικές Συναρτήσεις Splines (Restricted Cubic Splines - Rcs)” ως προεπιλεγμένο εργαλείο στη μοντελοποίηση παλινδρόμησης, ενώ οι Royston & Sauerbrei (2008) υποστηρίζουν τη χρήση “Κλασματικών Πολυωνύμων (Fractional Polynomials - FP)”. Για παράδειγμα, η εφαρμογή των λειτουργιών FP και rcs σε μια προσομοίωση με 5000 υποκείμενα: Το πραγματικό αποτέλεσμα του x_1 είναι μια γραμμική αύξηση από $x_1 = 0$ σε $x_1 = 17$ μια πιθανότητα 1 μεταξύ $x_1 = 17$ και $x_1 = 33$ και μια γραμμική μείωση μεταξύ $x_1 = 33$ και $x_1 = 50$. Για το μοντέλο FP, ένας γραμμικός όρος συν ένας τετραγωνικός όρος επιλέχθηκαν για το x_1. Αυτό το μοντέλο FP ακολουθεί καλά το πραγματικό σχήμα, αν και δεν έχει επιτευχθεί η πιθανότητα 1 και οι χαμηλές πιθανότητες έχουν υποτιμηθεί. Το μοντέλο rcs (με 5 κόμβους, 4 *df*) έφτασε την πιθανότητα-οροπέδιο 1, αλλά ελαφρώς υπερεκτίμησε τις χαμηλές πιθανότητες στα $x_1 = 0$ και $x_1 = 50$. Τα μοντέλα $y{\sim}fp(x_1) + fp(x_2)$ και $y{\sim}rcs(x_1) + rcs(x_2)$ είχαν βαθμολογίες Brier κάτω από 0,15, που ισοδυναμεί με τις MLTs με την καλύτερη απόδοση σε αυτήν την προσομοίωση (NN, SVM-Bessel). Έτσι, όπως αναμενόταν, ένα λογικά καθορισμένο μοντέλο LOGREG αποδίδει πολύ καλά στην προσομοίωση.

Δεύτερον, ενώ κάποια μορφή ομαλοποίησης είναι επιτακτική για τις MLT λόγω της ευελιξίας τους, παρόμοιες τεχνικές υπάρχουν για logreg για να επιβάλει ποινές (penalize) ή να συρρικνώσει τους συντελεστές μοντέλου. Παραδείγματα είναι οι ποινικοποιήσεις L1 (LASSO) ή L2 (κορυφογραμμής) ή Bayesian. Η μέθοδος LASSO χρησιμοποιεί ποινή L1 για τη συρρίκνωση των συντελεστών παλινδρόμησης στο μηδέν. Ως εκ τούτου, το LASSO συνδυάζει την επιλογή μεταβλητών με τη

συρρίκνωση, παρέχοντας παράλληλα επαρκείς προβλέψεις, όπως παρατηρήθηκε σε μια μεγάλη μελέτη προσομοίωσης για ασθενείς με οξύ έμφραγμα του μυοκαρδίου (Steyerberg et al., 2000). Παρόμοια με τη βελτίωση του RF σε σχέση με το "Δέντρο Ταξινόμησης και Παλινδρόμησης (Classification And Regression Tree – CART)" για πρόβλεψη, θα πρέπει να χρησιμοποιούνται ποινικοποιήσεις αντί για παραδοσιακές προσεγγίσεις για το logreg εάν γίνονται συγκρίσεις μεταξύ logreg και MLT.

Ένα σημαντικό πρόβλημα για τα μοντέλα πρόβλεψης είναι η αβεβαιότητα του μοντέλου. Συνήθως μπορούν να καθοριστούν διάφορα μοντέλα, τα οποία περιγράφουν εύλογα τα δεδομένα. Στην ιατρική έρευνα, μπορεί συχνά να υπάρχει μια σχετικά μεγάλη λίστα πιθανών προβλεπτών, π.χ. 49 για την εφαρμογή I (εγκεφαλικό επεισόδιο) (Kruppa et al., 2014). Αυτή η λίστα βασίστηκε προφανώς σε βάσιμους λόγους (μια συστηματική ανασκόπηση της βιβλιογραφίας), αλλά κάποια μείωση ενδέχεται να ήταν εφικτή θέτοντας αυστηρότερα κριτήρια στα στοιχεία που υποστηλώνουν ένα πιθανό προγνωστικό αποτέλεσμα, όπως η σταθερότητα ενός μεγέθους ουσιώδους επιδράσεως σε πολλές μελέτες. Δεν είναι αληθοφανές ότι ένα ιατρικό πρόβλημα έχει 49 εξίσου σημαντικούς προβλέπτες (όπου η «σημασία» μπορεί να εξαρτάται από την τεχνική μοντελοποίησης που χρησιμοποιείται). Για παράδειγμα, εντοπιστηκαν μόνο 3 προβλέπτες-κλειδιά για έκβαση 6 μηνών σε μια συστηματική ανασκόπηση της βιβλιογραφίας για ασθενείς με τραυματική εγκεφαλική βλάβη. Σε αυτό το πρόβλημα πρόβλεψης, η "Κλίμακα Κώματος Γλασκώβης (Glasgow Coma Scale - GCS)" - ειδικά η συνιστώσα Κίνηση (Motor) - και η αντιδραστικότητα των οφθαλμικών κορών προέβλεπαν έντονα τη θνησιμότητα 6 μηνών. Τα μοντέλα με αυτούς τους προβλέπτες-κλειδιά είχαν καλή απόδοση σε χρονικές και γεωγραφικές επικυρώσεις. Μόνο ελάχιστες βελτιώσεις σημειώθηκαν συμπεριλαμβάνοντας άλλα χαρακτηριστικά, όπως ευρήματα αξονικής τομογραφίας, ενώ πολλοί ιατροί θα θεωρούσαν αυτά τα χαρακτηριστικά ζωτικής σημασίας για την πρόβλεψη.

Επιπλέον, είναι ευρέως γνωστό ότι τα ιατρικά δεδομένα έχουν συνήθως κακό «λόγο σήματος προς θόρυβο» για τους προβλέπτες. Αυτό έχει δύο επιπτώσεις. Πρώτον, το μέγεθος του δείγματος και η ποινικοποίηση είναι βασικοί παράγοντες για την ακριβή μοντελοποίηση προβλέψεων. Αυτό ισχύει για τα μοντέλα παλινδρόμησης, και ακόμη περισσότερο για τα MLT. Τα MLT είναι πιο ευέλικτα από την παλινδρόμηση, γεγονός που τα κάνει πιο πεινασμένα στα δεδομένα. Μια τεχνική όπως η NN μπορεί να είναι ακραία στις απαιτήσεις δεδομένων, λόγω της πλήρως μη παραμετρικής φύσης της.

Δεύτερον, πιο ελαστικές προδιαγραφές του μοντέλου μπορεί συχνά να επαρκούν για την καταγραφή της κύριας δομής ενός προβλήματος πρόβλεψης. Η ακραία μη γραμμικότητα, όπως στην παρουσίαση της προσομοίωσης, παρατηρείται σπάνια στην ιατρική έρευνα. Πολύπλοκες αλληλεπιδράσεις υψηλότερης τάξης μπορεί περιστασιακά να υπάρχουν αλλά είναι αδύνατο να εντοπιστούν σε εύλογου μεγέθους σύνολα ιατρικών δεδομένων. Αυτό υποστηρίζεται από πρόσφατες μελέτες που αναφέρουν παρόμοια απόδοση του LOGREG έναντι του MLT (Van Calster et al., 2009).

Τα σύνολα ιατρικών δεδομένων είναι συχνά πολύ μικρού μεγέθους για να είναι σε θέση να αντιμετωπίσουν αξιόπιστα δύσκολα ερευνητικά ερωτήματα, όπως ο προσδιορισμός των προβλεπτών που είναι σημαντικοί και ποιοι όχι. Για παράδειγμα, ο αξιόπιστος προσδιορισμός χαρακτηριστικών που προβλέπουν τη θνησιμότητα μέσα από 49 χαρακτηριστικά μπορεί να απαιτεί πολύ μεγαλύτερο αριθμό συμβάντων από ό, τι συμβαίνει στο εκπαιδευτικό σύνολο των 1737 ασθενών στην εφαρμογή I (Kruppa et al., 2014). Επιπλέον, η οπισθοδρομική εξάλειψη (backward elimination) είναι μια τυπική προσέγγιση για επιλογή μεταβλητών στην ανάλυση παλινδρόμησης, που συνήθως αξιολογείται χρησιμοποιώντας $p < 0.05$ για προβλέπτες σε ένα μοντέλο πρόβλεψης. Πολλά μειονεκτήματα έχουν συζητηθεί στο παρελθόν, συμπεριλαμβανομένης της προκατειλημμένης εκτίμησης των συντελεστών παλινδρόμησης, της παραμόρφωσης της εκτίμησης της διακύμανσης και των τιμών *p* και της αστάθειας του επιλεγμένου συνόλου προβλεπτών. Για την εκτίμηση πιθανότητας το πιο σχετικό ζήτημα είναι ότι η βηματική επιλογή (stepwise selection) οδηγεί σε υποβέλτιστη πρόβλεψη: επιλέγονται μόνο οι πιο εμφανείς προβλέπτες, οπότε οι πληροφορίες από σχεδόν σημαντικούς ή σημαντικούς προβλέπτες χάνονται και τα αποτελέσματα είναι υπερβάλοντα, γεγονός που οδηγεί σε υπερβολικά ακραίες προβλέψεις.

Η συνετή μοντελοποίηση θα πρέπει να βρει μια ισορροπία μεταξύ εξωτερικών γνώσεων εκτός δεδομένων και με αυτό που μπορεί να μάθει από τα δεδομένα. Όσο μικρότερο είναι το διαθέσιμο σύνολο δεδομένων, τόσο περισσότερο η ανάλυση πρέπει να βασιστεί σε εξωτερικές πληροφορίες. Αυτό ισχύει κυρίως για τη λίστα των υποψήφιων προβλεπτών σε ένα μοντέλο, το οποίο σχετίζεται τόσο με το MLT όσο και το logreg. Αλλά ισχύει και για ζητήματα όπως το αν πρέπει να βασιστεί κανείς στην παραδοχή προσθετικότητας στο logreg, δηλαδή εάν πρέπει να ληφθούν υπόψη οι όροι στατιστικής αλληλεπίδρασης. Μερικοί παραδοσιακοί στατιστικολόγοι θα μπορούσαν

να θεωρήσουν την αξιολόγηση των αλληλεπιδράσεων ως καλή πρακτική μοντελοποίησης, ενώ άλλοι θα προειδοποιούσαν για υπερβολικό ταίριασμα λόγω της δυνατότητας συμπερίληψης ψευδών αλληλεπιδράσεων. Τα ευρήματα σε προηγούμενες μελέτες και το μέγεθος του δείγματος των δεδομένων που μελετήθηκαν είναι βασικές εκτιμήσεις για τέτοιες στρατηγικές.

Το MLT έχει διάφορες ελκυστικές ιδιότητες, όπως η εστίασή του στην ομαλοποίηση και στην εύρεση αλγορίθμων και μοντέλων ταξινόμησης που δουλεύουν, αντί να επικεντρώνεται έντονα στη θεωρία ενός υποτιθέμενου στοχαστικού μοντέλου δεδομένων. Η έρευνα κλινικής πρόβλεψης κινδύνου χρησιμοποιεί μια παρόμοια φιλοσοφία, εστιάζοντας σε ζητήματα απόδοσης όπως διάκριση, βαθμονόμηση, χρησιμότητα και επίπτωση. Ωστόσο, το MLT έχει επίσης διάφορα προβλήματα. Εάν στοχεύει κανείς σε έναν σημαντικό ρόλο των μοντέλων πρόβλεψης στην ιατρική, πρέπει να ακολουθήσει ένα πλαίσιο εργασίας που δεν περιλαμβάνει μόνο την ανάπτυξη μοντέλων, αλλά περιλαμβάνει μια διαδικασία επικύρωσης και ενημέρωσης των μοντέλων. Η ενημέρωση ενδέχεται να απαιτεί προσαρμογές στις τοπικές ρυθμίσεις. Στο logreg, η απλή ενημέρωση της μέσης πιθανότητας επιτυγχάνεται εύκολα αλλάζοντας την τεταγμένη (intercept) του μοντέλου ενώ αυτό είναι δύσκολο για το MLT.

Επιπλέον, η ερμηνευσιμότητα σε ένα κλινικό κοινό είναι συνήθως απαραίτητη (Kruppa et al., 2014). Τα μοντέλα λογιστικής παλινδρόμησης μπορούν να παρουσιαστούν με διαφάνεια, με γνώση των σχετικών επιδράσεων των προβλεπτών μέσω λόγων πιθανοτήτων και με νομογράμματα, διαγράμματα βαθμολογίας και άλλες απεικονίσεις. Τέτοιες παρουσιάσεις δεν είναι δυνατές για MLT, αν και έχουν γίνει προσπάθειες προς το σκοπό αυτό. Ωστόσο, παρατηρείται ότι τα μοντέλα εφαρμόζονται όλο και περισσότερο στο Διαδίκτυο. Για παράδειγμα, ένας υπολογιστής κινδύνου για την πιθανότητα μετάλλαξης που σχετίζεται με το σύνδρομο Lynch έχει πάνω από 1000 προσβάσεις το μήνα (Kastrinos et al., 2011). Ο υπολογισμός κινδύνου βάσει διαδικτύου μπορεί να επιτρέψει στο υποκείμενο μοντέλο να είναι αρκετά περίπλοκο, π.χ. ένα MLT.

Μια προσέγγιση NN μπορεί να είναι ελκυστική λόγω της θεωρητικής ιδιότητας της συνοχής, αλλά είναι πεινασμένη στα δεδομένα (απαιτεί τεράστια μεγέθη δειγμάτων) και στερείται διερμηνείας, παρόμοια με τα RF και SVM. Η συνοχή των RF και SVM μπορεί να μην αποδειχθεί πλήρως, αλλά η ευελιξία είναι μεγάλη. Αν και το logreg δεν είναι συνεπές στην εκτίμηση των πιθανοτήτων, η ευελιξία μπορεί να είναι ουσιαστική

με μια σύγχρονη στρατηγική μοντελοποίησης. Το αφελές ταίριασμα (fitting) των γραμμικών κύριων επιδράσεων και οι μέθοδοι αυτόματης επιλογής, όπως η οπισθοδρομική βηματική επιλογή με $p < 0.05$, είναι υποβέλτιστες προεπιλεγμένες υλοποιήσεις του logreg. Οι μη γραμμικοί μετασχηματισμοί μπορούν άνετα να γίνουν από συναρτήσεις rcs και FP, και οι μέθοδοι συρρίκνωσης ή ποινικοποίησης όπως το LASSO παρέχουν καλύτερη προβλεπτική απόδοση από την τυπική. Οι απαιτήσεις μεγέθους δειγμάτων για το logreg εξαρτώνται από το πόσα εξωτερικά αποδεικτικά στοιχεία είναι διαθέσιμα και από το πόσο προθυμοποιείται ο αναλυτής να βασιστεί σε τέτοια στοιχεία, π.χ. σχετικά με τη συνάφεια και τις επιπτώσεις των προβλεπτών. Η ερμηνευσιμότητα των μεγεθών επίδρασης είναι εύκολα δυνατή από ένα ιατρικά εκπαιδευμένο κοινό και η ενημέρωση του μοντέλου μπορεί να επιτευχθεί εύκολα με απλές ή πιο προηγμένες διαδικασίες.

Εν γένει, φαίνεται ότι το LOGREG θα παραμείνει η προεπιλεγμένη προσέγγιση μοντελοποίησης για την εκτίμηση πιθανότητας στην πρόβλεψη ιατρικού κινδύνου, ειδικά όταν εφαρμόζεται με σύγχρονες προσεγγίσεις. Το MLT μπορεί να έχει συμπληρωματικό ρόλο, σε εξαιρετικά περίπλοκα προβλήματα και να παρέχει σύγκριση με τα αποτελέσματα παλινδρόμησης.

3. Η Έννοια του Airbnb

3.1. Δημιουργία του Airbnb

Το ιστορικό δημιουργίας του Airbnb είναι ήδη γνωστή στη Silicon Valley αλλά και πέρα από αυτή: τον Οκτώβριο του 2007, δύο άνεργοι απόφοιτοι σχολών τέχνης που ζούσαν σε ένα διαμέρισμα τριών υπνοδωματίων στο Σαν Φρανσίσκο, μη έχοντας να πληρώσουν το ενοίκιο, αποφάσισαν μεταξύ σοβαρού και αστείου την ενοικίαση μερικών στρωμάτων κατά τη διάρκεια ενός μεγάλου σχεδιαστικού συνεδρίου που πραγματοποιήθηκε στην πόλη και που είχε σαν συνέπεια την υπερ-πληρότητα των ξενοδοχείων της. Σε ορισμένους κύκλους, αυτή η ιστορία έχει επιτύχει το ίδιο μυθικό ανάστημα με μερικές από τις ήδη υπάρχουσες θρυλικές ιδρυτικές ιστορίες: όταν ο Μπιλ Μπάουερμαν έχυσε υγρή ουρεθάνη στη συσκευή βάφλας της συζύγου του, οπότε γεννήθηκε το παπούτσι με σόλα τύπου βάφλας της Nike, ή όταν ο Bill Hewlett και ο Dave Packard έχτισαν έναν ταλαντωτή ήχου στο πλέον διάσημο γκαράζ της Packard.

Στην πραγματικότητα, η ιστορία του Airbnb ξεκινά λίγα χρόνια πριν, τρία χιλιάδες μίλια μακριά στο Providence, στο Rhode Island, σε ένα στούντιο στην πανεπιστημιούπολη της Σχολής Σχεδιασμού το καλοκαίρι του 2004. Οι Brian Chesky και Joe Gebbia, δύο φοιτητές – ο Gebbia ήταν στο τέταρτο έτος του πενταετούς διπλού πτυχίου στη βιομηχανική και γραφιστική και ο Chesky είχε μόλις αποφοιτήσει - ήταν μέρος ενός ερευνητικού έργου που χρηματοδοτήθηκε από την RISD με την Conair Corporation, εταιρεία γνωστή για τα στεγνωτήρες μαλλιών και άλλα προϊόντα προσωπικής φροντίδας. Οι εταιρείες συχνά συνεργάζονται με το RISD για πρόσβαση σε φοιτητές βιομηχανικού σχεδιασμού. Σύμφωνα με αυτό το συγκεκριμένο πρόγραμμα, η Conair «προσέλαβε» το σχολείο, το οποίο ανέθεσε σε μια ομάδα φοιτητών να εργάζονται ουσιαστικά αποκλειστικά στο σχεδιασμό προϊόντων για την εταιρεία κατά τη διάρκεια έξι εβδομάδων. Το μεγαλύτερο μέρος της εργασίας θα πραγματοποιείτο στην πανεπιστημιούπολη RISD, αλλά η εταιρεία θα είχε τα δικαιώματα για τα προϊόντα, και οι φοιτητές θα αποκτούσαν πραγματική εργασιακή εμπειρία και κάποιο επίδομα. Στο τέλος του προγράμματος, θα παρουσίαζαν τις ιδέες τους στα στελέχη της Conair.

Σε όρους εκκίνησης στη Silicon Valley, οι Chesky, Gebbia και Blecharczyk είχαν επιτύχει αυτό που είναι γνωστό ως «προϊόν/κατάλληλο για αγορά», το ιερό ποτήριο, όταν η ιδέα τους συνδύασε και τα δύο στοιχεία, μία καλή αγορά -με πολλούς

πραγματικούς, δυνητικούς πελάτες -και την απόδειξη ότι έχουν δημιουργήσει ένα προϊόν που μπορεί να ικανοποιήσει την αγορά αυτήν. Η εκλαΐκευση του όρου πιστώνεται συχνά στον Marc Andreessen, τον διάσημο επιχειρηματία της τεχνολογίας - που από επιχειρηματικός καπιταλιστής, μετατράπηκε σε φιλόσοφο-γκουρού για τις λεγεώνες ιδρυτών των νεοσύστατων εταιριών στη Silicon Valley. Χιλιάδες νεοσύστατες επιχειρήσεις δεν προσπάθησαν να φτάσουν σε αυτό το σημείο. Η προσαρμογή προϊόντος/αγοράς είναι ένα βασικό πρώτο επίτευγμα, χωρίς το οποίο, δεν υπάρχει εταιρία. Ένας άλλος τρόπος για να περιγραφεί το παραπάνω είναι ότι η εταιρεία προσφέρει να «κάνει κάτι που οι άνθρωποι θέλουν». Όπως κι αν ονομαστεί, οι Τσέσκι, Γκέμπια και Μπλερσκιζίκ είχαν φτάσει σε εκείνη την κρίσιμη συγκυρία τον Απρίλιο του 2009 όταν το «χαμόγελο της ελπίδας» είχε μετατραπεί σε μια ασταμάτητη ροή εσόδων. Προσέφεραν ένα προϊόν που οι άνθρωποι το ήθελαν. Και αυξανόταν: μέχρι τον Αύγουστο του 2009, τα 1.000 $ έσοδά την εβδομάδα, έφθασαν τα 10.000 $ για να εκτιναχθούν σχεδόν στα 100.000 $.

Το επίτευγμα έφθασε στο δύσκολο μέρος. Ο στόχος έπρεπε να μετατοπιστεί πιο μακροπρόθεσμα: χρειάζονταν ένα σχέδιο, έναν χάρτη πορείας, μια στρατηγική. Χρειάζονταν υπαλλήλους. Χρειάζονταν μια κουλτούρα. Είχαν το προϊόν, μα τώρα έπρεπε να χτίσουν την εταιρεία που θα παρήγαγε αυτό το προϊόν.

Αλλά ήταν ακόμα μόνο οι τρεις τους, εργάζονταν δεκαοκτώ ώρες την ημέρα, επτά ημέρες την εβδομάδα, μαζί και έκαναν σχεδόν όλα τα άλλα μαζί. «Ίσως όλοι έχουμε πάρει φόρμες», είπε αργότερα ο Τσέσκι κατά τη διάρκεια μιας συζήτησης για τον πολιτισμό με τον συνεργάτη της Sequoia και το μέλος του διοικητικού συμβουλίου της Airbnb, Alfred Lin για το μάθημα του Πανεπιστημίου του Στάνφορντ, «Πώς να ξεκινήσετε μια νεοσύστατη εταιρία» («How to start a Startup»). Άρχισαν να σκέφτονται την πιο πιεστική ανάγκη τους ήδη από τις ημέρες του Y Combinator, την πρόσληψη του πρώτου μηχανικού τους, αλλά τώρα η ανάγκη αυτή είχε γίνει επιτακτική. Ο Blecharczyk έκανε ακόμα μόνος του όλες τις τεχνικές εργασίες.

Είχαν επίσης αρχίσει να φαντάζονται το είδος της εταιρείας που ήθελαν να οικοδομήσουν και είχαν καταλήξει στο συμπέρασμα ότι η είσοδος των σωστών ανθρώπων θα είχε καθοριστικό μακροπρόθεσμο αντίκτυπο. Τέτοιες αποφάσεις δεν έπρεπε να ληφθούν ελαφρά τη καρδία. Ο Τσέσκι είχε διαβάσει αρκετά βιβλία για την εταιρική κουλτούρα και ένιωθε ότι τόσο αυτός όσο και οι συνάδελφοί του έπρεπε να

είναι προσεκτικοί σχετικά με το ποιόν θα έφερναν. "Νομίζω ότι η πρόσληψη του πρώτου μηχανικού σας είναι σαν να φέρετε DNA στην εταιρεία σας", είπε στους μαθητές του Στάνφορντ. Με άλλα λόγια, δεν έψαχναν κάποιον να δημιουργήσει τα επόμενα στοιχεία. Αν όλα πήγαιναν καλά, αυτό το άτομο θα κατέληγε να φέρει εκατοντάδες άτομα σαν αυτόν ή αυτήν. Έτσι, η απόκτηση της πρώτης πρόσληψης ήταν πολύ σημαντική.

Έκαναν μια λίστα με εταιρείες των οποίων την κουλτούρα ήθελαν να μιμηθούν. Τώρα πια έχουν τη δυνατότητα πρόσβασης σε εισαγωγές υψηλού επιπέδου μέσω του δικτύου Sequoia - ο Greg McAdoo είχε γίνει στενός σύμβουλος και όλοι έπαιρναν πρωινό μία ή δύο φορές την εβδομάδα μαζί στο Rocco's, ένα σημείο της γειτονιάς – ο Chesky, ο Gebbia και ο Blecharczyk μπόρεσαν να προσεγγίσουν εταιρείες όπως την Zappos, των οποίων η κουλτούρα φιλικότητας και «ζωηρότητας» ήταν ιδιαίτερα αξιοθαύμαστη, όπως συνέβαινε και με τις Starbucks, Apple, Nike κ.λ.π.. Κατά τη διάρκεια μίας τέτοιας συνάντησης, ζήτησαν από τον McAdoo μια σύσταση στον CEO της Zappos, Tony Hsieh, τον οποίο γνώριζε ο McAdoo από τότε που η Zappos ήταν εταιρεία χαρτοφυλακίου της Sequoia. Ο McAdoo έστειλε μια γρήγορη εισαγωγή μέσω e-mail καθώς ήταν στο δρόμο του για το αυτοκίνητό του, και την επόμενη μέρα όταν κάλεσε τον Chesky, εξεπλάγη όταν έμαθε ότι οι ιδρυτές βρίσκονταν ήδη στο Λας Βέγκας, περιοδεύοντας στα κεντρικά γραφεία της Zappos.

Το μέγεθος και η κλίμακα της Airbnb μπορούν να παρουσιαστούν με διάφορους τρόπους. Ο ευκολότερος είναι τα 140 εκατομμύρια «αφίξεων επισκεπτών» από την ημέρα της επινόησή της. Τα 3 εκατομμύρια ενεργές καταχωρίσεις της - το 80% των οποίων είναι εκτός της Βόρειας Αμερικής - κάνουν την Airbnb τον μεγαλύτερο πάροχο καταλυμάτων στον κόσμο, μεγαλύτερο από οποιαδήποτε αλυσίδα ξενοδοχείων. (Με την εξαγορά της Starwood, η Marriott International έχει το μεγαλύτερο απόθεμα οποιασδήποτε ξενοδοχειακής εταιρείας, με 1,1 εκατομμύρια δωμάτια.) Ωστόσο, η Airbnb δεν μοιάζει με ξενοδοχείο - ο αριθμός των καταχωρίσεών της αλλάζει κάθε δεδομένη ημέρα και διογκώνεται γύρω από μεγάλες εκδηλώσεις και μεγάλος αριθμός από τα παρεχόμενα δωμάτια αδειάζει κάθε βράδυ, ανάλογα με τα χρονοδιαγράμματα και τις προτιμήσεις συχνότητας των ενοικιαστών. Έτσι, ακόμη κι αν ο τεράστιος αριθμός καταχωρίσεων δεν συσχετίζεται με την πληρότητα ή τον όγκο συναλλαγών, ωστόσο δείχνει το εύρος και την κλίμακα.

Η εταιρεία λειτουργεί σε 191 χώρες - παντού εκτός από το Ιράν, τη Συρία και τη Βόρεια Κορέα, όπως θέλει να επισημαίνει - και σε 34.000 πόλεις. Δύο από τα πράγματα που αρέσουν περισσότερο στους επενδυτές της Airbnb είναι η αποτελεσματικότητά της και η ανάπτυξή της. Μπορεί να επεκταθεί με πολύ χαμηλό κόστος, έχει ξοδέψει λιγότερα από 300 εκατομμύρια δολάρια συνολικά για οκτώ χρόνια, σύμφωνα με εκτιμήσεις. Η κοινή οικονομία της Uber λέγεται ότι έχασε 1,2 δισεκατομμύρια δολάρια μόνο το πρώτο εξάμηνο του 2016. Και, οκτώ χρόνια μετά, η Airbnb συνεχίζει να μεγαλώνει. Από τη μελέτη και μόνο αυτή του (Gallagher, 2017), η εταιρεία λέγεται ότι θα προσθέτει 1,4 εκατομμύρια χρήστες την εβδομάδα, και οι 140 εκατομμύρια «αφίξεις επισκεπτών» αναμένεται να αυξηθούν σε 160 εκατομμύρια έως τις αρχές του 2017. Οι επενδυτές περίμεναν ότι η εταιρεία θα δει 1,6 δισεκατομμύρια δολάρια σε έσοδα και να αποκτήσει θετικές ταμειακές ροές το 2016.

3.2. Ενίσχυση και Αύξηση του Airbnb

Πρόκειται για ένα επιχειρηματικό έπος. Ο αγώνας που έδωσαν οι ιδρυτές για να την απογειώσουν, η τεχνολογία, το προϊόν και η κουλτούρα που δημιούργησαν, και ο τρόπος με τον οποίο έγινε γρήγορα μια μηχανή υψηλής απόδοσης αποτελεί μια ιστορία εκπληκτικής εταιρικής ευκινησίας. Το γεγονός ότι όλα αυτά επιτεύχθηκαν μέσα σε λίγα χρόνια και με πολύ μικρή προηγούμενη εμπειρία, είναι εντυπωσιακό.

Ωστόσο η μελέτη του τί έχει συμβεί στους τέσσερις τοίχους της ίδιας της εταιρείας από μόνη της θα ήταν η αιτία να χαθεί σχεδόν ολόκληρη την «ιστορία» της Airbnb. Η Airbnb – ως εταιρεία – απαριθμεί περίπου 2.500 άτομα, κυρίως στο Σαν Φρανσίσκο. Η Airbnb – ως κίνημα – απαριθμεί εκατομμύρια ανθρώπων στη γη.

Πολλά εκατομμύρια άνθρωποι έχουν χρησιμοποιήσει τουλάχιστον μία φορά το Airbnb. Η επιχείρηση είναι εποχιακή, αλλά η εταιρεία σημείωσε νέο νυχτερινό ζενιθ κατά τη διάρκεια του καλοκαιριού του 2016, όταν 1,8 εκατομμύρια άνθρωποι έμειναν σε καταλύματα της Airbnb σε μία μόνο νύχτα. Ωστόσο, ακόμη και με αυτούς τους αριθμούς, η διείσδυση της εταιρείας είναι ακόμα χαμηλή: πολλοί άνθρωποι δεν έχουν ακούσει καθόλου για αυτήν και τους γίνει γνωστή η ιδέα, τους ακούγεται ακόμα τόσο παράξενο όσο ακουγόταν στους πρώτους λίγους επενδυτές.

Πολλοί άνθρωποι, αναπήδησαν με την αναφορά της ιδέας στα πλαίσια της εργασίας του (Gallagher, 2017). Για μερικούς από αυτούς, υπάρχει ένας παράγοντας «ew». «Δεν θα το έκανα ποτέ αυτό», απάντησε φίλος ενός φίλου. "Τι γίνεται αν

καταλήξετε στα βρώμικα σεντόνια κάποιου;" Η αντίδραση ενός οδηγού ήταν τυπική. Δεν το είχε ακούσει, αλλά αφού του αναλύθηκε η ιδέα, κούνησε το κεφάλι του και είπε ότι απλά δεν θα το έκανε ποτέ. Πρώτον, είπε, έτσι ακριβώς μεταδίδονται οι κοριοί. Επιπλέον, επεσήμανε, αν ανοίξετε τις πόρτες σας σε αγνώστους, δεν έχετε ιδέα ποιον αφήνετε να μπει στο σπίτι σας. Θα μπορούσατε να έχετε έναν δολοφόνο» Είχε δίκιο. Θα μπορούσε. Πολλά πράγματα πήγαν στραβά: Φυσικά, υπήρχε η περίπτωση του EJ καθώς και άλλα άσχημα επεισόδια. Ωστόσο, οποιαδήποτε μελέτη του φαινομένου Airbnb πρέπει πρώτα να εξετάσει την ανάγκη που έχει εντοπίσει και το κενό που έχει γεμίσει. Επειδή δεν είναι δυνατή η προσέγγιση εκατομμυρίων πελατών ανά τον κόσμο χωρίς, όπως θα έλεγε ο Paul Graham, να παρέχεται κάτι που οι άνθρωποι θέλουν.

Στα πρώτα χρόνια της, η Airbnb, είχε τη φήμη ότι είναι ένας ιστότοπος όπου πλήθος χρηστών έψαχναν για φθηνές επιλογές και έμειναν στο σαλόνι κάποιου ή σε ένα εφεδρικό υπνοδωμάτιο. Αλλά με την πάροδο του χρόνου η εταιρία εξελίχθηκε. Εάν υπήρχαν τρεις φάσεις της Airbnb, θα μπορούσαν να κατηγοριοποιηθούν πολύ απλά ως η φάση του καναπέ-σέρφινγκ των πρώτων ημερών, τη φάση του ιγκλού και του κάστρου, όταν η ανάπτυξη άρχισε να απογειώνεται και η εταιρεία έγινε γνωστή για όλους τους περίεργους, ξεπερασμένους χώρους και τελευταία η φάση Gwyneth Paltrow, όταν η βάση χρηστών και οι παροχές της είχαν επεκταθεί σε τέτοιο βαθμό που η γνωστή ηθοποιός πέρασε διακοπές τον Ιανουάριο του 2016 σε μία από τις καταχωρήσεις της Airbnb των 8.000$ ανά διανυκτέρευση στην Punta Mita του Μεξικού και στη συνέχεια επέστρεψε λίγους μήνες αργότερα κλείνοντας μια βίλα στην Κυανή Ακτή με 10.000$ ανά διανυκτέρευση. Η σημασία της φάσης Paltrow ήταν διπλή: πρώτον, η Airbnb είχε γίνει πια μια νόμιμη επιλογή για τους πιο διάσημους και πιο εξελιγμένους ταξιδιώτες και, δεύτερον, είχε γίνει μια πλατφόρμα τόσο μεγάλη που ουσιαστικά είχε κάτι για όλους.

Σήμερα, το εύρος του αποθέματος της Airbnb αντικατοπτρίζει την ποικιλομορφία στην παγκόσμια αγορά κατοικιών. Τα τρία εκατομμύρια καταχωρίσεις της είναι όλες μοναδικές και το εύρος των διαθέσιμων ιδιοκτησιών και εμπειριών είναι δύσκολο να το φανταστεί κανείς. Μπορεί κάποιος με 20$ για να κοιμηθεί σε ένα στρώμα αέρα στην κουζίνα κάποιου ή να πληρώσει δεκάδες χιλιάδες δολλαρίων την εβδομάδα για μια βίλα στο Μεξικό όπως η Paltrow. Σε μία συγκεκριμένη ημέρα, οι επιλογές στη Νέα Υόρκη κυμαίνονταν από 64$ ανά διανυκτέρευση για ένα υπόγειο διαμέρισμα στην Τζαμάικα, Queens, έως 3.711$ για ένα πενταώροφο αρχοντικό στην

East Tenth Street. Στο Παρίσι, 24$ θα αντιστοιχούσαν σε ένα δωμάτιο με ένα διπλό κρεβάτι και έναν νιπτήρα στο νοτιοδυτικό προάστιο Fontenay-aux-Roses, αλλά με 8.956$ θα μπορούσε κανείς να περάσει μία νύχτα σε ένα τριπλό διαμέρισμα στο 16ο Τομέα με ιδιωτικό κήπο, θέα στον Πύργο του Άιφελ και "VIP υπηρεσίες ξενοδοχείου".

Κάποια στιγμή το 2013, η Airbnb άρχισε να σκέφτεται να επαναπροσανατολίσει ολόκληρη την αποστολή και το κέντρο βάρους της για να διαρθρώσει καλύτερα τα στοιχεία που έκαναν τη χρήση της πλατφόρμας της τόσο μοναδική. Σε μια διαδικασία με επικεφαλής τον Douglas Atkin, τον παγκόσμιο επικεφαλής της κοινότητας της εταιρείας που είχε ενταχθεί σε αυτήν νωρίτερα εκείνο το έτος, εστίασε αυτές τις πτυχές γύρω από μια και μόνη ιδέα, την έννοια του «ανήκειν».

Ο Atkin, ειδικός στη σχέση μεταξύ καταναλωτών και εμπορικών σημάτων και συγγραφέας του βιβλίου «The Culting of Brands», ήρθε στην παραπάνω ιδέα μετά από μήνες διασκέψεων με περίπου πεντακόσια μέλη της βάσης χρηστών της Airbnb σε όλο τον κόσμο, και μέχρι τα μέσα 2014 η εταιρεία είχε καταλήξει σε μια ολόκληρη επανατοποθέτηση γύρω από αυτήν την ιδέα. Η Airbnb είχε μια νέα αποστολή: να κάνει τους ανθρώπους σε όλο τον κόσμο να νιώθουν σαν να μπορούν «να ανήκουν οπουδήποτε» (“belong anywhere”). Απέκτησε νέο χρώμα: το ματζέντα. Και απέκτησε κι ένα νέο λογότυπο για να συμβολίσει τη νέα αποστολή: ένα χαριτωμένο, μικροσκοπικό σχήμα που ήταν αποτέλεσμα μηνών σύλληψης και εξευγενισμού με την ονομασία «Μπέλο». Ονομάστηκε έτσι από τον νέο επικεφαλής μάρκετινγκ της εταιρείας, Jonathan Mildenhall, ο οποίος προσχώρησε πρόσφατα από την Coca-Cola. Ο Mildenhall έπεισε τους ιδρυτές να επεκτείνουν το «να ανήκουν οπουδήποτε» (“belong anywhere”) από μια εσωτερική δήλωση αποστολής στην επίσημη ετικέτα της εταιρείας.

Τον Ιούλιο του 2014, η εταιρεία παρουσίασε την επωνυμία, καθώς και έναν επανασχεδιασμό της εφαρμογής και του ιστότοπού της για κινητά, σε μια μεγάλη εκδήλωση στην έδρα της. Ο Τσέσκι παρουσίασε την ιδέα σε ένα δοκίμιο στον ιστότοπο της Airbnb: Πριν από πολύ καιρό, έγραψε, οι πόλεις ήταν χωριά. Αλλά καθώς ήρθε η μαζική παραγωγή και η εκβιομηχάνιση, αυτό το προσωπικό συναίσθημα αντικαταστάθηκε από «μαζικές και απρόσωπες ταξιδιωτικές εμπειρίες» και στην πορεία, «οι άνθρωποι σταμάτησαν να εμπιστεύονται ο ένας τον άλλον». Η Airbnb,

δήλωσε, θα υπερασπίζεται κάτι πολύ μεγαλύτερο από το ταξίδι. θα υπερασπίζεται την κοινότητα και τις σχέσεις και τη χρησιμοποίηση της τεχνολογίας με σκοπό την προσέγγιση των ανθρώπων. Η Airbnb θα ήταν το μέρος που θα μπορούσαν να πάνε οι άνθρωποι για να γνωρίσουν την «παγκόσμια ανθρώπινη λαχτάρα του ανήκειν». Το ίδιο το Bélo σχεδιάστηκε προσεκτικά για να μοιάζει με καρδιά, δείκτη θέσης και το "A" του Airbnb. Σχεδιάστηκε για να είναι απλό, έτσι ώστε ο καθένας να μπορεί να το σχεδιάσει. Αντί να το προστατεύσει με δικηγόρους και εμπορικά σήματα, η εταιρεία κάλεσε τους ανθρώπους να σχεδιάσουν τις δικές τους εκδόσεις του λογότυπου - οι οποίες, όπως ανακοινώθηκε, θα αποτελούσαν τέσσερα πράγματα: άτομα, μέρη, αγάπη και Airbnb.

3.3. Ανάπτυξη του Airbnb και Δεδομένα – DataSet

Από την ίδρυσή της το 2008, η Airbnb έγινε ένας από τους μεγαλύτερους διαδικτυακούς παρόχους διαμονής και επιτομή της οικονομίας πλατφόρμας. Προσφέρει προς ενοικίαση περισσότερα από 7 εκατομμύρια σπίτια, διαμερίσματα και δωμάτια σε όλες σχεδόν τις χώρες (Airbnb, 2019). Η πλατφόρμα έχει αποκτήσει δημόσια και επιστημονική προσοχή λόγω των αναταραχών που έφερε στη βιομηχανία φιλοξενίας, των επιπτώσεων στις αγορές κατοικιών και των νομικών συγκρούσεων σχετικά με τη στέγαση, τη φορολογία και τους κανονισμούς καταναλωτών (Guttentag, 2015; Hassanli et al., 2019).

Οι μελέτες σχετικά με την Airbnb έχουν συχνά υποτιμήσει την ποικιλομορφία της δραστηριότητας της όσον αφορά τόσο τους τύπους προσφορών όσο και τις γεωγραφικές τοποθεσίες τους. Οι ερευνητές τείνουν να αναγνωρίζουν την Airbnb ως μια ποιοτικά νέα μορφή διαμονής, συνδέοντάς την με ομότιμη ενοικίαση διαμερισμάτων ή δωματίων από μη εμπορικούς παρόχους μέσω άμεσης αλληλεπίδρασης (Dolnicar, 2019). Η δραστηριότητα επαγγελματιών φιλοξενούντων και επιχειρήσεων διαμονής, συμπεριλαμβανομένων αυτών που παρείχαν τις υπηρεσίες τους πριν από την ένταξή τους στην Airbnb, δεν αναγνωρίζεται ή αντιμετωπίζεται ως ανωμαλία. Επίσης, οι μελέτες στην πλατφόρμα που προσφέρει διαμονή σε περισσότερες από 100.000 πόλεις και 191 χώρες (Airbnb, 2019) επικεντρώνονται συνήθως σε περιορισμένο αριθμό μεγάλων πόλεων στις ανεπτυγμένες χώρες (Guttentag, 2019). Η αναγνώριση της τυπολογικής και γεωγραφικής ποικιλομορφίας της προσφοράς διαμονής της Airbnb είναι απαραίτητη για τον υπολογισμό των

επιπτώσεων της τρέχουσας και μελλοντικής ανάπτυξης πλατφορμών στην τουριστική οικονομία.

Διεθνείς μελέτες σχετικά με τη χρήση και τις επιπτώσεις της Airbnb εμποδίζονται από την έλλειψη δημόσιων προσβάσιμων δεδομένων. Ο Τύπος και οι ακαδημαϊκές δημοσιεύσεις αναφέρουν στοιχεία σχετικά με τον αριθμό, τη χρήση και τη διανομή των προσφορών της Airbnb που παρέχονται στα δελτία τύπου και τις εκθέσεις της πλατφόρμας. Τέτοιες πηγές δεδομένων εξαρτώνται από τη στρατηγική μάρκετινγκ της Airbnb, είναι φτωχές πληροφοριακά και δεν διαθέτουν μεθοδολογικές λεπτομέρειες. Ως εναλλακτική λύση σε αυτά τα στατιστικά στοιχεία, οι ερευνητές χρησιμοποιούν ευρέως διαγραμμένα δεδομένα σε προσφορές της πλατφόρμας, συμπεριλαμβανομένων συνόλων δεδομένων από το Inside Airbnb ή το AirDNA (Guttentag, 2019). Αυτό, ωστόσο, περιορίζει συνήθως το γεωγραφικό πεδίο ανάλυσης σε μεμονωμένες πόλεις. Επομένως, υπάρχει ανάγκη για περισσότερες διεθνούς κλίμακας μελέτες, τόσο περιγραφικές (π.χ. διεθνείς συγκρίσεις) όσο και επεξηγηματικές (π.χ. λόγοι για διαφορές μεταξύ χωρών) που θα έθεταν ένα ευρύτερο πλαίσιο για μελλοντικές τοπικές μελέτες και θα βοηθούσαν στην αξιολόγηση των επιπτώσεων της δραστηριότητας της Airbnb στους προορισμούς.

3.3.1. Δεδομένα

Τα δεδομένα για τις καταχωρίσεις της Airbnb ελήφθησαν από την πλατφόρμα χρησιμοποιώντας σενάριο διαγραφής ιστού (Slee, 2018). Η διαγραφή πραγματοποιήθηκε δύο φορές σε διάστημα ενός έτους: τον Σεπτέμβριο και τον Οκτώβριο του 2018 (4,7 εκατομμύρια καταχωρίσεις) και τον Σεπτέμβριο του 2019 (5,7 εκατομμύρια καταχωρίσεις). Μόνο οι καταχωρίσεις που ήταν διαθέσιμες τους επόμενους μήνες αποθηκεύτηκαν στο σύνολο δεδομένων, με συνέπεια την υποτίμηση του συνολικού αριθμού προσφορών έως και 20% (Adamiak et al., 2019 για περισσότερες λεπτομέρειες σχετικά με την αξιοπιστία των αποτελεσμάτων διαγραφής ιστού). Τα σύνολα δεδομένων περιλάμβαναν τις γεωγραφικές συντεταγμένες των καταχωρήσεων, πληροφορίες σχετικά με τους τύπους ιδιοκτησίας (ολόκληρο το σπίτι/ιδιοκτησία, ιδιωτικό δωμάτιο, κοινόχρηστο δωμάτιο ή δωμάτιο ξενοδοχείου), τις τιμές αναφοράς ανά διανυκτέρευση, τον αριθμό των κριτικών και τη μέση βαθμολογία. Με βάση τον αριθμό ταυτοποίησης των κεντρικών φιλοξενούντων, οι προσφορές χωρίστηκαν σε λίστες με απλή φιλοξενία, όπου ένας χρήστης πλατφόρμας προσφέρει

μόνο μία ιδιοκτησία και λίστες με πολλά σημεία φιλοξενίας όπου ο ίδιος χρήστης προσφέρει περισσότερες από μία θέσεις ενοικίασης.

Ένα από τα μειονεκτήματα των δεδομένων που έχουν διαγραφεί από τον ιστό είναι ότι περιλαμβάνουν πληροφορίες για προσφορές και όχι κρατήσεις, ενώ ορισμένες από τις καταχωρήσεις ενδέχεται να δημιουργηθούν κατά λάθος ή να μην έχουν χρησιμοποιηθεί καθόλου (Coles et al., 2018). Με σκοπό το φιλτράρισμα των ανενεργών καταχωρίσεων, το σύνολο δεδομένων περιορίστηκε σε καταχωρίσεις που είχαν λάβει τουλάχιστον μία κριτική μεταξύ των δύο καταργήσεων ή εμφανίστηκαν μόνο στο νεότερο σύνολο δεδομένων και είχαν τουλάχιστον μία κριτική (υποθέτοντας ότι ήταν νέες καταχωρίσεις το 2019 και είχαν ήδη ενοικιαστεί). Περίπου το 62,4% των συνολικών καταχωρίσεων από το σύνολο δεδομένων του 2019 πληρούσε ένα από αυτά τα κριτήρια και χρησιμοποιήθηκε στην ανάλυση. Δεδομένου ότι δεν αφήνουν όλοι οι επισκέπτες τις κριτικές τους τα μέρη της Airbnb, ο αριθμός των ενεργών καταχωρίσεων ενδέχεται να είναι μικρότερος

Στη συνέχεια, οι ενεργές προσφορές ανατέθηκαν σε 167 χώρες και εξαρτημένες περιοχές με βάση τη διεύθυνση κειμένου στην περιγραφή κάθε καταχώρισης. Υπάρχουν πολλές χώρες χωρίς καταχωρίσεις στην πλατφόρμα Airbnb: πρόκειται για ορισμένες αφρικανικές χώρες, μικρές νησιωτικές χώρες και χώρες όπου οι αμερικανικές εταιρείες δεν μπορούν να λειτουργήσουν λόγω των κυρώσεων της κυβέρνησης των ΗΠΑ, π.χ. Ιράν και Βόρεια Κορέα (το ίδιο εφαρμόζεται και το έγκλημα σε μια περιοχή). Τα περισσότερα εξαρτώμενα εδάφη, τα οποία διαθέτουν καταχώριση Airbnb, επίσης εξαιρέθηκαν από τη μελέτη εξαιτίας της έλλειψης συγκριτικών στατιστικών δεδομένων (π.χ. υπερπόντια διαμερίσματα της Γαλλίας: Γουαδελούπη, Μαρτινίκα και Ιλ ντε Ρεϋνιόν με 3000–5000 ενεργές καταχωρήσεις το καθένα).

Η ανάλυση και παρουσίαση των αποτελεσμάτων της εργασίας του (Adamiak, 2019) πραγματοποιήθηκε με τη χρήση του ArcGIS Pro and RStudio με πακέτα: ggplot2, gridExtra, ggalluvial, rgdal και sf (Auguie, 2017; Pebesma , 2018). Κατά την παρουσίαση των εθνικών δεδομένων, οι χώρες ομαδοποιήθηκαν σε περιοχές βάσει της περιφερειοποίησης του UNWTO (UNWTO, 2019).

Οι προσφορές της Airbnb είναι ως επί το πλείστον ενοικιαζόμενα σπίτια και διαμερίσματα, και οι πολυ-οικοδεσπότες προσφέρουν πάνω από το ήμισυ της παροχής

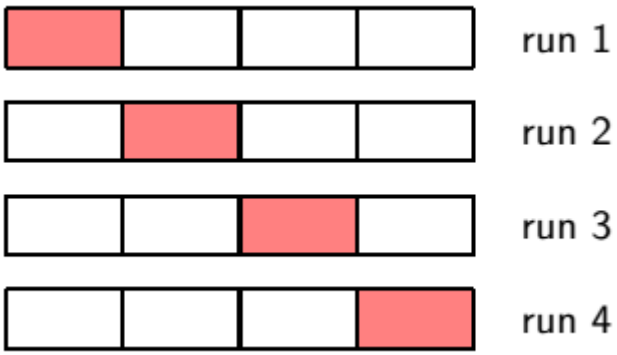
της πλατφόρμας. Η Airbnb δραστηριοποιείται στις περισσότερες χώρες του κόσμου, ωστόσο το ήμισυ των παροχών της βρίσκεται στην Ευρώπη. Η χρήση της πλατφόρμας ως ομότιμη αγορά είναι σχετικά πιο δημοφιλής στην Αμερική, τη Δυτική και τη Βόρεια Ευρώπη και την Ωκεανία, ενώ στον υπόλοιπο κόσμο η αποτελεί πρωτίστως μια πλατφόρμα επαγγελματικής ενοικίασης κατοικιών και εξυπηρέτησης. Ο συνολικός αριθμός των καταχωρίσεων της Airbnb στις χώρες εξαρτάται από το επίπεδο οικονομικής ανάπτυξης και το μέγεθος της εισερχόμενης τουριστικής ροής. Ο τελευταίος παράγοντας επηρεάζει περισσότερο τον αριθμό των επαγγελματικών προσφορών. Το ένα τρίτο της παγκόσμιας προσφοράς Airbnb βρίσκεται σε μεγάλες πόλεις, ενώ ένα άλλο τρίτο στις παράκτιες περιοχές. Η τοποθεσία των προσφορών σε κάθε χώρα, ιδίως εκείνων που προσανατολίζονται επαγγελματικά, εξαρτάται από τη διανομή των κύριων τουριστικών αξιοθέατων. Υπάρχουν επίσης διαφορές στη συχνότητα χρήσης και τις τιμές των καταχωρίσεων μεταξύ και εντός των χωρών.

4. Πρακτικό Μέρος

4.1. Cross Validation

Σε πολλές εφαρμογές, η παροχή δεδομένων για εκπαίδευση και δοκιμές είναι περιορισμένη, και προκειμένου τα μοντέλα που θα αναπτυχθούν να είναι αξιόπιστα, πρέπει να χρησιμοποιηθούν όσο το δυνατόν περισσότερα από τα διαθέσιμα δεδομένα για εκπαίδευση. Ωστόσο, εάν το σύνολο επικύρωσης είναι μικρό, θα δώσει μία σχετικά εσφαλμένη εκτίμηση της προγνωστικής απόδοσης. Μια λύση σε αυτό το δίλημμα είναι για χρήση διασταυρούμενης επικύρωσης όπως αυτή φαίνεται στο ακόλουθο σχήμα. Η λύση αυτή επιτρέπει μια αναλογία (S - 1) / S των διαθέσιμων δεδομένων που πρόκειται να χρησιμοποιηθούν για εκπαίδευση κατά τη χρήση όλων των δεδομένων για την αξιολόγηση της απόδοσης. Όταν τα δεδομένα είναι ιδιαίτερα σπάνια, μπορεί να είναι σκόπιμο να ληφθεί υπόψη η περίπτωση S = N, όπου N είναι ο συνολικός αριθμός σημείων δεδομένων, που δίνει η τεχνική leave-one-out.

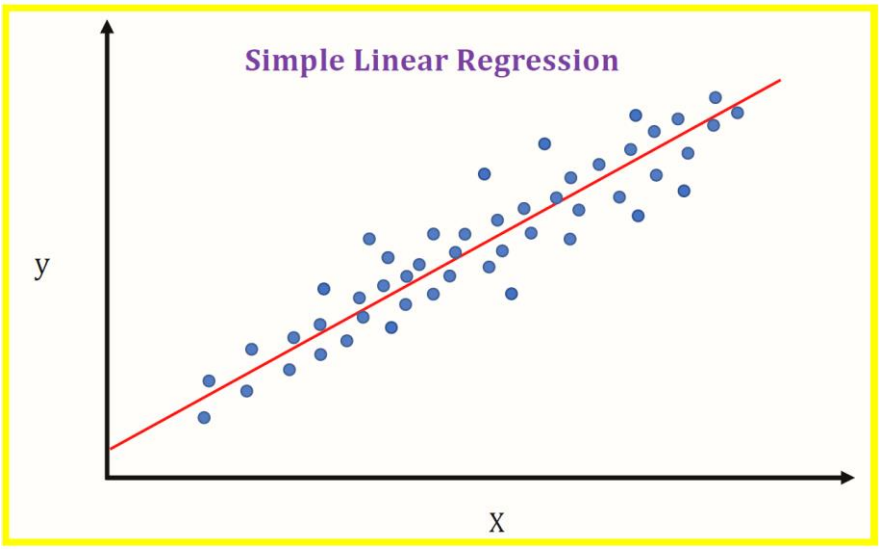
Ένα σημαντικό μειονέκτημα της διασταυρούμενης επικύρωσης είναι ότι ο αριθμός των δοκιμών που πρέπει να εκτελεστούν αυξάνεται κατά έναν συντελεστή S, και αυτό μπορεί να αποδειχθεί προβληματικό για μοντέλα στα οποία η εκπαίδευση είναι από μόνη της υπολογιστικά ακριβή. Ένα άλλο πρόβλημα με τεχνικές όπως η διασταυρούμενη επικύρωση που χρησιμοποιούν ξεχωριστά δεδομένα για την αξιολόγηση της απόδοσης είναι ότι ενδέχεται να διαθέτουν πολλαπλές παραμέτρους πολυπλοκότητας για ένα μεμονωμένο μοντέλο (για παράδειγμα, ενδέχεται να υπάρχουν αρκετές παράμετροι κανονικοποίησης). Η διερεύνηση συνδυασμών ρυθμίσεων για τέτοιες παραμέτρους θα μπορούσε, στη χειρότερη περίπτωση, να απαιτεί έναν αριθμό εκπαιδευτικών εκτελέσεων που είναι εκθετικός στον αριθμό των παραμέτρων. Καθίσταται λοιπόν σαφές ότι απαιτείται μια καλύτερη προσέγγιση. Στην ιδανική περίπτωση, αυτό θα έπρεπε να βασίζεται μόνο στα δεδομένα εκτέλεσης και να επιτρέπει τη σύγκριση πολλαπλών υπερπαραμέτρων και τύπων μοντέλου σε μία μόνο εκτέλεση. Πρέπει λοιπόν να βρεθεί ένα μέτρο απόδοσης που εξαρτάται μόνο από τα δεδομένα εκπαίδευσης.



Σχήμα: Η τεχνική της διασταυρούμενης επικύρωσης S-fold, που απεικονίζεται εδώ για την περίπτωση του S = 4, περιλαμβάνει τη λήψη των διαθέσιμων δεδομένων και την κατανομή τους σε ομάδες S (στην απλούστερη περίπτωση αυτές είναι ίσου μεγέθους) Στη συνέχεια, το S - 1 των ομάδων χρησιμοποιείται για να εκπαιδεύσει ένα σύνολο μοντέλων που στη συνέχεια αξιολογούνται στην υπόλοιπη ομάδα. Η διαδικασία αυτή επαναλαμβάνεται για όλες τις πιθανές επιλογές S για την ομάδα που έχει παραμείνει, που υποδεικνύεται εδώ από τα κόκκινα μπλοκ και στη συνέχεια τα αποτελέσματα των επιδόσεων από τις διαδρομές S υπολογίζονται κατά μέσο όρο.

4.2. Γραμμική Παλινδρόμηση

Ο στόχος της παλινδρόμησης είναι να προβλεφθεί η τιμή μιας ή περισσότερων συνεχών μεταβλητών στόχου t δεδομένης της τιμής ενός D-διαστάσεων διανύσματος X των μεταβλητών εισόδου. Το πολυώνυμο είναι ένα συγκεκριμένο παράδειγμα μιας ευρείας κατηγορίας συναρτήσεων που ονομάζονται μοντέλα γραμμικής παλινδρόμησης, τα οποία μοιράζονται την ιδιότητα να είναι γραμμικές συναρτήσεις των ρυθμιζόμενων παραμέτρων. Η απλούστερη μορφή μοντέλων γραμμικής παλινδρόμησης είναι επίσης γραμμικές συναρτήσεις των μεταβλητών εισόδου. Ωστόσο, μπορεί να αναπτυχθεί μια πιο χρήσιμη κατηγορία συναρτήσεων λαμβάνοντας γραμμικούς συνδυασμούς ενός σταθερού συνόλου μη γραμμικών συναρτήσεων των μεταβλητών εισόδου, γνωστών ως βασικών συναρτήσεων. Τέτοια μοντέλα είναι γραμμικές συναρτήσεις των παραμέτρων, οι οποίες τους παρέχουν απλές αναλυτικές ιδιότητες, αλλά μπορεί να είναι μη γραμμικές σε σχέση με τις μεταβλητές εισόδου.



Έστω ένα εκπαιδευτικό συνόλο δεδομένων που περιλαμβάνει παρατηρήσεις N $\{x_n\}$, όπου n = 1,. . . , N, μαζί με τις αντίστοιχες τιμές-στόχους $\{t_n\}$, ο στόχος είναι να προβλεφθεί η τιμή t για μια νέα τιμή x. Στην απλούστερη προσέγγιση, αυτό μπορεί να γίνει δημιουργώντας απευθείας μια κατάλληλη συνάρτηση y(x) της οποίας οι τιμές για νέες εισόδους x αποτελούν τις προβλέψεις για τις αντίστοιχες τιμές του t. Γενικότερα, από μια πιθανοτική προοπτική, στόχος είναι να μοντελοποιηθεί η προγνωστική κατανομή p(t|x) επειδή αυτό εκφράζει την αβεβαιότητά σχετικά με την τιμή του t για κάθε τιμή του x. Από αυτήν την υπό όρους κατανομή είναι δυνατόν να γίνουν προβλέψεις του t, για οποιαδήποτε νέα τιμή του x, με τέτοιο τρόπο ώστε να ελαχιστοποιείται η αναμενόμενη τιμή μιας κατάλληλα επιλεγμένης συνάρτησης απώλειας. Μια κοινή επιλογή συνάρτησης απώλειας για μεταβλητές πραγματικής αξίας είναι η τετραγωνική απώλεια, για την οποία η βέλτιστη λύση δίνεται από την υπό όρους προσδοκία του t.

Αν και τα γραμμικά μοντέλα έχουν σημαντικούς περιορισμούς ως πρακτικές τεχνικές για την αναγνώριση προτύπων, ιδιαίτερα για προβλήματα που αφορούν χώρους εισόδου υψηλών διαστάσεων, έχουν ωραίες αναλυτικές ιδιότητες και αποτελούν τη βάση για πιο εξελιγμένα μοντέλα.

Το πιο απλό γραμμικό μοντέλο για παλινδρόμηση είναι αυτό που περιλαμβάνει γραμμικό συνδυασμό για μεταβλητές εισόδου:

$$y(x, w) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

Όπου $x = (x_1 \dots x_D)^T$, το οποίο είναι γνωστό ως γραμμική παλινδρόμηση

Χρησιμοποιώντας μη γραμμικές συναρτήσεις βάσης, επιτρέπεται στη συνάρτηση y (x, w) να είναι μη γραμμική συνάρτηση του διανύσματος εισόδου x. Οι συναρτήσεις ονομάζονται γραμμικά μοντέλα επειδή αυτή η συνάρτηση είναι γραμμική σε w. Αυτή η γραμμικότητα στις παραμέτρους θα απλοποιήσει σημαντικά την ανάλυση αυτής της κατηγορίας μοντέλων.

4.3. Data Augmentation

Η ταξινόμηση κειμένου αποτελεί θεμελιώδες καθήκον στην επεξεργασία φυσικής γλώσσας (NLP). Τόσο η μηχανική όσο και η βαθιά μάθηση έχουν επιτύχει υψηλή ακρίβεια σε εργασίες που κυμαίνονται από την ανάλυση συναισθημάτων (Tang et al., 2015) έως την ταξινόμηση θεμάτων (Tong and Koller, 2002). Ωστόσο η υψηλή απόδοση συχνά εξαρτάται από το μέγεθος και την ποιότητα των δεδομένων εκπαίδευσης, τα οποία είναι συχνά κουραστικά στη συλλογή. Η αυτόματη αύξηση δεδομένων χρησιμοποιείται συνήθως στην όραση του υπολογιστή (Simard et al., 1998; Krizhevsky et al., 2017) και στην ομιλία (Cui et al., 2015; Ko et al., 2015) και μπορεί να βοηθήσει στην εκπαίδευση πιο ισχυρών μοντέλων, ειδικά όταν χρησιμοποιούνται μικρότερα σύνολα δεδομένων. Ωστόσο, επειδή η κατάληξη σε γενικευμένους κανόνες για τον μετασχηματισμό της γλώσσας είναι δύσκολη, οι καθολικές τεχνικές αύξησης δεδομένων στο NLP δεν έχουν διερευνηθεί διεξοδικά.

Προηγούμενη εργασία έχει προτείνει ορισμένες τεχνικές για την αύξηση των δεδομένων στο NLP. Μια δημοφιλής μελέτη δημιούργησε νέα δεδομένα μεταφράζοντας προτάσεις σε γαλλικά και πάλι πίσω στα αγγλικά (Yu et al., 2018). Άλλες εργασίες έχουν χρησιμοποιήσει δεδομένα θορύβου ως εξομάλυνση (Xie et al., 2017) και μοντέλα προγνωστικής γλώσσας για αντικατάσταση συνωνύμων (Kobayashi, 2018). Αν και αυτές οι τεχνικές είναι έγκυρες, δεν χρησιμοποιούνται συχνά στην πράξη, επειδή έχουν υψηλό κόστος εφαρμογής σε σχέση με το κέρδος απόδοσης

Λειτουργία	*Πρόταση*
None	*A sad, superior human comedy played out on the back roads of life*
SR	*A lamentable, superior human comedy played out on the backward road of life.*
RI	*A sad, superior human comedy played out on funniness the back roads of life.*
RS	*A sad, superior human comedy played out on roads back the of life.*
RD	*A sad, superior human out on the roads of life.*

Πίνακας: Προτάσεις που δημιουργήθηκαν με τη χρήση SR(αντικατάσταση συνωνύμου), RI(Τυχαία εισαγωγή), RS(Τυχαία εναλλαγή) και RD(τυχαία διαγραφή)

Ένα απλό σύνολο τεχνικών γενικής αύξησης δεδομένων για το NLP ονομάζεται EDA (εύκολη αύξηση δεδομένων).

Για μια συγκεκριμένη πρόταση στο σετ εκπαίδευσης, επιλέγεται και εκτελείται τυχαία μία από τις ακόλουθες λειτουργίες:

- **Αντικατάσταση συνωνύμου (Synonym Replacement-SR):** Επιλέγονται τυχαία n λέξεις από την πρόταση που δεν είναι λέξεις κλειδιά. Ακολουθεί αντικατάσταση κάθε μίας από αυτές τις λέξεις με ένα από τα συνώνυμα που επιλέγονται τυχαία.
- **Τυχαία εισαγωγή (Random Insertion-RI):** Πρέπει να βρεθεί ένα τυχαίο συνώνυμο μιας τυχαίας λέξης στην πρόταση που δεν είναι μια λέξη κλειδί. Ακολουθεί εισαγωγή αυτού του συνωνύμου σε μια τυχαία θέση στην πρόταση. Η διαδικασία αυτή επαναλαμβάνεται n φορές.
- **Τυχαία εναλλαγή (Random Swap-RS):** Πραγματοποιείται τυχαία επιλογή δύο λέξεων μέσα στην πρόταση και ανταλλαγή των θέσεών τους. Η διαδικασία αυτή επαναλαμβάνεται n φορές..
- **Τυχαία διαγραφή (Random Deletion-RD):** Πραγματοποιείται τυχαία αφαίρεση κάθε λέξης στην πρόταση με πιθανότητα p.

Βιβλιογραφία

A. S. Sabitha and D. Mehrotra, "User centric retrieval of learning objects in LMS," in *Proc. 2012 Third Int. Conf. Comput. Commun. Technol.*, Nov. 2012, pp. 14-19

A. Salazar, J. Gosalbez, I. Bosch, R. Miralles, and L. Vergara, "A case study of knowledge discovery on academic achievement, student desertion and student retention," in *Proc. ITRE 2004 2nd Int. Conf. Inf. Technol. Res. Educ.*, 2004, pp. 150-154.

A. Stes and P. Van Petegem, "Profiling approaches to teaching in higher education: A cluster-analytic study," *Stud. in High. Educ.*, vol. 39, issue 4, pp. 1-15, 2014.

Adamiak, C., Szyda, B., Dubownik, A., & García-Álvarez, D. (2019). Airbnb offer in Spain—spatial analysis of the pattern and determinants of its distribution. *ISPRS International Journal of Geo-Information*, *8*(3), 155.

Aggarwal, C. C. (2018). Neural networks and deep learning. *Springer*, *10*, 978-3.

Auguie, B., & Antonov, A. (2017). gridExtra: miscellaneous functions for "grid" graphics. *R package version*, *2*(1).

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157-166.

Bo, G., & Xianwu, H. (2006). SVM multi-class classification. *Journal of Data Acquisition & Processing*, *21*(3), 334-339.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152).

Breiman, L. (1999). Random Forests; UC Berkeley TR567. *University of California: Berkeley, CA, USA*.

Caropreso, M. F., & Matwin, S. (2006, June). Beyond the bag of words: A text representation for sentence selection. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 324-335). Springer, Berlin, Heidelberg.

Chapelle, O., & Zien, A. (2005, January). Semi-supervised classification by low density separation. In *AISTATS* (Vol. 2005, pp. 57-64).

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., ... & Ma, J. (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena, 151, 147-160.

Coles, P. A., Egesdal, M., Ellen, I. G., Li, X., & Sundararajan, A. (2017). Airbnb usage across New York City neighborhoods: Geographic patterns and regulatory implications. *Forthcoming, Cambridge Handbook on the Law of the Sharing Economy*.

Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(9), 1469-1477.

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. *International Journal of Computer Applications*, *28*(2), 37-40.

Dhuliawala, S., Kanojia, D., & Bhattacharyya, P. (2016, May). Slangnet: A wordnet like resource for english slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4329-4332).

Dolnicar, S. (2019). A review of research into paid online peer-to-peer accommodation: Launching the Annals of Tourism Research Curated Collection on peer-to-peer accommodation. *Annals of Tourism Research*, *75*, 248-264.

Dou, J., Yamagishi, H., Zhu, Z., Yunus, A. P., & Chen, C. W. (2018). TXT-tool 1.081-6.1 A comparative study of the binary logistic regression (BLR) and artificial neural network (ANN) models for GIS-based spatial predicting landslides at a regional scale. In *Landslide dynamics: ISDR-ICL landslide interactive teaching tools* (pp. 139-151). Springer, Cham.

Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, *5*(2), 112.

Dziadek, J., Henriksson, A., & Duneld, M. (2017). Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, *235*, 241.

F. Jing and K. Shiying, "Application of data mining for emotional intelligence based on cluster analysis," in *Proc. 2010 Int. Conf. Artif. Intell. Educ.*, Oct. 2010, pp. 512-515.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Gowda, H. S., Suhil, M., Guru, D. S., & Raju, L. N. (2016, December). Semi-supervised text categorization using recursive K-means clustering. In *International Conference on Recent Trends in Image Processing and Pattern Recognition* (pp. 217-227). Springer, Singapore.

Gupta, G., & Malhotra, S. (2015). Text document tokenization for word frequency count using rapid miner (taking resume as an example). *International Journal of Computer Applications*, *975*, 8887.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current issues in Tourism*, *18*(12), 1192-1217.

Guttentag, D. (2019). Progress on Airbnb: a literature review. *Journal of Hospitality and Tourism Technology*.

Han, E. H. S., & Karypis, G. (2000, September). Centroid-based document classification: Analysis and experimental results. In *European conference on principles of data mining and knowledge discovery* (pp. 424-431). Springer, Berlin, Heidelberg.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Harrell, F. E. (2001). Regression modeling strategies, with applications to linear models, survival analysis and logistic regression. *GET ADDRESS: Springer*.

Hassanli, N., Small, J., & Darcy, S. (2019). The representation of Airbnb in newspapers: a critical discourse analysis. *Current Issues in Tourism*, 1-13

Helm, A. (2003). Recovery and reclamation: A pilgrimage in understanding who and what we are. *Psychiatric and mental health nursing: The craft of caring*, 50-55.

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, *63*(322), 677-691.

Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299-310.

Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), 61-70.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*(1), 1503-1509.

K. Govindarajan, T. S. Somasundaram, and V. S. Kumar, "Particle swarm optimization (pso)-based clustering for improving the quality of learning using cloud computing," in *Proc. 2013 IEEE 13th Int. Conf. Adv. Learn. Technol.*, Jul. 2013, pp. 495-497.

Kastrinos, F., Steyerberg, E. W., Mercado, R., Balmaña, J., Holter, S., Gallinger, S., ... & Thibodeau, S. N. (2011). The PREMM1, 2, 6 model predicts risk of MLH1, MSH2, and MSH6 germline mutations based on cancer history. *Gastroenterology*, *140*(1), 73-81.

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, *18*(11), 1457-1466.

Kim, Y. H., Hahn, S. Y., & Zhang, B. T. (2000, July). Text filtering by boosting naive Bayes classifiers. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 168-175).

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Kowsari, K. (2014). *Investigation of fuzzyfind searching with golay code transformations* (Doctoral dissertation, M. Sc. Thesis, The George Washington University, Department of Computer Science).

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017, December). Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (pp. 364-371). IEEE.

Kowsari, K., Yammahi, M., Bari, N., Vichr, R., Alsaby, F., & Berkovich, S. Y. (2015). Construction of fuzzyfind dictionary using golay coding transformation for searching applications. *arXiv preprint arXiv:1503.06483*.

Krishnapuram, B., Carin, L., Figueiredo, M. A., & Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, *27*(6), 957-968.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., & Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal*, *56*(4), 534-563.

Kruppa, J., Liu, Y., Diener, H. C., Holste, T., Weimar, C., König, I. R., & Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal*, *56*(4), 564-583.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Larson, R. R. (2010). Introduction to information retrieval.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature 521 (7553), 436-444. *Google Scholar Google Scholar Cross Ref Cross Ref*.

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, *17*(12), 1131-1142.

Lock, G. (2002). Acute mesenteric ischemia: classification, evaluation and therapy. *Acta gastro-enterologica Belgica*, *65*(4), 220-225.

Manevitz, L. M., & Yousef, M. (2001). One-class SVMs for document classification. *Journal of machine Learning research*, *2*(Dec), 139-154.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442-451.

Mawardi, V. C., Susanto, N., & Naga, D. S. (2018). Spelling correction for text documents in Bahasa Indonesia using finite state automata and Levinshtein distance method. In *MATEC Web of Conferences* (Vol. 164, p. 01047). EDP Sciences.

N. A. Rashid, M. N. Taib, S. Lias, and N. Sulaiman, "Classification of learning style based on Kolb‟s learning style inventory and EEG using cluster analysis approach," in *Proc. 2010 2nd Int. Congr. On Eng. Educ. (ICEED)*, pp. 64-68, 2010.

Nigam, K., McCallum, A., & Mitchell, T. M. (2006). Semi-Supervised Text Classification Using EM.

P. Moreno-Clari, M. Arevalillo-Herraez, and V. Cerveron-Lleo, “Data analysis as a tool for optimizing learning management systems,” in *Proc. Ninth IEEE Int. Conf. Adv. Learn. Technol.*, Jul. 2009, pp. 242-246.

Pahwa, B., Taruna, S., & Kasliwal, N. (2018). Sentiment Analysis-Strategy for Text Pre-Processing. *Int. J. Comput. Appl*, *180*, 15-18.

Pebesma, E. J. (2018). Simple features for R: Standardized support for spatial vector data. *R J.*, *10*(1), 439..

Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, *27*(2), 157-172.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*.

Qu, Z., Song, X., Zheng, S., Wang, X., Song, X., & Li, Z. (2018, January). Improved Bayes method based on TF-IDF feature and grade factor feature for chinese information classification. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 677-680). IEEE.

Romero, C., Ventura, S., Delgado, J. A., & De Bra, P. (2007, September). Personalized links recommendation based on data mining in adaptive educational hypermedia systems. In *European conference on technology enhanced learning* (pp. 292-306). Springer, Berlin, Heidelberg.

Royston, P., & Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables* (Vol. 777). John Wiley & Sons.

S. V Lahane, M. U. Kharat, and P. S. Halgaonkar, “Divisive approach of clustering for educational data,” in *Proc. 2012 Fifth Int. Conf. Emerg. Trends Eng. Technol. (ICETET 2012)*, 2012, pp. 191- 195.

S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, "A clustering methodology of web log data for learning management systems," *Educ. Technol. Soc.*, vol. 15, no. 2, pp. 154- 167, 2012.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Sampson, G. (2005). *The'Language Instinct'Debate: Revised Edition*. A&C Black.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine learning*, *39*(2), 135-168.

Shen, D., Sun, J. T., Li, H., Yang, Q., & Chen, Z. (2007, January). Document summarization using conditional random fields. In *IJCAI* (Vol. 7, pp. 2862-2867).

Shi, L., Mihalcea, R., & Tian, M. (2010, October). Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1057-1067).

Simard, P. Y., LeCun, Y. A., Denker, J. S., & Victorri, B. (1998). Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade* (pp. 239-274). Springer, Berlin, Heidelberg.

Singh, J., & Gupta, V. (2016). Text stemming: Approaches, applications, and challenges. *ACM Computing Surveys (CSUR)*, *49*(3), 1-46.

Slee, T. (2018). Data collection for Airbnb listings.

Spirovski, K., Stevanoska, E., Kulakov, A., Popeska, Z., & Velinov, G. (2018, June). Comparison of different model's performances in task of document classification. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-12).

Steyerberg, E. W., Eijkemans, M. J., Harrell Jr, F. E., & Habbema, J. D. F. (2000). Prognostic modelling with logistic regression analysis: a comparison of

selection and estimation methods in small data sets. *Statistics in medicine*, *19*(8), 1059-1079.

Sutskever, I., Martens, J., & Hinton, G. E. (2011, January). Generating text with recurrent neural networks. In *ICML*.

Tang, D., Qin, B., & Liu, T. (2015, September). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422-1432).

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, *2*(Nov), 45-66.

V. Subbian, "Role of MOOCs in integrated STEM education: A learning perspective," in *Proc. 2013 IEEE Integr. STEM Educ. Conf.*, Mar. 2013, pp. 1-4.

Van Calster, B., Condous, G., Kirk, E., Bourne, T., Timmerman, D., & Van Huffel, S. (2009). An application of methods for the probabilistic three-class classification of pregnancies of unknown location. *Artificial intelligence in medicine*, *46*(2), 139-154

Vapnik, V., & Chervonenkis, A. Y. (1964). A class of algorithms for pattern recognition learning. *Avtomat. i Telemekh*, *25*(6), 937-945.

Verma, T., Renu, R., & Gaur, D. (2014). Tokenization and filtering process in RapidMiner. *International Journal of Applied Information Systems*, *7*(2), 16-18.

Whitney, D. L., & Evans, B. W. (2010). Abbreviations for names of rock-forming minerals. *American mineralogist*, *95*(1), 185-187.

Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*(Aug), 975-1005.

Xu, B., Ye, Y., & Nie, L. (2012, June). An improved random forest classifier for image classification. In *2012 IEEE International Conference on Information and Automation* (pp. 795-800). IEEE.

Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Zaiane, O. (2001). Web usage mining for a better web-based learning environment.

Zaíane, O. R. (2002, December). Building a recommender agent for e-learning systems. In *International Conference on Computers in Education, 2002. Proceedings.* (pp. 55-59). IEEE.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*(3), 1169-1180.

Zhou, S., Chen, Q., & Wang, X. (2014). Fuzzy deep belief networks for semi-supervised sentiment classification. *Neurocomputing*, *131*, 312-322.