ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Δ.Π.Μ.Σ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Εξαμηνιαία Εργασία

Τεχνητά Νευρωνικά Δίκτυα και Μηχανική Μάθηση

Ανδριοσοπούλου Γεωργία (03400042)
Ζωγραφάκης Δημήτριος (03400050)
Μαστακούρης Ανδρέας (03400062)

Αθήνα,
15/03/2020

Περιεχόμενα

Εισαγωγή

1.1 Ακριβής Διατύπωση Προβλήματος

Στα πλαίσια του παρόντος μαθήματος ζητείται η επιλογή ενός συνόλου δεδομένων και η ανάλυσή του εφαρμόζοντας αλγορίθμους μηχανικής μάθησης, με στόχο την εξαγωγή γνώσης από τα δεδομένα αυτά. Ύστερα από τη σχετική αναζήτηση των διαθέσιμων δεδομένων, το σύνολο που επιλέγεται ονομάζεται "What's Cooking?", το οποίο δημιουργήθηκε στην ιστοσελίδα Kaggle. Όσον αφορά το περιεχόμενό του, το υπό εξεταζόμενο σύνολο δεδομένων περιλαμβάνει καταγραφές συνταγών με αρκετά διαφορετικά υλικά και για κάθε μια συνταγή παρέχεται ως επισημείωση η αντίστοιχη κουζίνα από την οποία προέρχεται. Συνεπώς, το πρόβλημα που τίθεται είναι ένα πρόβλημα ταξινόμησης κάθε συνταγής στην αντίστοιχη κουζίνα προέλευσης της.

Στην παρούσα εργασία λοιπόν πραγματοποιείται επεξεργασία των πληροφοριών που προσφέρει το συγκεκριμένο σύνολο δεδομένων, οργανώνοντας τα επιμέρους τμήματα της ανάλυσης ως εξής: στο Κεφάλαιο 1 παρουσιάζεται το πρόβλημα που καλείται να αντιμετωπίσει η συγκεκριμένη εργασία και καθορίζονται επίσης οι στόχοι της ανάλυσης. Στο Κεφάλαιο 2 πραγματοποιείται διερευνητική ανάλυση των δεδομένων με κατασκευή γραφημάτων και εξηγείται ταυτόχρονα η προεπεξεργασία που πραγματοποιείται στα δεδομένα. Στο Κεφάλαιο 3 ακολουθεί παρουσίαση ορισμένων αλγορίθμων μηχανικής μάθησης μαζί με τα αντίστοιχα αποτελέσματα που επιτυγχάνουν για σύγκριση των αποδόσεων τους. Παράλληλα πραγματοποιείται μείωση της διαστατικότητας των δεδομένων - λόγω του μεγάλου αριθμού χαρακτηριστικών που δημιουργούνται - και εξετάζεται πώς η τεχνική αυτή επηρεάζει την απόδοση των αλγορίθμων. Τέλος, στο Κεφάλαιο 4 παρουσιάζονται συγκεντρωτικά τα συμπεράσματα όπως αυτά προέκυψαν από όλη τη διαδικασία της ανάλυσης και αναφέρονται ενδεικτικά τρόποι επέκτασης των αποτελεσμάτων.

1.2 Στόχοι Εργασίας

Το σύνολο δεδομένων "What's Cooking?" έχει προκύψει από την καταγραφή 39.774 διαφορετικών συνταγών για 20 διαφορετικές κουζίνες (Ελληνική, Ιταλική κ.ο.κ). Στα πλαίσια τη συγκεκριμένης εργασίας τίθονται λοιπόν οι εξής στόχοι:

1. Κατανόηση και αποσαφίνηση των δεδομένων που προσφέρονται, το οποίο επιτυγχάνεται με την κατασκευή ποικίλων γραφημάτων. Στο στάδιο αυτό ανήκει και η προεπεξεργασία του συνόλου δεδομένων.

2. Εφαρμογή αλγορίθμων μηχανικής μάθησης και σύγκριση των αποτελεσμάτων τους. Συγκεκριμένα οι μέθοδοι που επιστρατεύονται είναι: Πολυωνυμικός Αφελής Ταξινομητής (Multinomial Naive Bayes), αλγόριθμος k-κοντινότερου γείτονα (k-NN), Λογιστική Παλινδρόμηση (Logistic Regression), Μηχανές Υποστηριστικών Διανυσμάτων (Support Vector Machine), Δένδρα Απόφασης (Decision Trees) και Τυχαία Δάση (Random Forests).

3. Μείωση της διαστατικότητας των δεδομένων υλοποιώντας την τεχνική Ανάλυση σε Κύριες Συνιστώσες (Principal Components Analysis) και εκτίμηση της επιρροής της συγκεκριμένης τεχνικής στην απόδοση των αλγορίθμων μηχανικής μάθησης.

Διερευνητική Ανάλυση

2.1 Προεπεξεργασία Συνόλου Δεδομένων

Το σύνολο δεδομένων που επιλέγεται προέρχεται από το διαγωνισμό του Kaggle το έτος 2018. Το μέγεθος ανέρχεται στις 39.774 καταγραφές και για κάθε μία καταγραφή αναγράφονται τα συστατικά που απαιτούνται για την εκτέλεση της συνταγής, όπως επίσης και η χώρα προέλευσης της. Τα στοιχεία που είναι διαθέσιμα και η μορφή αυτών περιγράφονται στον Πίνακα 2.1:

Πίνακας 2.1: Περιγραφή του Συνόλου Δεδομένων.

Στήλη	Περιγραφή	Τύπος
id	Μοναδικό κλειδί Αναγνώρισης	Φυσικός Αριθμός
cuisine	Χώρα Προέλευσης Συνταγής	Συμβολοσειρά
ingredients	Υλικά της Συνταγής	Συμβολοσειρά

Ακολούθως πραγματοποιείται καθαρισμός του συνόλου δεδομένων, ώστε να είναι έτοιμα για τη μετέπειτα επεξεργασία τους. Συγκεκριμένα τα βήματα που ακολουθούνται είναι τα εξής:

1. Μετατροπή όλων των κεφαλαίων γραμμάτων σε πεζά ώστε η διαφορά του πρώτου γράμματος να μη δημιουργεί διαφορετικές λέξεις.

2. Λημματοποίηση (συνάρτηση Lemmatize του ntlk) των δεδομένων, δηλαδή διατήρηση της ριζικής προέλευσης κάθε λέξης. Με αυτόν τον τρόπο λέξεις με την ίδια σημασία δε θεωρούνται διαφορετικές (π.χ. οι λέξεις eggs και egg αντιμετωπίζονται ισοδύναμα) και κατά συνέπεια το ποσοστό σωστής ταξινόμησης των συνταγών μπορεί να βελτιωθεί [1].

3. Για τον ίδιο ακριβώς λόγο πραγματοποιείται απομάκρυνση αριθμητικών συμβόλων (π.χ 1%) και σημείων στίξης εκτός του κόμματος, το οποίο χρησιμοποιείται για το διαχωρισμό των υλικών σε κάθε συνταγή.

4. Αφαίρεση καταγραφών με λιγότερα από 2 υλικά. Προφανώς, αν κάποια καταγραφή αποτελείται από μόνο ένα υλικό τότε δεν μπορεί να θεωρηθεί συνταγή και για το λόγο αυτό απομακρύνεται ως θόρυβος.

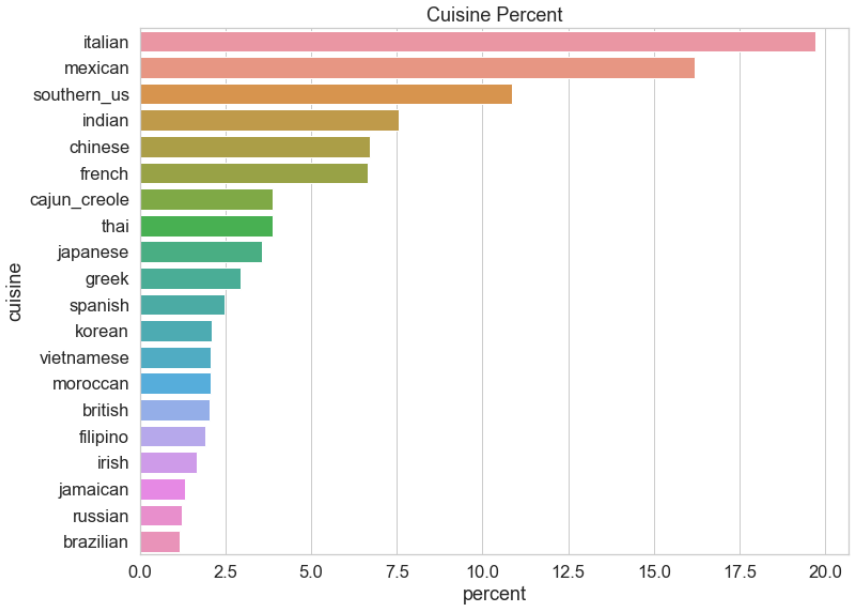
5. Έλεγχος ελλίπουσων τιμών στο σύνολο δεδομένων (ύστερα από έλεγχο δε βρέθηκε κάποια).

Προφανώς, η προσαρμογή των αλγορίθμων μηχανικής μάθησης στα δεδομένα απαιτεί την κατάλληλη κωδικοποίησή τους σε αριθμητικές μεταβλητές. Η διαδικασία αυτή επιτυγχάνεται μέσω της τεχνικής One Hot Encoding και η οποία παρουσιάζεται στην Ενότητα 3.1.

2.2 Παρουσίαση Γραφημάτων

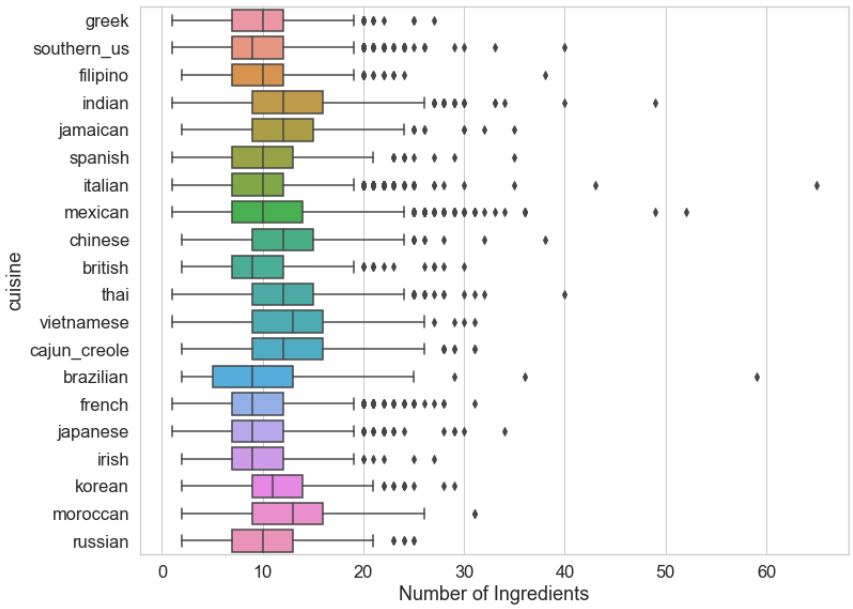
Στη συγκεκριμένη ενότητα παρατίθενται ορισμένα γραφήματα όπου σε πρωταρχικό στάδιο βοηθούν στην εξαγωγή γνώσης από τα συγκεκριμένα δεδομένα. Αρχικά, στο Σχήμα 2.1 παρουσιάζεται

η συχνότητα εμφάνισης των 20 διαφορετικών κατηγοριών κουζίνας στο σύνολο δεδομένων που εξετάζεται. Συγκεκριμένα, παρατηρείται ότι περίπου το 20% των συνολικών συνταγών ανήκουν στην ιταλική κουζίνα, η οποία αποτελεί την συχνότερα εμφανιζόμενη. Από το συγκεκριμένο γράφημα παρατηρείται επίσης η ύπαρξη συνταγών με ποσοστό εμφάνισης μικρότερο του 2% για συγκεκριμένες κουζίνες, όπως για παράδειγμα η Φιλλιπινέζικη, η Ρώσικη κ.ο.κ.



Σχήμα 2.1: Ποσοστό εμφάνισης κάθε κουζίνας στο σύνολο δεδομένων.

Ακολούθως, στο Σχήμα 2.2 κατασκευάζονται τα θηκογράμματα για το πλήθος των υλικών ανά κουζίνα. Τα γραφήματα αυτά απεικονίζουν ταυτόχρονα τη μικρότερη παρατήρηση του δείγματος, το πρώτο τεταρτημόριο (25%), τη διάμεσο, το τρίτο τεταρτημόριο (75%), καθώς επίσης και τη μέγιστη τιμή του δείγματος. Οι αποστάσεις μεταξύ των διαφόρων τμημάτων του θηκογράμματος αντικατοπτρίζουν τη διασπορά και την ασυμμετρία των δεδομένων. Επιπλέον, τυχόν παράτυπα σημεία του δείγματος (outliers) απεικονίζονται ως ξεχωριστά σημεία.

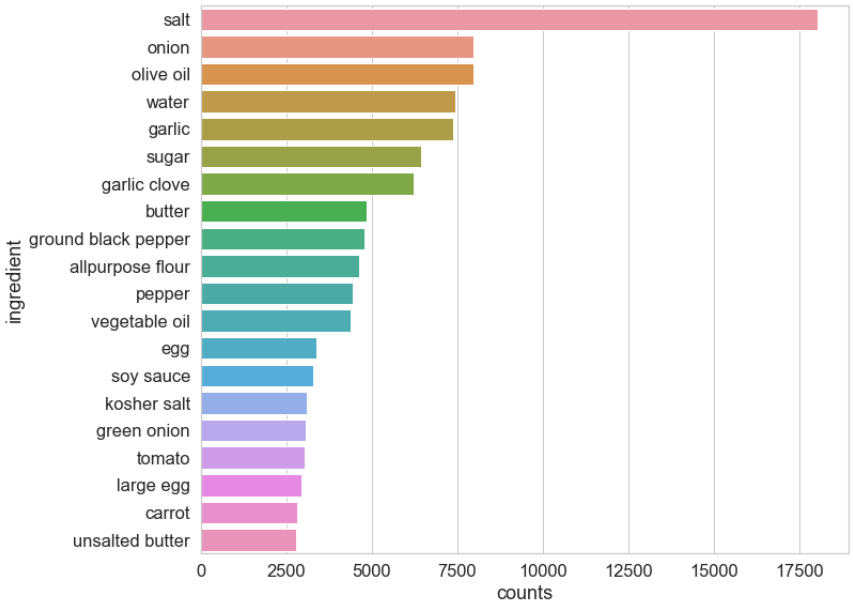


Σχήμα 2.2: Θηκογράμματα πλήθους υλικών ανά κουζίνα.

Σύμφωνα με τα αποτελέσματα του 2.2 παρατηρείται ότι η διάμεσος για όλες τις κουζίνες βρίσκεται στο εύρος από 9 έως 13. Ωστόσο παρατηρείται ύπαρξη συνταγών με πολύ μεγαλύτερο πλήθος

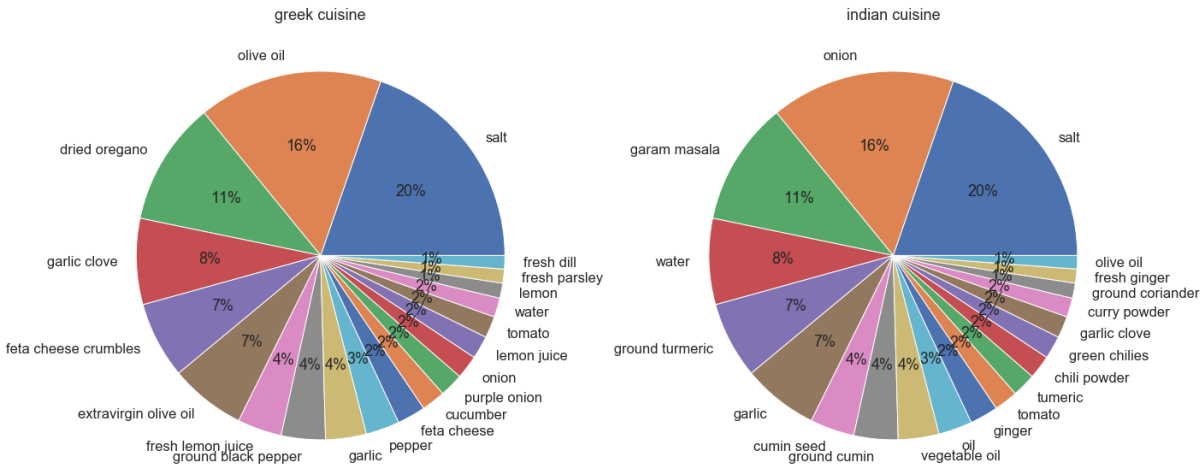
συστατικών, όπως για παράδειγμα στην περίπτωση μιας συγκεκριμένης ιταλικής συνταγής που χρησιμοποιεί περισσότερα από 60 υλικά.

Εν συνεχεία, στο Σχήμα 2.3 απεικονίζεται η συχνότητα εμφάνισης κάθε υλικού στο σύνολο των συνταγών που έχει καταγραφεί. Παρατηρείται λοιπόν ότι το πιο συχνό στοιχείο είναι το αλάτι με συχνότητα εμφάνισης σε περισσότερες από 17.500 συνταγές. Ακολουθούν τα κρεμμύδια, το λάδι, το νερό και το σκόρδο, τα οποία συμπεριλαμβάνονται σε περίπου 7.500 συνταγές.



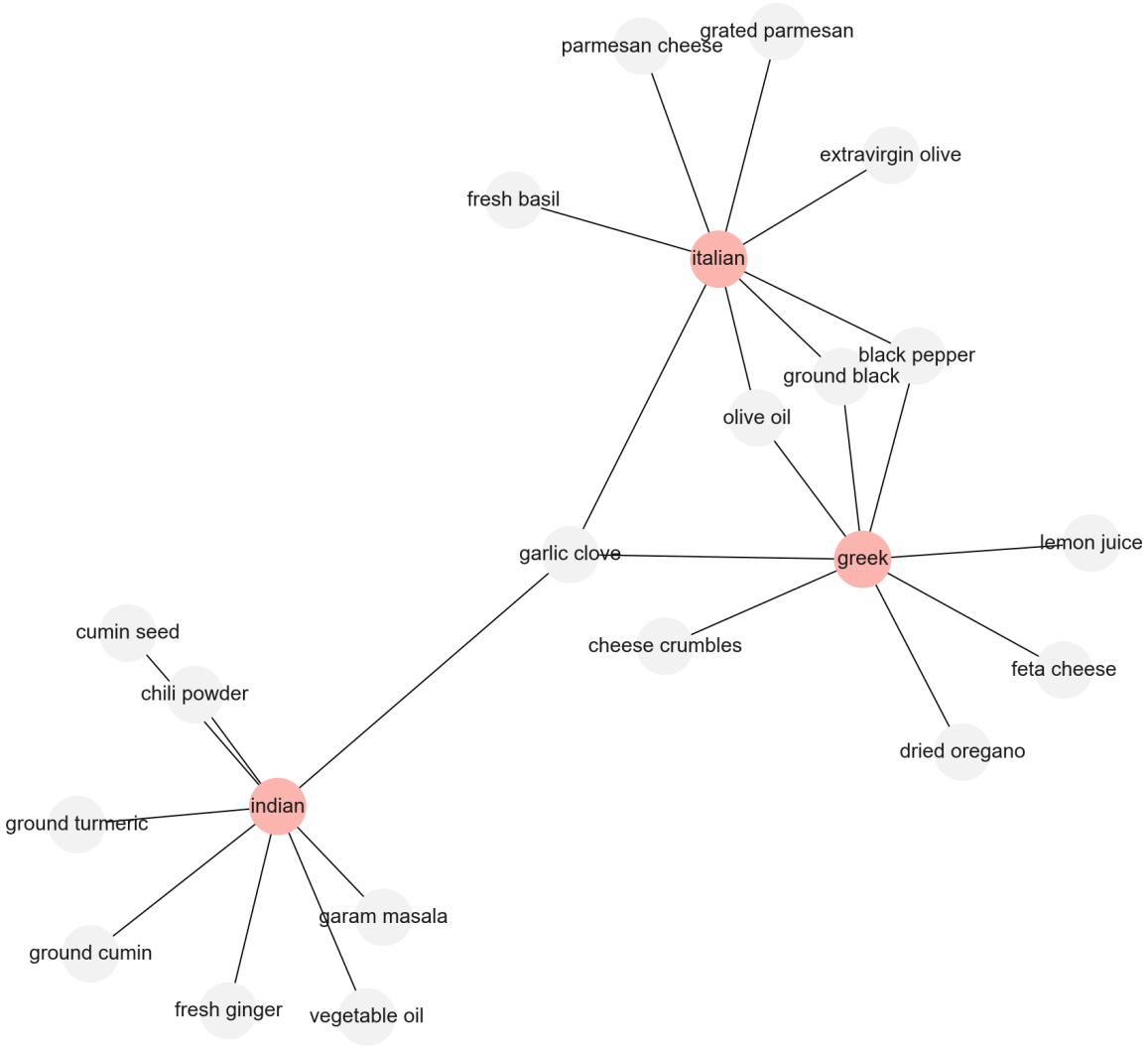
Σχήμα 2.3: Κορυφαία 20 υλικά στο Σύνολο Δεδομένων.

Επέκταση του παραπάνω γραφήματος αποτελεί η εύρεση του μέσου ποσοστού συμμετοχής για κάθε συστατικό ανά διαφορετικό είδος κουζίνας. Στο Σχήμα 2.4 αναπαριστάται ενδεικτικά αυτή η πληροφορία για την Ελληνική και την Ινδική κουζίνα, χρησιμοποιώντας γραφήματα τύπου πίτας. Από τις συγκεκριμένες αναπαραστάσεις εξάγονται τα συστατικά με τη μεγαλύτερη και τη μικρότερη συχνότητα εμφάνισης για τις 2 αυτές κουζίνες. Όμοια με πριν, παρατηρείται η έντονη παρουσία του αλατιού και στις 2 κουζίνες, με δεύτερο πιο συχνό συστατικό να είναι το ελαιόλαδου για την Ελληνική κουζίνα και τα κρεμμύδια για την Ινδική. Επιπρόσθετα, μπορούν να εντοπισθούν εκείνα τα συστατικά που χρησιμοποιούνται με λιγότερη συχνότητα στις 2 κουζίνες.



Σχήμα 2.4: Συχνότητα εμφάνισης υλικών για Ελληνική (αριστερά) και Ινδική (δεξιά) κουζίνα.

Τέλος, κατασκευάζεται ένα ενδιαφέρον γράφημα που απεικονίζει την σύνδεση μεταξύ των διαφορετικών κουζίνων. Στο Σχήμα 2.5 απεικονίζεται η συσχέτιση της Ελληνικής κουζίνας, της Ιταλικής και της Ινδικής, η οποία προκύπτει συγκρίνοντας το πλήθος των κοινών συστατικών τους. Για κάθε κουζίνα εξάγονται τα συχνότερα bigrams και δημιουργείται γράφημα με τα 8 πρώτα (διαλέχθηκε το 8 ως παράμετρος για λόγους αναγνωσιμότητας). Από το συγκεκριμένο γράφημα διαπιστώνεται η ικανοποιητική συσχέτιση της Ελληνικής με την Ιταλική κουζίνα, εφόσον είναι κοινά 4 διαφορετικά bigrams (black pepper, ground black, olive oil, cheese crumbles) από τα επιμέρους 8 που εξετάστηκαν. Αντιθέτως, η συσχέτιση των κουζίνων αυτών με την Ινδική δεν είναι τόσο έντονη, καθώς το μόνο κοινό bigram φαίνεται να είναι το garlic clove.



Σχήμα 2.5: Κοινά συστατικά Ελληνικής, Ιταλικής και Ινδικής Κουζίνας.

Μηχανική Μάθηση

3.1 Ανασκόπηση Ταξινομητών

Στο συγκεκριμένο κεφάλαιο προσαρμόζονται μέθοδοι μηχανικής μάθησης στο σύνολο δεδομένων που επιλέχθηκε, με στόχο την πρόβλεψη των κουζίνων από τις οποίες προέρχονται ορισμένες συνταγές με συγκεκριμένα μαγειρικά υλικά. Η εφαρμογή των μεθόδων απαιτεί αρχικά τη μετατροπή του συνόλου δεδομένων με τρόπο ώστε κάθε ένα υλικό να αποτελεί ένα ξεχωριστό χαρακτηριστικό και στη συνέχεια την κωδικοποίηση των χαρακτηριστικών αυτών σε αριθμητικές μεταβλητές. Εφόσον τα χαρακτηριστικά αυτά αποτελούν αποκλειστικά κατηγορικές ονομαστικές (nominal) μεταβλητές, η κωδικοποίηση βασίζεται στη μέθοδο One Hot Encoding, δημιουργώντας συνολικά 2.845 χαρακτηριστικά.

Η κωδικοποίηση πραγματοποιείται με χρήση της συνάρτησης TfidfVectorizer (γλώσσα Python), η οποία ουσιαστικά εξάγει το κωδικοποιημένο σύνολο δεδομένων σε μορφή πίνακα ο οποίος, για το συγκεκριμένο πρόβλημα, είναι "αραιός". Με άλλα λόγια, αποτελείται από αρκετά μηδενικά εφόσον τα υλικά κάθε μίας εξεταζόμενης συνταγής - δηλαδή κάθε δείγματος - αποτελούν ένα αρκετά μικρό ποσοστό συγκριτικά με το σύνολο των υλικών που χρησιμοποιούνται από όλες τις κουζίνες. Εφόσον λοιπόν οι "αραιοί" πίνακες είναι από τη φύση τους συμπιεσμένοι (Compressed Sparse Row Matrix), η κωδικοποίηση αυτή λοιπόν συνεισφέρει αρκετά στη βελτίωση της ταχύτητας των αλγορίθμων που παρουσιάζονται ακολούθως.

Πριν την προσαρμογή των ταξινομητών στα κωδικοποιημένα χαρακτηριστικά, τα δεδομένα χωρίζονται με τυχαίο τρόπο σε δεδομένα εκπαίδευσης (train set, 80% των αρχικών δεδομένων) και δεδομένα ελέγχου (test set, 20% των αρχικών δεδομένων). Ακολούθως, οι αλγόριθμοι που προσαρμόζονται στα δεδομένα εκπαίδευσης και ελέγχου είναι οι εξής:

- Πολυωνυμικός Αφελής Ταξινομητής (Multinomial Naive Bayes)

  Ο αφελής ταξινομητής βασίζεται στην ιδέα εκτίμησης της αποστεριόρι πιθανότητας ενός δείγματος να ανήκει σε μία κλάση. Μια από τις βασικές υποθέσεις αποτελεί η ανεξαρτησία όλων των χαρακτηριστικών μεταξύ τους, το οποίο σαν υπόθεση μπορεί να θεωρηθεί αφελής, δεδομένου ότι σπάνια κάτι τέτοιο ισχύει. Ο αλγόριθμος αυτός βρίσκει ιδιαίτερη απήχηση σε προβλήματα ταξινόμησης κατηγορικών μεταβλητών και ιδιαίτετα ανάλυση κειμένου. Εν γένει, αποτελεί έναν γρήγορο και εύκολο αλγόριθμο.

- Κ-κοντινότερος Γείτονας (K-Nearest Neighbour)

  Ο συγκεκριμένος αλγόριθμος υποθέτει ότι παρόμοια αντίκειμενα βρίσκονται σε κοντικές αποστάσεις και κατά συνέπεια κάθε κλάση δημιουργεί τη δική της ξεχωριστή ομάδα. Για την εύρεση της ομοιότητας των χαρακτηριστικών κάθε κλάσης εκτιμάται η απόσταση μεταξύ αυτών και με τον τρόπο αυτό προκύπτουν εν συνεχεία τα επίπεδα απόφασης. Αποτελεί έναν εύκολο αλγόριθμο, αλλά επηρεάζεται σημαντικά από το μέγεθος του συνόλου δεδομένων.

- Λογιστική Παλινδρόμηση (Logistic Regression)

  Η λογιστική παλινδρόμηση αποτελεί έναν από τους πιο βασικούς και διάσημους αλγορίθμους για προβλήματα ταξινόμησης. Η μέθοδος αυτή, συγκριτικά με τη γραμμική παλινδρόμηση, χαρακτηρίζεται ως λιγότερη ευάλωτη σε παράτυπα σημεία και συνεπώς πιο εύρωστη. Χρησιμοποιεί σιγμοειδή συνάρτηση, την οποία προσπαθεί να προσαρμόσει στις παρατηρήσεις.

- Μηχανές Υποστηρικτικών Διανυσμάτων (Support Vector Machines)

  Οι μηχανές υποστηρικτικών διανυσμάτων προσπαθούν να βρουν ένα υπερεπίπεδο, το οποίο διαχωρίζει τα δεδομένα σε δύο κλάσεις με τον καλύτερο δυνατό τρόπο. Στην περίπτωση που το πρόβλημα υποθέτει παραπάνω από μία κλάσεις, τότε εξέταζεται κάθε μία κλάση ξεχωριστά έναντι όλων των υπολοίπων μαζί (one vs rest). Αποτελεί μία από τις πιο αποδοτικές μεθόδους, επιτυγχάνοντας υψηλά ποσοστά σωστής ταξινόμησης, αλλά δεν ενδείκνυται για μεγάλα σύνολα δεδομένων. Στην παρούσα εργασία ο συγκεκριμένος ταξινομητής χρησιμοποιείται με πυρήνα RBF. Τέλος, να τονισθεί ότι παρόλο που το σύνολο δεδομένων ίσως είναι αρκετά μεγάλο για αυτόν τον ταξινομητή, η κλειστή μορφή του συνόλου δεδομένων που δημιουργείται με χρήση της συνάρτησης TfidfVectorizer επιτρέπει τη χρησιμοποίησή του χωρίς μεγάλη καθυστέρηση για την εξαγωγή των προβλέψεων ταξινόμησης.

- Δένδρα Απόφασης (Decision Tree)

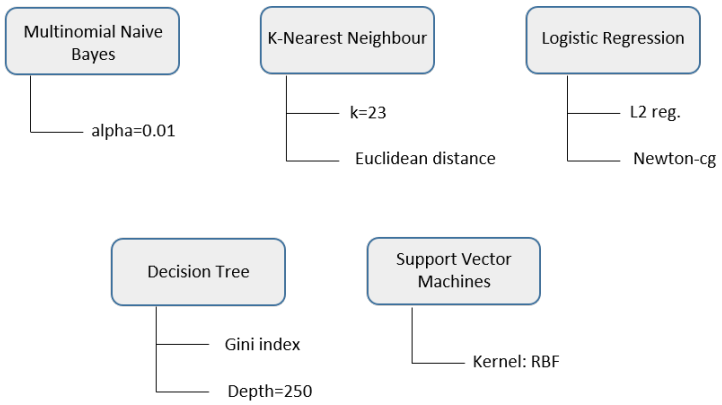
  Τα δένδρα απόφασης ανήκουν στις μη παραμετρικές μεθόδους επιβλεπόμενης μάθησης και προσπαθούν να κατασκευάσουν ένα μοντέλο με μορφή δέντρου. Διαχειρίζεται τόσο κατηγορικά όσο και αριθμητικά δεδομένα και έχουν αναπτυχθεί αρκετοί τρόποι για τον τρόπο εξαγωγής ενός δένδρου απόφασης. Στην παρούσα υλοποίηση ως κριτήριο χρησιμοποιείται ο συντελεστής ”gini”. Ένα από τα μειονεκτήματα του συγκεκριμένου αλγορίθμου είναι η τάση για υπερεκπαίδευση (overfitting), το οποίο όμως μπορεί να αποφευχθεί με ορισμένες μεθόδους (μείωση βάθους, τεχνική κλαδέματος κτλ).

- Τυχαία Δάση (Random Forests)

  Ο αλγόριθμος αυτός βασίζεται στην τεχνική bagging, όπου το σύνολο εκπαίδευσης χωρίζεται σε τυχαία υποσύνολα (με πιθανές επικαλύψεις) και σε κάθε ένα από αυτά εφαρμόζεται ο ταξινομητής των δένδρων απόφασης. Η τελική απόφαση προκύπτει μέσω ψηφοφορίας των προβλέψεων των επιμέρους ταξινομητών.

3.2 Υλοποίηση Ταξινομητών

Φάση Εκπαίδευσης. Στα πλαίσια της υλοποίησης ενός ταξινομητή απαιτείται η εύρεση εκείνου του συνδυασμού υπερπαραμέτρων που οδηγεί στην υψηλότερη απόδοση. Η διαδικασία αυτή μπορεί να εκτελεσθεί με τη βοήθεια της συνάρτησης ”grid search” της γλώσσας Python. Σε ορισμένες περιπτώσεις η μέθοδος αυτή έχει εξαιρετικά μεγάλο υπολογιστικό κόστος, αφού υπολογίζει τις αποδόσεις του μοντέλου για κάθε πιθανό συνδυασμό υπερπαραμέτρων. Ωστόσο, στο συγκεκριμένο πρόβλημα που εξετάζεται μπορεί να εφαρμοστεί χωρίς δυσκολία και τα αποτελέσματα για κάθε ένα μοντέλο παρουσιάζονται στο Σχήμα 3.1.



Σχήμα 3.1: Επιλεγμένες υπερπαράμετροι μοντέλων.

Οι επιλογές του Σχήματος 3.1 βασίζονται στην εξέταση διαφορετικού συνδυασμού υπερπαραμέτρων για κάθε ταξινομητή. Συγκεκριμένα, οι συνδυασμοί που εξετάζονται είναι οι εξής:

1. Αφελής ταξινομητής: Η μέθοδος ”grid search” εφαρμόζεται για ρύθμιση της υπερπαραμέτρου που υποδηλώνεται ως alpha. Ουσιαστικά η παράμετρος αυτή συνδέεται με την ικανότητα του ταξινομητή να κατηγοριοποιεί σωστά στο στάδιο εκπαίδευσης, όταν συναντά λέξεις που δεν έχει ξαναδεί. Οι τιμές του συντελεστή που εξετάζονται είναι: 1, 0.1, 0.01 και 0, με το βέλτιστο αποτέλεσμα να επιτυγχάνεται για alpha=0.01.

2. Ταξινομητής Κοντινότερου Γείτονα: Όσον αφορά τον ταξινομητή των k-NN, πραγματοποιείται σε πρώτο στάδιο εκτίμηση της ακρίβειας στο σύνολο ελέγχου όταν χρησιμοποιούνται k=15, 17, 19, 21, 23, 25 και 29 κοντινότεροι γείτονες. Τα αποτελέσματα απεικονίζονται στο Σχήμα 3.2 και παρατηρείται ότι το υψηλότερο ποσοστό ακρίβειας επιτυγχάνεται για k=23. Έπειτα, θεωρώντας k=23, εξετάζεται η επίδραση της απόστασης που χρησιμοποιεί ο αλγόριθμος και η σύγκριση αυτή πραγματοποιείται για τις αποστάσεις: ”euclidean” και ”manhattan”. Το βέλτιστο αποτέλεσμα προκύπτει όταν χρησιμοποιείται ως μετρική η Ευκλείδια απόσταση.

3. Λογιστική Παλινδρόμηση: Αναφορικά με τη λογιστική παλινδρόμηση συγκρίνονται οι αλγόριθμοι που επιλύουν το πρόβλημα βελτιστοποίησης και συγκεκριμένα οι ”newton-cg”, ”sag”, ”saga” και ”lbfgs”. Οι αλγόριθμοι αυτοί επιδέχονται ως όρο κανονικοποίησης (regularization) L2 (Ridge Regression) ή καθόλου. Επίσης, χρησιμοποιείται διασταυρούμενη επικύρωση 5 τμημάτων (5-fold cross validation) για την επαναληπτική διαδικασία. Τελικά, το καλύτερο μοντέλο προκύπτει όταν χρησιμοποιείται L2 κανονικοποίηση σε συνδυασμό με τον αλγόριθμο ”newton-cg”.

4. Δένδρο Απόφασης: Σχετικά με τα δένδρα απόφασης εξετάζεται η επίδραση δύο βασικών παραγόντων και πιο συγκεκριμένα του κριτηρίου που χρησιμοποιείται για τη δημιουργία κόμβων, καθώς επίσης και του μέγιστου βάθους του δένδρου. Τα δύο διαθέσιμα κριτήρια είναι ο συντελεστής ”gini” και η εντροπία (entropy), ενώ τα μέγιστα βάθη δένδρων που εξετάζονται είναι: 150, 200, 250 και 300. Τελικά, η μέγιστη απόδοση του ταξινομητή προκύπτει όταν ως κριτήριο χρησιμοποιείται ο συντελεστής ”gini” και το μέγιστο βάθος του δένδρου επιλέγεται το 250.

Να σημειωθεί ότι για την περίπτωση των μηχανών υποστηρικτικών διανυσμάτων δεν εφαρμόζεται η μέθοδος εύρεσης του καλύτερου συνδυασμού υπερπαραμέτρων λόγω του μεγάλου υπολογιστικού κόστους που απαιτείται.

Φάση Ελέγχου. Σε δεύτερο στάδιο παρουσιάζεται η απόδοση κάθε ταξινομητή, ο οποίος έχει εκπαιδευτεί με τις υπερπαραμέτρους που επιλέχθηκαν παραπάνω. Αρχικά, για κάθε ένα πραγματοποιείται εκτίμηση της ακρίβειας (accuracy), της σταθερής ακρίβειας (precision), της ανάκλησης (recall) και του F1 score. Τα μεγέθη αυτά εκτιμώνται με βάση την εξής συλλογιστική. Αν θεωρηθεί ως κατηγορία 1 η κουζίνα η οποία επιδιώκεται να προβλεφθεί για μία συγκεκριμένη συνταγή και σαν κατηγορία 0 όλες οι υπόλοιπες, τότε ορίζονται εκτιμώνται τα εξής μεγέθη:

i. True (ή False) Positive: Το πλήθος των συνταγών της κατηγορίας 1 (ή της 0) που κατηγοριοποιούνται σωστά (ή λανθασμένα) στην κατηγορία 1.

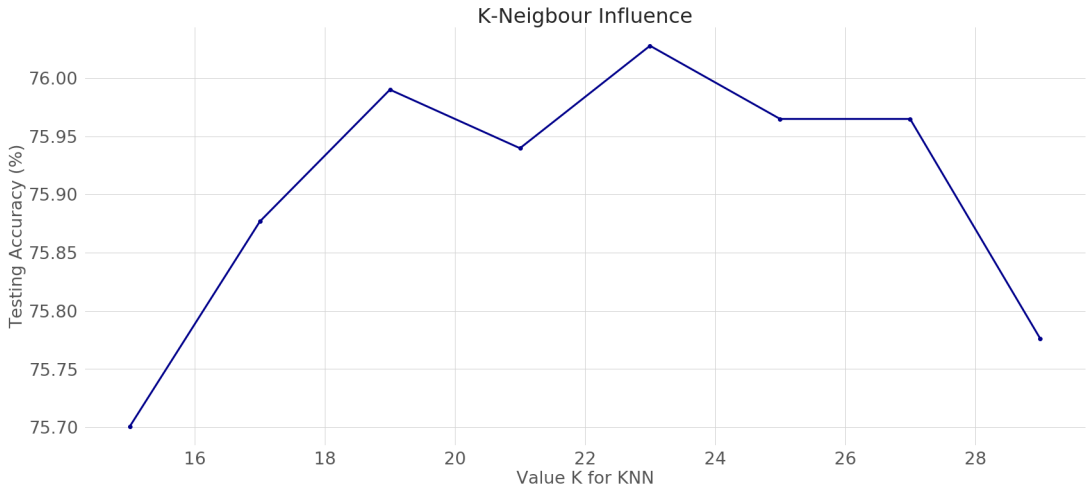
ii. True (ή False) Negative : Το πλήθος των συνταγών της κατηγορίας 0 (ή της 1) που κατηγοριοποιούνται σωστά (ή λανθασμένα) στην κατηγορία 0.

Χρησιμοποιώντας τους παραπάνω συντελεστές στη συνέχεια πραγματοποιείται ερμηνεία των μεγεθών που χρησιμοποιούνται για την αξιολόγηση των ταξινομητών.

- **Ακρίβεια (Accuracy)**

  Το συγκεκριμένο μέτρο εκφράζει το συνολικό αριθμό σωστών προβλέψεων προς το συνολικό πλήθος των προβλέψεων. Εν γένει όμως δεν αποτελεί αξιόπιστο μέτρο αξιολόγησης για μη ισορροπημένα σύνολα δεδομένων, δηλαδή για περιπτώσεις που το πλήθος των δειγμάτων μιας κατηγορίας είναι πολύ περισσότερα έναντι των άλλων. Το συγκεκριμένο μέτρο εκτιμάται ως εξής:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$



Σχήμα 3.2: Επίδραση πλήθους κοντινότερων γειτόνων

- **Σταθερή Ακρίβεια (Precision)**

  Η σταθερή ακρίβεια αντικατοπτρίζει το πλήθος των δειγμάτων που σωστά έχουν κατηγοριοποιηθεί στην κλάση 1 προς το συνολικό πλήθος δειγμάτων που έχει κατηγοριοποιήσει στην κατηγορία αυτή ο ταξινομητής. Συνεπώς υπολογίζεται ως εξής:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

- **Ανάκληση (Recall)**

  Η ανάκληση εκφράζει το πλήθος των δειγμάτων που σωστά έχουν κατηγοριοποιηθεί στην κλάση 1 προς το συνολικό πλήθος δειγμάτων της κατηγορίας 1 του συνόλου δεδομένων. Έτσι, μπορεί να υπολογιστεί από την ακόλουθη εξίσωση:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

- **F_1 score**

  Τέλος, το συγκεκριμένο μέτρο αξιολόγησης αποτελεί τον αρμονικό μέσο της σταθερής ακρίβειας και της ανάκλησης.

$$F_1\ score = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{3.4}$$

Η αξιολόγηση των ταξινομητών πραγματοποιείται χρησιμοποιώντας τα μεγέθη των Εξ. 3.1, 3.2, 3.3 και 3.4. Βέβαια, εφόσον ο αριθμός των διαφορετικών κουζίνων ανέρχεται στις 20, το εξεταζόμενο πρόβλημα θεωρείται ένα πρόβλημα πολυταξινόμησης (multiclass classification) και κατά συνέπεια εκτιμώνται είτε οι ”macro average” είτε οι ”micro average” τιμές των μεγεθών αυτών. Με τον πρώτο τρόπο η κάθε μετρική αξιολόγησης εκτιμάται ανεξάρτητα για κάθε κλάση και εν συνεχεία υπολογίζεται ο μέσος όρος, δηλαδή κάθε κλάση αντιμετωπίζεται ισοδύναμα. Αντίθετα, στην περίπτωση του ”micro average” συνεκτιμάται η συνεισφορά κάθε κλάσης για τον υπολογισμού της μέσης μετρικής.

Συγκεντρωτικά λοιπόν, στον Πίνακα 3.1 παρουσιάζονται οι τιμές των παραπάνω μετρικών αξιολόγησης για κάθε ταξινομητή που εκπαιδεύεται. Τα αποτελέσματα αυτά προκύπτουν συγκρίνοντας τις προβλέψεις των ταξινομητών για τα δεδομένα του συνόλου ελέγχου (test set) με τις πραγματικές τιμές αυτών.

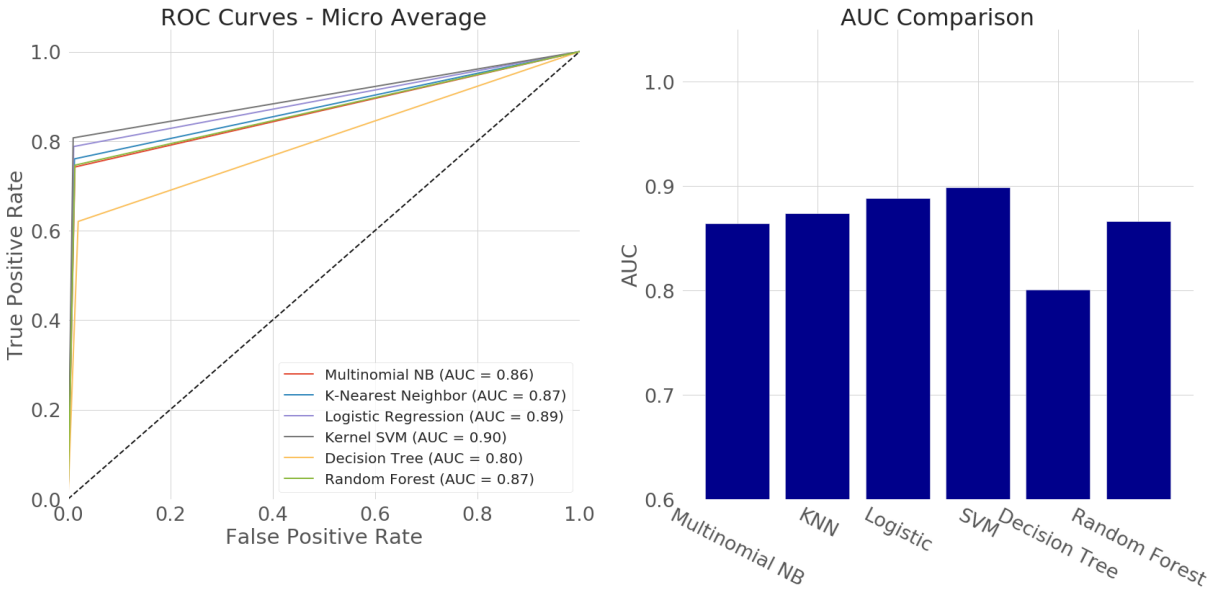
Η πιο διαδεδομένη και ασφαλής μέθοδος αξιολόγησης μιας μεθόδου είναι οι καμπύλες ROC (Receiver Operating Characteristics), όπως επίσης και το εμβαδόν κάτω από τις καμπύλες αυτές (Area Under Curve-AUC). Αποτελούν ουσιαστικά τους 2 πιο σημαντικές μετρικές για τον έλεγχο της απόδοσης

Πίνακας 3.1: Αξιολόγηση ταξινομητών με διάφορες μετρικές.

	Multinomial Naive Bayes	**k-Nearest Neighbour**	**Logistic Regression**	**Kernel SVM**	**Decision Tree**	**Random Forest**
Macro Average Metrics [%]						
Accuracy	74.22	76.03	78.79	81.00	62.05	74.62
Precision	73.73	76.45	76.69	81.25	51.54	79.09
Recall	60.41	62.85	66.90	69.87	50.49	58.03
F_1-Score	65.09	67.63	70.68	74.28	50.89	64.18
Micro Average Metrics [%]						
All Metrics	79.79	76.03	78.79	80.73	62.05	74.62

ταξινομητών, ανεξάρτητα από τη φύση των δεδομένων που εξετάζονται. Ο οριζόντιος άξονας αναπαριστά το False Positive Rate, δηλαδή το πλήθος των δειγμάτων που κατηγοριοποιήθηκαν λανθασμένα στην κατηγορία 1 προς το συνολικό πλήθος των δειγμάτων που πραγματικά ανήκουν στην κατηγορία 0. Ο κατακόρυφος άξονας αναπαριστά την ανάκληση (ή ισοδύναμα το True Positive Rate). Αν ένας ταξινομητής κατηγοριοποιεί όλα τα δεδομένα σωστά, τότε οι τιμές της ανάκλησης θα ισούνται με 1 για όλες τις τιμές του False Positive Rate και κατά συνέπεια το εμβαδόν κάτω από την καμπύλη (AUC) θα ισούται επίσης με 1. Αντιθέτως, όσο πιο λανθασμένες προβλέψεις εξάγει η μέθοδος για τις 2 κατηγορίες τόσο μικρότερο είναι και το εμβαδόν του κάτω από την καμπύλη ROC.

Στο Σχήμα 3.3 απεικονίζονται αριστερά οι καμπύλες ROC και δεξιά οι τιμές των εμβαδών κάτω από αυτές για όλους τους ταξινομητές που εξετάζονται. Σύμφωνα με τα αποτελέσματα προκύπτει ότι ο λιγότερο αποδοτικός αλγόριθμος για το συγκεκριμένο πρόβλημα είναι τα δένδρα απόφασης, με τιμή AUC=0.8. Συγκεκριμένα ο ταξινομητής αυτός κατέχει τη χαμηλότερη απόδοση σε όλο το εύρος του False Positive Rate, με τη μεγαλύτερη απόκλιση - έναντι των υπολοίπων - να εμφανίζεται για χαμηλές τιμές του False Positive Rate. Όσον αφορά τους υπόλοιπους 5 ταξινομητές προκύπτει ότι οι αποδόσεις τους δε διαφέρουν σημαντικά, με την καλύτερη δυνατή να επιτυγχάνεται από τις μηχανές υποστηρικτικών διανυσμάτων (RBF-Kernel SVM) με τιμή AUC=0.90. Ακολουθεί η μέθοδος της λογιστικής παλινδρόμησης, με την τιμή AUC να ανέρχεται στο 0.89.



Σχήμα 3.3: Σύγκριση αποδόσεων ταξινομητών.

Σημείωση: Από τις μεθόδους συνόλων (Ensemble methods) παρουσιάσθηκε μόνο η μέθοδος Random Forest που στηρίζεται στη φιλοσοφία του λεγόμενου bagging. Ωστόσο, για το συγκεκριμένο πρόβλημα εξετάσθηκαν επίσης και οι μέθοδοι Voting και Boosting. Τα αποτελέσματα όμως έδειξαν ότι πέραν του μεγάλου υπολογιστικού τρόπου που απαιτείται, δεν οδηγούν σε υψηλότερες αποδόσεις συγκριτικά με αυτές των μεμονομένων ταξινομητών. Για τους λόγους αυτούς λοιπόν οι μέθοδοι αυτοί δεν αναλύονται στην παρούσα εργασία, αλλά οι υλοποιήσεις τους συμπεριλαμβάνονται στο αρχείο κώδικα που έχει κατατεθεί.

3.3 Μέθοδος Κύριων Συνιστωσών

Όπως προαναφέρθηκε στην Ενότητα 3.1, ο αριθμός των χαρακτηριστικών που προκύπτουν ύστερα από την κωδικοποίηση ανέρχεται στα 2.845. Παρόλο λοιπόν που η επεξεργασία τους βασίστηκε στην κλειστή μορφή τους ύστερα από χρήση της συνάρτησης TfidfVectorizer και η ταχύτητα των αλγορίθμων θεωρείται ικανοποιητική, στην υποενότητα αυτή εξετάζεται πώς μπορεί να επηρεάσει την ακρίβεια των παραπάνω υπολογισμών η μείωση της διαστατικότητας των δεδομένων. Η μείωση αυτή μπορεί να επιτευχθεί με διάφορες τεχνικές, ενώ αυτή που επιστρατεύεται στην παρούσα περίπτωση είναι η μέθοδος των Κυρίων Συνιστωσών (Principal Component Analysis) [2].

Φάση Εκπαίδευσης: Χρησιμοποιώντας τη μέθοδο αυτή καθορίζεται ένα n-διάστατο (όπου n=2.845 λόγω του αρχικού πλήθους των χαρακτηριστικών) ορθοκανονικό σύστημα συντεταγμένων. Στο σύστημα αυτό προβάλλονται τα αρχικά χαρακτηριστικά, δημιουργώντας έτσι καινούρια διανύσματα χαρακτηριστικών, τα οποία θεωρούνται ανεξάρτητα μεταξύ τους. Για την εφαρμογή της μεθόδου απαιτείται η αποσύνθεση σε κύριες συνιστώσες (PCA decomposition) του δειγματικού πίνακα συνδιασπορών που εκτιμάται από τα διανύσματα των χαρακτηριστικών των δεδομένων εκπαίδευσης. Η αποσύνθεση αυτή οδηγεί στην εξαγωγή των ιδιοτιμών (κατά φθίνουσα σειρά) και των αντίστοιχων ιδιοδιανυσμάτων του πίνακα συνδιασπορών, οι διευθύνσεις των οποίων καθορίζουν το ορθοκανονικό σύστημα συντεταγμένων (ή αλλιώς καθορίζουν τις κύριες συνιστώσες). Οι εξισώσεις που διέπουν την προαναφερθείσα συλλογιστική παρουσιάζονται συγκεντρωτικά παρακάτω.

Για την ανάλυση που ακολουθεί, από κάθε διάνυσμα χαρακτηριστικών που χρησιμοποιείται έχει αφαιρεθεί η μέση τιμή του και συμβολίζεται ως $\boldsymbol{\alpha}_{o,k}$.

Η ανάλυση των κύριων συνιστωσών πραγματοποιείται στον $n \times n$ (όπου n=2.845 λόγω του αρχικού πλήθους των χαρακτηριστικών) δειγματικό πίνακα συνδιασπορών $\boldsymbol{P}$ των χαρακτηριστικών των δεδομένων εκπαίδευσης:

$$\boldsymbol{P} := \frac{1}{p-1} \sum_{k=1}^{p} \boldsymbol{\alpha}_{o,k}\, \boldsymbol{\alpha}_{o,k}^{T} \tag{3.5}$$

Σύμφωνα με την PCA, ο παραπάνω πίνακας αποσυνθέτεται σε γινόμενο 3 πινάκων:

$$\boldsymbol{P} = \boldsymbol{U}\ \boldsymbol{S}^2\ \boldsymbol{U}^T \tag{3.6}$$

όπου:

$$\boldsymbol{S}^2 := diag(s_1^2,\ s_2^2,\ \ldots,\ s_n^2) \quad (n \times n), \qquad \boldsymbol{U} := [\boldsymbol{u}_1\ \boldsymbol{u}_2\ \ldots\ \boldsymbol{u}_n] \quad (n \times n) \tag{3.7}$$

με $diag(s_1^2, s_2^2, \ldots, s_n^2)$ να αποτελεί ένα διαγώνιο πίνακα που περιέχει τις θετικές ιδιοτιμές s_i^2 σε φθίνουσα σειρά του δειγματικού πίνακα συνδιασπορών $\boldsymbol{P}$, και $\boldsymbol{U}$ ο πίνακας που περιέχει τα αντίστοιχα ιδιοδιανύσματα $\boldsymbol{u}_i$ του $\boldsymbol{P}$. Τα διανύσματα είναι ορθοκανονικά, έτσι ώστε $\boldsymbol{u}_i^T\, \boldsymbol{u}_j = \delta_{ij}$, με δ_{ij} να υποδηλώνει το δέλτα του Kronecker .

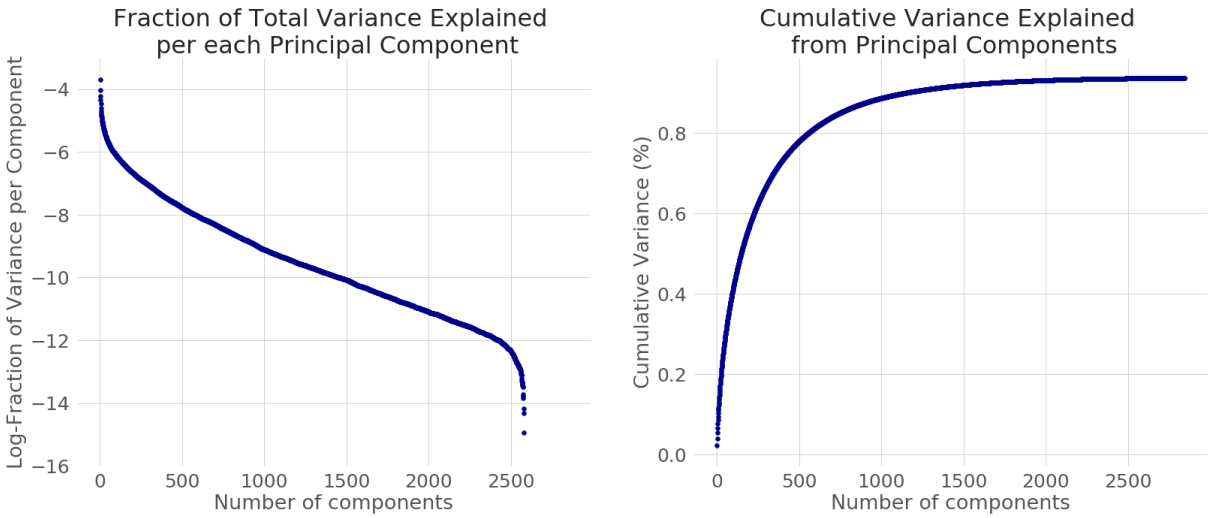
Επόμενο ερώτημα θεωρείται η επιλογή του m-διάστατου συστήματος συντεταγμένων (με $m < n$) στο οποίο πρέπει θα προβληθούν τα αρχικά χαρακτηστικά των δεδομένων εκπαίδευσης. Εναλλακτικά, το ερώτημα που τίθεται είναι πόση θα είναι η τιμή του $\boldsymbol{q}$ η οποία αναφέρεται στις πρώτες $\boldsymbol{q}$ στήλες του πίνακα $\boldsymbol{U}$ που εξηγούν το μεγαλύτερο ποσοστό της μεταβλητότητας των χαρακτηριστικών και οι οποίες πρέπει να διατηρηθούν. Η τιμή του $\boldsymbol{q}$ λοιπόν επιλέγεται έτσι ώστε να αποτελεί τον ελάχιστο αριθμό των ιδιοτιμών που εξηγούν ένα επιδιωκόμενο ποσοστό γ της μεταβλητότητας των χαρακτηριστικών των δεδομένων εκπαίδευσης, δηλαδή:

$$q := \min_{l} \left\{ l \in [1, \ldots, n] \, | \gamma \leq \frac{\sum_{j=1}^{l} s_j^2}{\sum_{j=1}^{n} s_j^2} \times 100\,(\%) \right\} \tag{3.8}$$

όπου ο αριθμητής του κλάσματος της Εξ. 3.8 ισούται με το άθροισμα της διασποράς των l συνιστωσών των χαρακτηριστικών στο περιστραμμένο σύστημα συντεταγμένων, ενώ ο παρονομαστής με το συνολικό άθροισμα της διασποράς όλων των χαρακτηριστικών. Τέλος, η εκπαίδευση των ταξινομητούν πραγματοποιείται στα χαρακτηριστικά του περιστραμμένου m-διάστατου συστήματος συντεταγμένων των δεδομένων εκπαίδευσης.

Φάση Ελέγχου: Εφόσον έχει επιλεχθεί το m-διάστατο σύστημα συντεταγμένων πάνω στο οποίο έχουν προβληθεί τα διανύσματα των χαρακτηριστικών των δεδομένων εκπαίδευσης, αν αποκτηθεί κάποια συνταγή χωρίς να είναι γνωστή η κουζίνα από την οποία προέρχεται, τότε το διάνυσμα των χαρακτηριστικών της συνταγής αυτής προβάλλεται στο περιστραμμένο m-διάστατο σύστημα συντεταγμένων. Στη συνέχεια, προσαρμόζονται οι εκπαιδευμένοι ταξινομητές στα περιστραμμένα χαρακτηριστικά αυτά, η διάσταση των οποίων θα είναι μικρότερη από την αρχική διάσταση των χαρακτηριστικών.

Σύμφωνα λοιπόν με τα προαναφερθέντα, εκτιμάται ο δειγματικός πίνακας συνδιασπορών των χαρακτηριστικών των δεδομένων εκπαίδευσης και στη συνέχεια εξάγονται τα ιδιοδιανύσματα του πίνακα αυτού. Η συνεισφορά κάθε ιδιοδιανύσματος στη συνολική διασπορά των δεδομένων απεικονίζεται στο Σχήμα 3.4, όπου παρατηρείται ότι για την ερμηνεία του 90% της διασποράς των δεδομένων εκπαίδευσης απαιτείται η διατήρηση περίπου των 1.000 πρώτων ιδιοδιανυσμάτων από τα συνολικά 2.845. Επιπλέον, η διατήρηση 2.000 ιδιοδιανυσμάτων αντιστοιχίζεται στην ερμηνεία του 99% της διασποράς των χαρακτηριστικών εκπαίδευσης, ενώ τέλος η συνεισφορά των τελευταίων 300 ιδιοδιανυσμάτων είναι μηδαμινή.

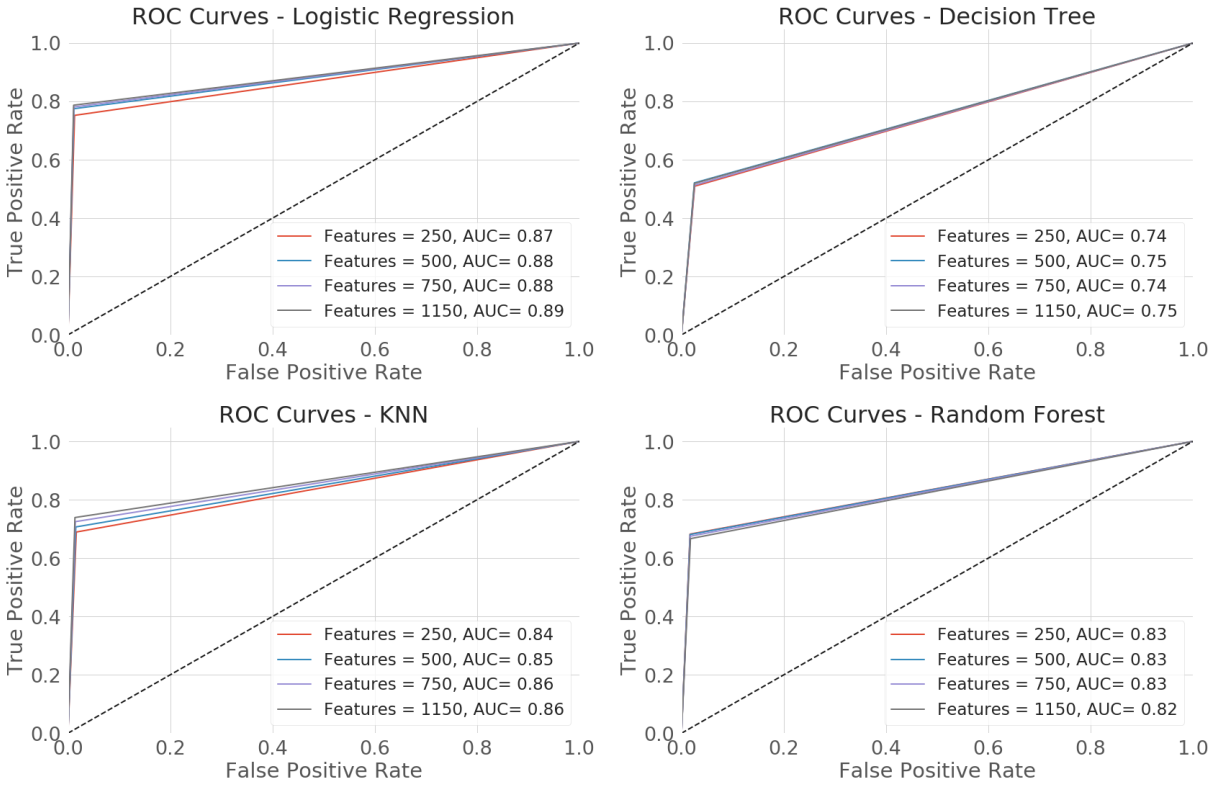


Σχήμα 3.4: Ημιλογαριθμικό γράφημα της συνεισφοράς κάθε ιδιοδιανύσματος στη συνολική μεταβλητότητα των 2.845 ιδιοδιανυσμάτων των χαρακτηριστικών της φάσης εκπαίδευσης

Εφόσον το ποσοστό της διασποράς που ερμηνεύεται από ένα συγκεκριμένο αριθμό ιδιοδιανυσμάτων μπορεί να επηρεάσει τις προβλέψεις των ταξινομητών, στη συνέχεια εξετάζεται η επίδραση αυτή για τους ταξινομητές k-NN, Logistic Regression, Decision Tree και Random Forest. Συγκεκριμένα, με χρήση καμπύλων ROC και των συντελεστών AUC ελέγχεται πώς επιδρά η διατήρηση των πρώτων 250, 500, 750 και 1.150 ιδιοδιανυσμάτων των δεδομένων εκπαίδευσης στην πρόβλεψη των αντίστοιχων δεδομένων ελέγχου. Να σημειωθεί ότι τα ποσοστά της διασποράς των χαρακτηριστικών που ερμηνεύουν τα προαναφερθέντα ιδιοδιανύσματα ισούνται με 64.8%, 81.2%, 89.2% και 95.3% αντίστοιχα.

Σύμφωνα λοιπόν με τα αποτελέσματα του Σχήματος 3.5, παρατηρείται ότι η επίδραση του πλήθους των διατηρούμενων ιδιοδιανυσμάτων δεν επηρεάζει αισθητά την απόδοση των ταξινομητών. Συγκεκριμένα, παρότι στις μεθόδους Logistic Regression και k-NN παρατηρείται μια μικρή αύξηση

του συντελεστή AUC (από 0.87 σε 0.89 και από 0.84 σε 0.86 αντίστοιχα) με αύξηση του αριθμού των ιδιοδιανυσμάτων, η διαφορά αυτή δεν μπορεί να θεωρηθεί σημαντική. Επιπροσθέτως, το φαινόμενο αυτό δεν παρατηρείται στις μεθόδους Decision Tree και Random Forest, καθώς οι αποδόσεις των ταξινομητών δε φαίνεται να επηρεάζονται. Κατά συνέπεια, η διατήρηση λίγων χαρακτηριστικών μπορεί να επιφέρει σχεδόν ίδια ποσοστά ταξινόμησης, εξασφαλίζοντας ταυτόχρονα μικρότερο υπολογιστικό κόστος.



Σχήμα 3.5: Καμπύλες ROC για την επίδραση του αριθμού των ιδιοδιανυσμάτων στις αποδόσεις των επιμέρους ταξινομητών

Σύνοψη

4.1 Συμπεράσματα

Στην παρούσα εργασία πραγματοποιήθηκε διερευνητική ανάλυση του συνόλου δεδομένων "What's Cooking?" με στόχο την εκπαίδευση αλγορίθμων μηχανικής μάθησης για την ταξινόμηση μίας συνταγής στην κουζίνα προέλευσής της. Ορισμένα συμπεράσματα που μπορούν να εξαχθούν από την ανάλυση που προηγήθηκε είναι τα εξής:

1. Όσον αφορά την προεπεξεργασία των δεδομένων, χρησιμοποιήθηκε ένα πλήθος τεχνικών που περιγράφηκε αναλυτικά στην Ενότητα 2.1, έτσι ώστε τα δεδομένα να "καθαριστούν" με τον καλύτερο δυνατό τρόπο.

2. Η βέλτιστη απόδοση προέκυψε από τις Μηχανές Υποστηρικτικών Διανυσμάτων όταν χρησιμοποιούνται σε συνδυασμό με πυρήνα RBF, προσεγγίζοντας την τιμή AUC $= 0.90$.

3. Οι αποδόσεις της Λογιστικής Παλινδρόμησης, των 23 - Κοντινότερων Γειτόνων και του Πολωνυμικού Αφελή Ταξινομητή είναι παρόμοιες και πολύ κοντά σε αυτή των Μηχανών Υποστηρικτικών Διανυσμάτων.

4. Η χειρότερη επίδοση προέκυψε για τα Δένδρα Απόφασης, αγγίζοντας τιμή AUC $= 0.80$. Βελτιωμένη απόδοση - συγκριτικά με τα Δένδρα Απόφασης - έχουν τα Τυχαία Δάση, τα οποία οδηγούν σε τιμή AUC $= 0.87$.

4. Η μείωση της διαστατικότητας των δεδομένων δε φαίνεται να επηρεάζει σημαντικά τις προβλέψεις των μεθόδων ακόμα και στην περίπτωση που τα διατηρούμενα ιδιοδιανύσματα ερμηνεύουν το 60% της διασποράς. Κατά συνέπεια, η διατήρηση λίγων χαρακτηριστικών μπορεί να επιφέρει σχεδόν ίδια ποσοστά ταξινόμησης με μικρότερο υπολογιστικό κόστος.

4.2 Μελλοντικές Επεκτάσεις

Μια από τις πιθανές επεκτάσεις της συγκεκριμένης εργασίας αποτελεί η υλοποίηση νευρωνικών δικτύων και σύγκριση των απόδόσεων με τις τεχνικές που αναλύθηκαν. Συγκεκριμένα, μπορεί να εξετασθεί η απόδοση ενός πολυεπίπεδου νευρωνικού δικτύου (Multilayer Rercepton) ή ενός δικτύου τύπου LSTM [3]. Αρκετή βιβλιογραφία για ταξινόμηση με λέξεις υπάρχει επίσης βασισμένη στη χρήση συνελικτικών δικτύων, τα οποία δύναται να εφαρμοστούν στο συγκεκριμένο πρόβλημα [4]. Οι λέξεις κωδικοποιούνται ως διανύσματα (word embeddings) και ο τελικός πίνακας που προκύπτει εισάγεται στο αντίστοιχο embedding επίπεδο. Τέλος, θα μπορούσε να εξετασθεί μία εναλλακτική μέθοδος μείωσης της διαστατικότητας των δεδομένων έναντι της Ανάλυσης σε Κύριες Συνιστώσες (PCA). Συγκεκριμένα, εφόσον τα χαρακτηστικά τις παρούσης εργασίας είναι κατηγορικές μεταβλητές, εφικτή είναι η υλοποίηση της μεθόδου Ανάλυσης Αντιστοιχιών (Correspondense Analysis) σε συνδυασμό με νευρωνικά δίκτυα [5].

Κατάλογος πινάκων

Κατάλογος σχημάτων

Βιβλιογραφία

[1] T. Joachims. Transductive Inference for Text Classification using Support Vector Machines, *Proceedings of the 16th International Conference on Machine Learning (ICML)*, 200–209, 1999.

[2] K.J. Vamvoudakis–Stefanou, J.S. Sakellariou, S.D. Fassois, Vibration–based damage detection for a population of nominally identical structures: Unsupervised Multiple Model (MM) statistical time series type methods, *Mechanical Systems and Signal Processing*, **111**, 149–171, 2018. https://doi.org/10.1016/j.ymssp.2018.03.054.

[3] J. Nowak, A. Taspinar, R. Scherer. LSTM recurrent neural networks for short text and sentiment classification, *Rutkowski L., Korytkowski M., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J. (eds), Artificial Intelligence and Soft Computing (ICAISC)*, 553–562, 2017. Also in *Lecture Notes in Computer Science (LNCS)*, **10246**. https://doi.org/10.1007/978-3-319-59060-8_50.

[4] N. Kalchbrenner, E.Grefenstette, P. Blunsom. A convolutional neural network for modelling sentences, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, **1**, 655–665, Baltimore, Maryland, USA, 2014. doi:10.3115/v1/P14-1062.

[5] H. Hsu, S. Salamatian, F. P. Calmon. Correspondence analysis using neural networks, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, **89**, Naha, Okinawa, Japan, 2019.