ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Εκτίμηση της δημοτικότητας των ειδήσεων με βάση τα χαρακτηριστικά τους

Νεφέλη-Γεωργία Αργυροπούλου

Αθήνα, Μάιος 2018

Τριμελής Εξεταστική Επιτροπή

Επιβλέπων: Βαρλάμης Ηρακλής, Επίκουρος Καθηγητής

2ο Μέλος Επιτροπής: Μιχαήλ Δημήτριος, Επίκουρος Καθηγητής

3ο Μέλος Επιτροπής: Τσερπές Κωνσταντίνος, Επίκουρος Καθηγητής

Η Νεφέλη-Γεωργία Αργυροπούλου,

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τόσο την οικογένεια και τους φίλους μου, όσο και το καθηγητή μου κ. Βαρλάμη, για την απεριόριστη υπομονή που έδειξαν για την ολοκλήρωση αυτής της πτυχιακής.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη στα Ελληνικά

Δεδομένου ότι στις μέρες μας, η συνεχής ανάπτυξη του Παγκόσμιου Ιστού είναι ραγδαία έχει ως αποτέλεσμα μια πληθώρα άρθρων να δημοσιεύεται καθημερινά. Θα ήταν άκρως ιδανικό, λοιπόν, αν μπορούσε κάποιος να προβλέψει τη δημοτικότητα ενός άρθρου, πριν ακόμη αυτό δημοσιευθεί. Αυτό το ερώτημα προσπαθήσαμε να απαντήσουμε και σε αυτή τη μελέτη: κατά πόσο υπάρχουν χαρακτηριστικά, τα οποία θα μπορούσαν να κρίνουν τη δημοτικότητά του άρθρου.

Για να επιτευχθεί αυτό, χρησιμοποιήθηκε ένα dataset 40.000 περίπου άρθρων τα οποία προερχόντουσαν από το Mashable.com και αφορούσαν τη περίοδο 2 χρόνων. Αυτά τα άρθρα, πέρα από το μεγάλο όγκο τους, είχαν και πάρα πολλά χαρακτηριστικά, συγκεκριμένα 61. Χρειάστηκε, λοιπόν, να πραγματοποιηθεί μια επεξεργασία ώστε να μειωθούν οι εγγραφές. Έπειτα, δημιουργήθηκαν 3 μοντέλα: ένα στο οποίο εφαρμοζόντουσαν οι αλγόριθμοι σε όλο το σύνολο και δυο ακόμη τα οποία χρησιμοποιούσαν διαφορετικό τρόπο μείωσης των γνωρισμάτων, είτε με PCA είτε με βάση το Information Gain.

Οι αλγόριθμοι που χρησιμοποιήθηκαν ήταν οι: kNN, Neural Nets και Naïve Bayes. Αυτό που παρατηρήσαμε είναι ότι ο kNN, λειτουργεί καλύτερα όταν επιλεγόταν η μέθοδος του PCA, σε αντίθεση με τον Naïve Bayes και τα Neural Nets που είχαν καλύτερα αποτελέσματα στο μοντέλο στο οποίο χρησιμοποιήθηκε το Information Gain.

Λέξεις κλειδιά: Δημοτικότητα ειδήσεων, Δημοφιλές άρθρο, Rapid Miner, Εξόρυξη Γνώσης

Abstract

Since Web development is rapid these days, there are many articles that are published every day. It would be extremely helpful, if someone could predict the popularity of an article before it is published. That is the question we tried to answer in this study: whether there are features that could predict its article's popularity.

To achieve this, a dataset of about 40.000 articles from Mashable.com was used, covering a 2 year period. These articles, in addition to their large amount, had also too many features, 61 specifically. Therefore, a preprocessing had to be made in order to reduce the entries. Then, 3 models were created: one in which the algorithms were applied across the whole dataset, and two more that used different methods of decreasing the attributes, either with PCA or based on their Information Gain.

The algorithms that applied were: kNN, Neural Nets and Naïve Bayes. What has been noticed is that kNN, had better results when PCA was applied. In contrast, Naïve Bayes and Neural Nets got higher percentages at the model that used the method of Information Gain.

Keywords: News Popularity, Online News, Data Mining

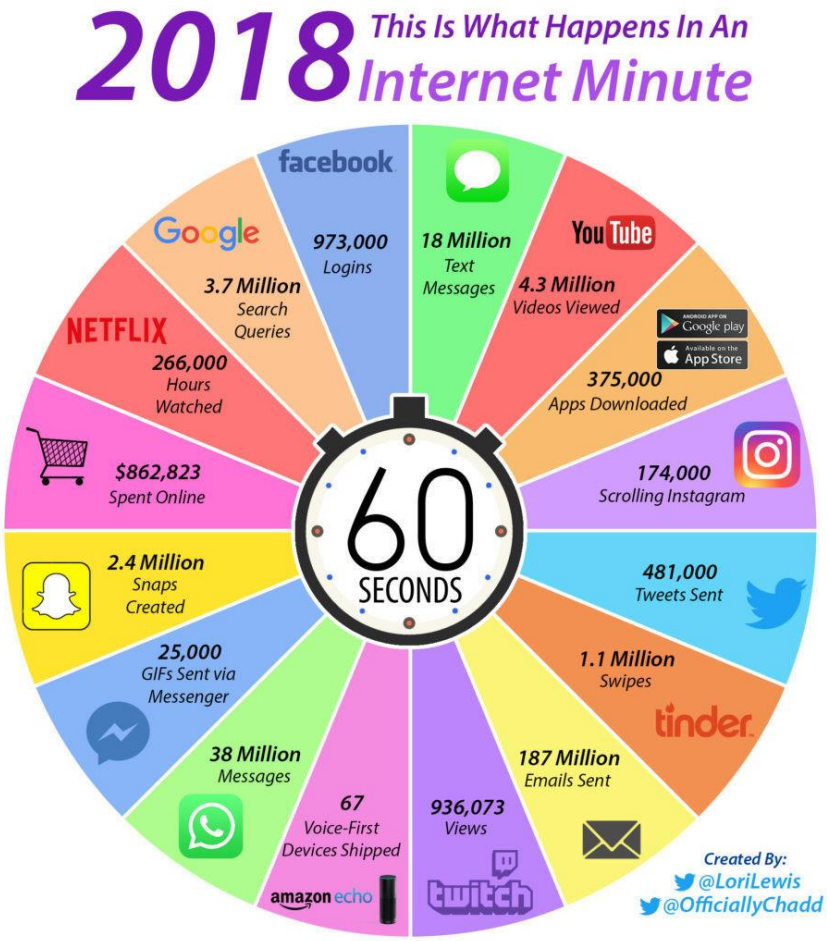
ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

Το γεγονός ότι ο Παγκόσμιος Ιστός αριθμεί πλέον πάνω από 4 δισεκατομμύρια χρήστες (31 Δεκεμβρίου 2017), μας κάνει να αναρωτιόμαστε πόσα άρθρα μπορεί να δημοσιεύονται καθημερινά στο Internet. Όταν, λοιπόν, ένας χρήστης γράφει ένα άρθρο, το πιο πιθανό αποτέλεσμα, δεδομένου αυτού του τεράστιου αριθμού χρηστών, είναι να χαθεί μέσα σε αυτό τον ωκεανό πληροφοριών. Ένα, ακόμη, αποτέλεσμα αυτού του «ωκεανού» είναι ότι τα άρθρα πλέον έχουν μικρή «διάρκεια ζωής» και λόγω αυτού, για να πούμε ότι ένα άρθρο είναι δημοφιλές, πρέπει να καταφέρει να διαδοθεί σε πολλά άτομα, μέσα σε μικρό χρονικό διάστημα.

Απόδειξη για τον «ωκεανό» του διαδικτύου, αποτελεί η επόμενη εικόνα η οποία μας δείχνει τι συμβαίνει μέσα σε 60” στο Internet.



Ως εκ τούτου, εάν υπήρχε ένας τρόπος να ξέρουμε από πριν τι χρειάζεται να προσέξουμε κατά τη συγγραφή του κειμένου ώστε να κάνουμε το άρθρο μας πιο ελκυστικό και ως αποτέλεσμα να μπορέσει να διαδοθεί πιο γρήγορα και, σε όσο το δυνατόν γίνεται, περισσότερα άτομα, θα μπορούσαμε να διαμορφώσουμε κατάλληλα τις σχετικές ενέργειές μας.

Γι' αυτό το λόγο, μια μεγάλη ερευνητική πρόκληση είναι το κατά πόσο μπορεί να προβλέψει κάποιος τη δημοτικότητα ενός άρθρου πριν αυτό ακόμη δημοσιευθεί ώστε να μπορέσει να πάρει τις κατάλληλες αποφάσεις. Δυστυχώς όμως, με τις υπάρχουσες τεχνικές, κάτι τέτοιο δεν είναι τόσο απλό και εύκολο. Επιπλέον, κάποιες έρευνες (Arapakis, 2014) μπόρεσαν να προβλέψουν, με αρκετά μεγάλη ακρίβεια, μονάχα τα μη δημοφιλή άρθρα κάτι το οποίο δεν είναι ένα ιδιαίτερα επιθυμητό αποτέλεσμα. Από την άλλη πλευρά όμως, μερικοί ερευνητές (Bandari, 2012) ανακάλυψαν πως η πηγή δημοσίευσης είναι ένα αρκετά σημαντικό χαρακτηριστικό που μπορεί να επηρεάσει την τελική δημοτικότητα του άρθρου στο Twitter. Τέλος, το 2015 δημιουργήθηκε (Fernandes, 2015) ένα εργαλείο το οποίο είχε τη δυνατότητα τόσο να προβλέπει τη δημοτικότητα του άρθρου, όσο και να προτείνει αλλαγές στο κείμενο και στη δομή, με στόχο τη βελτίωση της διάδοσής του.

1.1 Στόχος της εργασίας

Σκοπός αυτής της εργασίας είναι να δούμε αν τελικά υπάρχουν στοιχεία που μπορούν να κάνουν ένα άρθρο πιο επιτυχημένο και δημοφιλές. Τέτοια στοιχεία θα μπορούσαν να είναι π.χ. η μέρα δημοσίευσής του, η ύπαρξη βίντεο και φωτογραφιών, η αναφορά σημαντικών προσώπων, η αναφορά άλλων άρθρων στο κείμενο κ.λπ. Για τους σκοπούς της εργασίας, χρησιμοποιήσαμε ένα σύνολο 39.797 δεδομένων, τα οποία προήλθαν από την ιστοσελίδα Mashable.com. Μέσω, λοιπόν, του RapidMiner Studio και τεχνικών κατηγοριοποίησης,

προσπαθήσαμε να εξετάσουμε, αν κάποιο από αυτά τα χαρακτηριστικά που είχε το κάθε άρθρο, ήταν ικανά να «προδώσουν» τη δημοτικότητα τους. Τα αρχικά δεδομένα περιείχαν 61 χαρακτηριστικά. Χρειάστηκε, λοιπόν, μια προεπεξεργασία ώστε να μειωθούν και να μπορούν να είναι πιο εύκολα διαχειρίσημα από τους αλγορίθμους. Έτσι, δημιουργήσαμε 3 μοντέλα στα οποία εφαρμόσαμε αλγορίθμους κατηγοριοποίησης (K-NN, Neural Nets και Naïve Bayes). Τα μοντέλα, διέφεραν ως προς τη μέθοδο μείωσης των γνωρισμάτων:

1. Μοντέλο 1=εφαρμογή αλγορίθμων σε όλο το dataset
2. Μοντέλο 2=χρήση top 10 γνωρισμάτων με βάση το Information Gain τους
3. Μοντέλο 3=εφαρμογή μεθόδου PCA με 10 principal components

Στη συνέχεια, φάνηκε ότι ο αλγόριθμος kNN είχε καλύτερα αποτελέσματα όταν εφαρμοζόταν η μέθοδος του PCA για τη μείωση των γνωρισμάτων. Αντίθετα, ο Naïve Bayes και τα Neural Nets παρουσίαζαν καλύτερες μετρικές αξιολόγησης με την εφαρμογή του Information Gain.

ΚΕΦΑΛΑΙΟ 2: Θεωρητικό Υπόβαθρο

Η έρευνα πάνω στο αντικείμενο της πρόβλεψης δημοτικότητας ενός άρθρου, έχει ξεκινήσει από το 2010 και συνεχίστηκε το 2012 και το 2014. Σε αυτή, λοιπόν, την ενότητα θα αναλύσουμε 3 από τις πιο σημαντικές έρευνες που έχουν πραγματοποιηθεί πάνω σε αυτό το θέμα.

2.1 Πρόβλεψη δημοτικότητας ειδήσεων στα Social Media

Οι Roja Bandari et al. (2012), έκαναν μια από τις πιο σημαντικές έρευνες πάνω στο θέμα της πρόβλεψης δημοτικότητας των ειδήσεων. Τα δεδομένα που χρησιμοποίησαν ήταν δημοσιευμένα άρθρα στο Twitter και τα αξιολόγησαν με βάση τα retweets τους.

Ο βασικός τους στόχος, ήταν να διαπιστώσουν ποια από τα ακόλουθα 4 χαρακτηριστικά παίζουν μεγαλύτερο ρόλο στη διάδοση ενός άρθρου:

1. η κατηγορία του θέματος του
2. η γλώσσα στην οποία είναι γραμμένο το άρθρο (απλή ή επίσημη)
3. η ύπαρξη αναφορών σε σημαντικά πρόσωπα
4. η πηγή που δημοσίευσε το άρθρο

Για να μπορέσουν να εξετάσουν αν μπορεί να προβλεφθεί η διάδοση ενός άρθρου μέσα από αυτά τα χαρακτηριστικά, υλοποίησαν διάφορους αλγορίθμους για classification και regression (Linear Regression, Knn, SVM και Decision Trees).

Τα συμπεράσματα στα οποία κατέληξαν είναι τα εξής:

- Άλλες πηγές είναι οι «παραδοσιακά» δημοφιλής και άλλες εντοπίστηκαν μέσα στα δεδομένα.

- Πολύ σημαντικό ρόλο παίζει η πηγή που δημοσίευσε το άρθρο.
- Η κατηγορία του θέματος δε βοηθάει τόσο στο να προβλεφθεί η συνολική δημοτικότητα του άρθρου, αλλά στο αν το άρθρο θα αναδημοσιευθεί.

Μετά, λοιπόν, από αυτά τα παραπλέυρως συμπεράσματα και με τη χρήση των παραπάνω αλγορίθμων, οι ερευνητές έδειξαν ότι δεν είναι εφικτό να βρεθεί ο ακριβής αριθμός retweets ενός άρθρου, αλλά μπορεί να προσδιορισθεί με αρκετά μεγάλη επιτυχία, ένα εύρος δημοτικότητας αυτού.

2.2 Πρόβλεψη δημοτικότητας ειδήσεων πριν τη δημοσίευσή τους

Οι Ioannis Arapakis et al., προσπάθησαν κι αυτοί με τη σειρά τους, το 2014 να ελέγξουν αν είναι εφικτό να προβλεφθεί η δημοτικότητα των ειδήσεων μετά την δημοσίευσή τους. Βασίστηκαν σε ένα μεγάλο βαθμό στην έρευνα που αναφέραμε παραπάνω και προσπάθησαν να προσθέσουν και δικά τους στοιχεία.

Χρησιμοποίησαν περίπου 13.000 άρθρα που δημοσιεύθηκαν στο Yahoo News και τα αξιολόγησαν, πέρα από τον αριθμό των tweets, με το πλήθος των views. Έπειτα, υλοποίησαν τους ίδιους αλγορίθμους με την προηγούμενη έρευνα αλλά κατέληξαν σε διαφορετικά αποτελέσματα καθώς έλεγξαν τη δημοτικότητα του άρθρου μετά από 1 εβδομάδα δημοσίευσης και όχι μετά από 4 ημέρες όπως οι προηγούμενοι ερευνητές.

Τα συμπεράσματα στα οποία κατέληξαν είναι τα εξής:

- Τα μη δημοφιλή άρθρα, είναι πιο εύκολο να προβλεφθούν σε σχέση με τα δημοφιλή.
- Ήταν πιο εύκολο να προβλεφθούν τα tweets, παρά τα views ενός άρθρου.

Εν κατακλείδι, το γενικότερο συμπέρασμα της έρευνας, ήταν ότι με τα εργαλεία εκείνης της περιόδου (2014), δεν ήταν εφικτό να προβλεφθεί το αν ένα άρθρο θα γίνει δημοφιλές πριν τη δημοσίευσή του.

2.3 Εργαλείο για πρόβλεψη δημοτικότητας ειδήσεων και βελτιστοποίησης του άρθρου

Οι Kelwin Fernandes et al., πίστευαν πως το πιο σημαντικό για έναν αρθρογράφο, είναι ναι μεν να γνωρίζει από πριν εάν το άρθρο του θα γίνει δημοφιλές αλλά και να ξέρει τι θα μπορούσε να αλλάξει σε αυτό ώστε να πάει καλύτερα. Γι' αυτό το λόγο δημιούργησαν ένα εργαλείο το οποίο είχε τις ακόλουθες δυνατότητες:

- να προβλέπει τη δημοτικότητα ενός άρθρου, προτού δημοσιευθεί
- να προτείνει αλλαγές για τη βελτίωση του περιεχομένου και της δομής

Το dataset που χρησιμοποίησαν, είναι το ίδιο στο οποίο βασίζεται και η εν λόγω εργασία. Για να μπορέσουν να ελέγξουν την ακρίβεια της πρόβλεψης για ένα άρθρου, δοκίμασαν 5 αλγορίθμους: Random Forest, Adaptive Boosting, SVM, k-NN και Naïve Bayes.

Μέσω του Random Forest, ο οποίος αποδείχτηκε ο πιο αποδοτικός αλγόριθμος με ποσοστό καμπύλης ROC 73%, κατέληξαν στο ότι:

- η ύπαρξη keywords (λέξεις-κλειδιά) σε ένα άρθρο είναι άκρως σημαντική
- αλλάζοντας μερικά στοιχεία του άρθρου, μπορεί να αλλάξει και η μετέπειτα δημοτικότητά του

Οι ερευνητές, λοιπόν, θεώρησαν ότι αυτό το εργαλείο είναι εξαιρετικά χρήσιμο στους αρθρογράφους του Mashable και δήλωσαν ότι στο μέλλον θα ήθελαν να εμβαθύνουν ακόμα περισσότερο στα χαρακτηριστικά του περιεχομένου που μπορούν να επηρεάσουν ένα άρθρο.

ΚΕΦΑΛΑΙΟ 3: Έννοιες και εργαλείο

Σε αυτή εδώ την ενότητα, θα αναλύσουμε λίγο παραπάνω όλες τις σχετικές έννοιες που αφορούν την εν λόγω εργασία ενώ θα αναφερθούμε και στο εργαλείο με το οποίο πραγματοποιήθηκε όλη η έρευνα.

3.1 Εξόρυξη Γνώσης

Σύμφωνα με τον καθηγητή Jiawei Han του Πανεπιστημίου του Illinois, “We are drowning in data, but starving for knowledge”, δηλαδή βρισκόμαστε μπροστά σε ένα τεράστιο αριθμό δεδομένων χωρίς να κατανοούμε τι πραγματικά σημαίνουν αυτά για εμάς ή πως θα μπορούσαν να αποκτήσουν πραγματική αξία. Λόγω αυτής της αναγκαιότητας, λοιπόν, δημιουργήθηκε η έννοια της «Εξόρυξης της Γνώσης». Η έννοια αυτή αποτελεί την αυτοματοποιημένη εξαγωγή ενδιαφέρουσας και χρήσιμης πληροφορίας από μεγάλο όγκο δεδομένων. Για να φτάσουμε σε αυτή τη χρήσιμη πληροφορία, χρειάζεται να βρεθεί ένα μοτίβο ή αλλιώς μια σχέση μεταξύ των δεδομένων, η οποία να είναι κατανοητή από τον άνθρωπο και να ισχύει σε οποιαδήποτε νέα δεδομένα τυχόν προκύψουν στην εκάστοτε έρευνα.

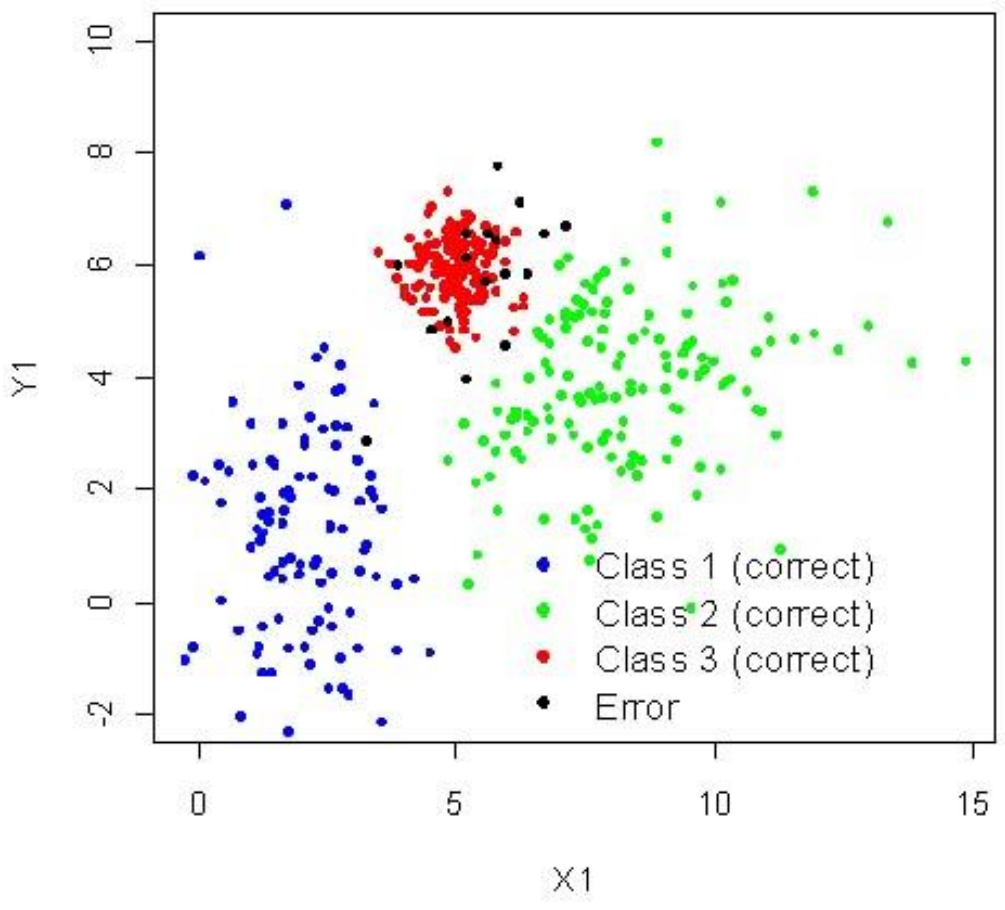
Για να μπορέσουμε, όμως, να φτάσουμε στην ανακάλυψη ενός μοτίβου και κατ’ επέκταση, στην Εξόρυξη της Γνώσης, υπάρχουν 2 βασικές μέθοδοι:

1. Κατηγοριοποίηση: διαδικασία εγγραφής ενός στοιχείου σε μια προ υπάρχουσα κλάση-κατηγορία. Στόχος είναι η σωστή πρόβλεψη της κλάσης αυτής.
2. Συσταδοποίηση: διαδικασία ανακάλυψης ομοιοτήτων στα υπάρχοντα δεδομένα. Στόχος είναι η δημιουργία νέων κλάσεων.
3. Κανόνες συσχέτισης: διαδικασία εύρεσης σχέσεων μεταξύ των μεταβλητών.

3.1.1 Κατηγοριοποίηση

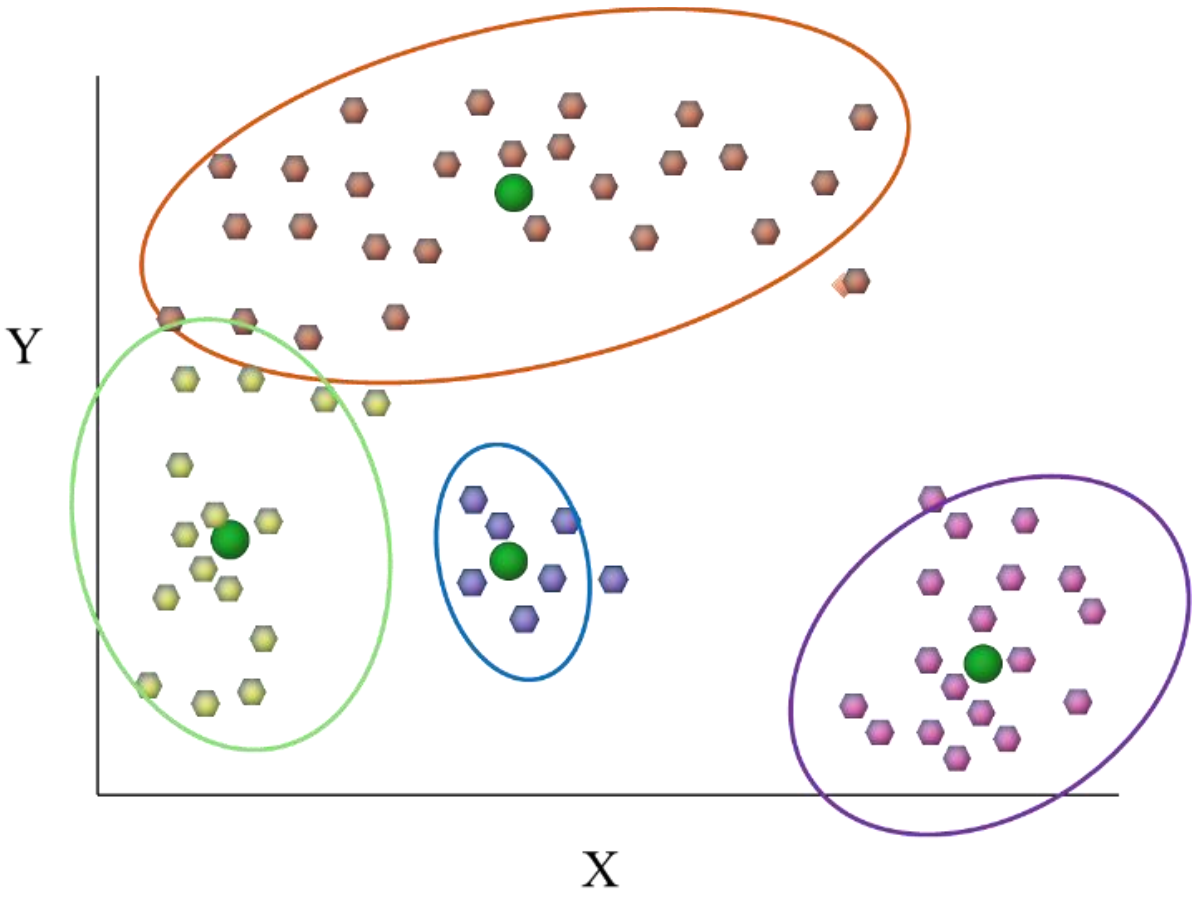
Όπως αναφέρθηκε και παραπάνω, η διαδικασία της κατηγοριοποίησης αποτελεί την ενσωμάτωση ενός στοιχείου σε μια υπάρχουσα κλάση-κατηγορία. Δηλαδή, σε ένα σύνολο δεδομένων, κάθε στοιχείο αποτελείται από κάποια γνωρίσματα. Ένα από αυτά τα γνωρίσματα, αποτελεί την κλάση, δηλαδή το χαρακτηριστικό το οποίο επιθυμούμε να προβλέψουμε. Σκοπός είναι να δημιουργηθεί ένα μοντέλο, το οποίο να μπορεί να βρίσκει αυτόματα την κλάση νέων δεδομένων.

Για τη δημιουργία του μοντέλου, χρησιμοποιούνται συνήθως αλγόριθμοι όπως δέντρα απόφασης, νευρωνικά δίκτυα κ.α. Αυτοί οι αλγόριθμοι μπορούν να επιλέξουν την επιθυμητή κλάση είτε ανάμεσα σε 2 (όπως συμβαίνει και στην περίπτωση της εργασίας αυτής λόγω δημοφιλίας ή μη ενός άρθρου) είτε ανάμεσα σε πολλές.



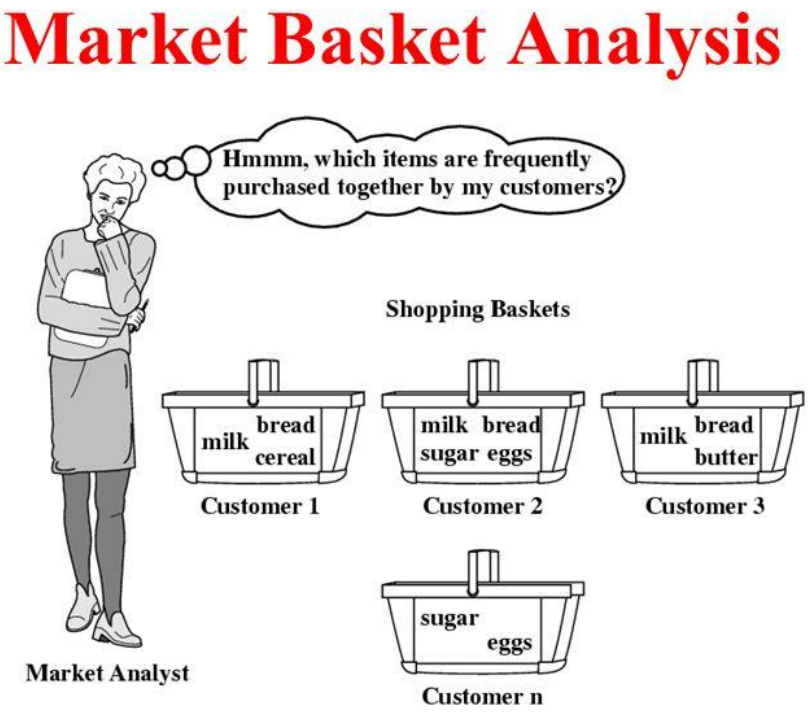
3.1.2 Συσταδοποίηση

Η συσταδοποίηση ή αλλιώς clustering, έχει ως στόχο την ανακάλυψη ομοιοτήτων μεταξύ των στοιχείων σε ένα σύνολο δεδομένων και τη δημιουργία διακριτών ομάδων-συστάδων. Αυτό σημαίνει ότι στοιχεία που βρίσκονται στην ίδια συστάδα είναι όμοια, ενώ στοιχεία που είναι σε διαφορετικές, θεωρούνται ως ανόμοια. Στη περίπτωση, λοιπόν, προσθήκης νέων στοιχείων στο σύνολο των δεδομένων, οι αλγόριθμοι συσταδοποίησης θα μπορέσουν να το κατατάξουν στις ανάλογες συστάδες, ανάλογα με τις ομοιότητες που διαθέτουν με τα άλλα στοιχεία. Ο αλγόριθμος που χρησιμοποιείται κατά κύριο λόγο για τη δημιουργία συστάδων, είναι ο K-means. Εκτός αυτού, υπάρχουν και αλγόριθμοι ιεραρχικής συσταδοποίησης οι οποίοι προσπαθούν να παράγουν μια ιεραρχία από συστάδες.



3.1.3 Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης αποτελούν μια από τις πιο σημαντικές μεθόδους εξόρυξης δεδομένων σε μεγάλες βάσεις. Οι πληροφορίες που μπορούν να προκύψουν από αυτές τις συσχετίσεις, μπορούν να είναι άκρως ενδιαφέρουσες και να έχουν ισχύ και εφαρμογή σε διάφορους τομείς της ζωής. Η αφορμή για την περαιτέρω μελέτη των κανόνων συσχέτισης, ήταν η ανάλυση του καλαθιού αγοράς στα σουπερμάρκετ, δηλαδή αν υπάρχει κάποια σχέση μεταξύ των προϊόντων που αγοράζουν οι καταναλωτές. Για παράδειγμα, αν όταν ένας πελάτης αγοράζει ψωμί, αν αγοράζει και βούτυρο. Αυτές οι πληροφορίες είναι εξαιρετικά χρήσιμες, σε θέματα marketing.

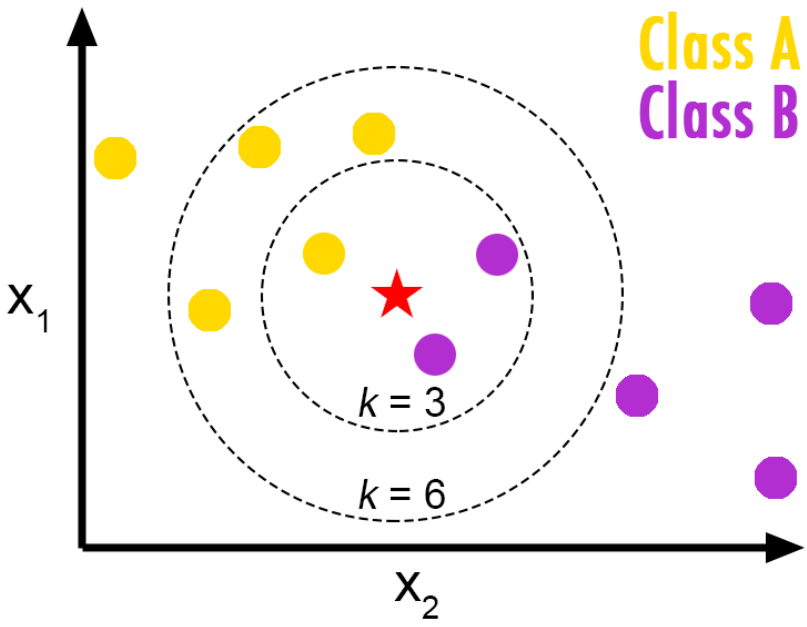


3.2 Αλγόριθμοι υλοποίησης μοντέλου

Σε αυτό το σημείο, θα αναφερθούμε στους αλγορίθμους που χρησιμοποιήθηκαν στην εν λόγω εργασία. Δεδομένου ότι ο σκοπός ήταν να βρούμε αν ένα άρθρο θεωρείται δημοφιλές ή όχι με βάση κάποιο χαρακτηριστικό, εφαρμόσαμε την τεχνική της κατηγοριοποίησης και τους αλγορίθμους: kNN, Naïve Bayes, Δέντρα Απόφασης (Decision Trees) και Random Forest.

Μια μικρή περιγραφή του εκάστοτε αλγορίθμου, παρουσιάζεται παρακάτω:

- **kNN:** η πλήρης ονομασία είναι k-nearest neighbor και βασίζεται στην Ευκλείδεια απόσταση. Αυτό σημαίνει ότι ο αλγόριθμος προσπαθεί να βρει που ανήκει ένα στοιχείο με βάση τους k-«κοντινότερους γείτονες» (όπου k=ακέραιος αριθμός), όπως φαίνεται και στην παρακάτω εικόνα. Ο συγκεκριμένος αλγόριθμος θεωρείται από τους πιο απλούς και μπορεί να χρησιμοποιηθεί τόσο σε μοντέλα κατηγοριοποίησης όσο και συσταδοποίησης.



- **Naïve Bayes:** Η βασική ιδέα του αλγορίθμου είναι να υπολογιστεί η πιθανότητα να ισχύουν δυο χαρακτηριστικά ταυτόχρονα. Για κάθε πιθανή κατηγορία, ο αλγόριθμος υπολογίζει την εκ των υστέρων πιθανότητα να ανήκει ένα στοιχείο σε αυτήν. Η κατηγορία η οποία έχει τη μεγαλύτερη πιθανότητα, είναι και αυτή στην οποία κατατάσσεται εν τέλει το στοιχείο.

Ο αλγόριθμος αυτό βασίζεται στο θεώρημα του Bayes κατά το οποίο ισχύει ότι:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Όπου:

P(h|D): η πιθανότητα να συμβεί το h, δεδομένου του D

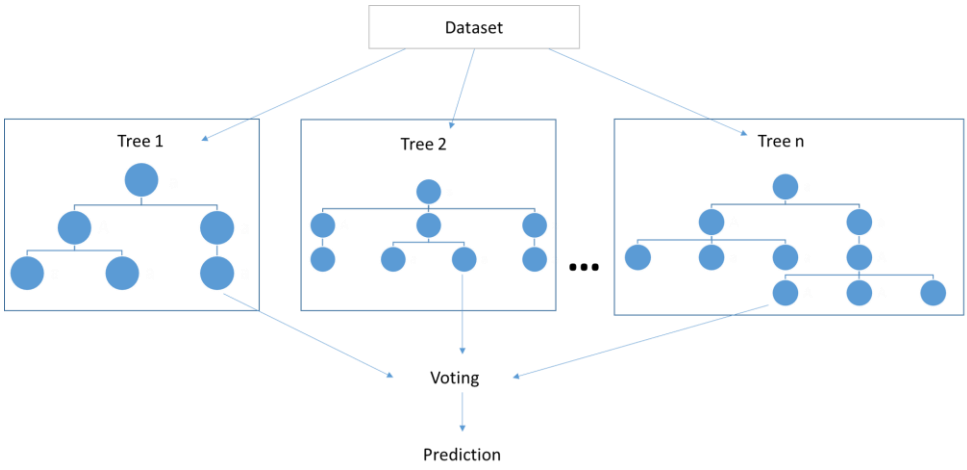
P(D|h): η πιθανότητα να συμβεί το D, δεδομένου του h

P(h): η πιθανότητα να συμβεί το h

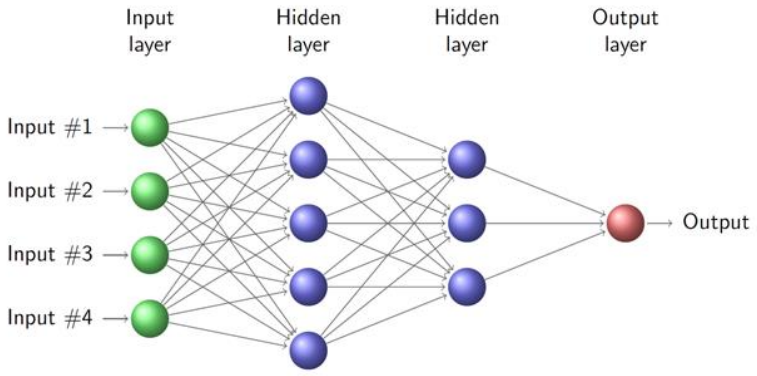
P(D): η πιθανότητα να συμβεί το D

Ο εν λόγω αλγόριθμος, είναι κι αυτός από τους αρκετά απλούς και κατανοητούς αλγορίθμους ο οποίος όμως είναι εξαιρετικά αποτελεσματικός.

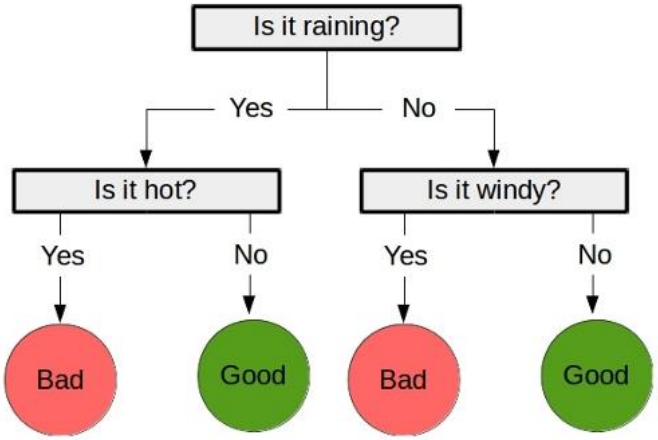
❖ **Random Forest:** δημιουργία πολλών δέντρων απόφασης. Κάθε δέντρο καταλήγει σε μια προτιμητέα κλάση και στο τέλος, το «τυχαίο δάσος» επιλέγει την κλάση με τις περισσότερες προτιμήσεις. Έχει τη δυνατότητα να χειριστεί μεγάλο όγκο δεδομένων ακόμα και αν υπάρχουν ελλείψεις σε αυτόν.



❖ **Neural Nets:** μοντέλα που αναπαριστούν τα ανθρώπινα νευρωνικά δίκτυα τα οποία μεταφέρουν πληροφορίες. Έχουν τη δυνατότητα να «μαθαίνουν» από τα υπάρχοντα δεδομένα και να αναγνωρίζουν μοτίβα μέσα σε αυτά. Επιπλέον, το μεγάλο πλεονέκτημα είναι η ανοχή τους στο θόρυβο που μπορεί να υπάρχει σε ένα σύνολο δεδομένων.



- **Δέντρα Απόφασης (Decision Tree):** ο χρήστης δημιουργεί ένα «δέντρο» κατά το οποίο οι εσωτερικοί κόμβοι αντιστοιχούν σε συνθήκες και έλεγχο αυτών και τα φύλλα στις σχετικές κατηγορίες/κλάσεις στις οποίες πρέπει να καταχωρηθεί ένα νέο στοιχείο. Βασικό χαρακτηριστικό αυτού του αλγορίθμου είναι ότι μπορεί να κατασκευάσει πολλά διαφορετικά δέντρα για ένα σύνολο δεδομένων.



3.3 PCA (Ανάλυση Πρωταρχικών Συνιστωσών)

Η μέθοδος PCA αποτελεί ένα εργαλείο γραμμικής συμπίεσης των δεδομένων, η οποία συνίσταται όταν το σύνολο είναι πολύ μεγάλο. Αυτό κρίνεται αναγκαίο, καθώς η πλειοψηφία των αλγορίθμων έχει καλύτερα αποτελέσματα, και ποιότητας αλλά και χρόνου, όταν οι διαστάσεις είναι λιγότερες. Τα δεδομένα μεταφέρονται σε ένα νέο σύνολο συντεταγμένων, μέσω του ορθογώνιου μετασχηματισμού, και δημιουργούν νέες συνιστώσες οι οποίες όμως διατηρούν την αρχική διακύμανσή τους. Ο αριθμός των συνιστωσών μπορεί να είναι είτε ίσος είτε μικρότερος από τα αρχικά γνωρίσματα των δεδομένων. Το πλεονέκτημα της εν λόγω μεθόδου, είναι ότι παρόλης της μείωσης του συνόλου των μεταβλητών, το 90% των δεδομένων παραμένει ίδιο άρα και ολόκληρη η πληροφορία.

3.4 Information Gain (Κέρδος Πληροφορίας)

Όταν ένα dataset έχει μεγάλο αριθμό γνωρισμάτων, όπως το δικό μας, χρειάζεται να βρεθεί με κάποιο τρόπο, ποιο απ' όλα αυτά τα γνωρίσματα μας δίνει τη περισσότερη πληροφορία. Η μετρική αυτή ονομάζεται Information Gain και υπολογίζεται μέσω της εντροπίας. Ο τύπος που μας δίνει το Information Gain ενός γνωρίσματος είναι:

$$\text{Information gain} = (\text{Entropy of distribution before the split}) - (\text{entropy of distribution after it})$$

Με λίγα λόγια, το κέρδος πληροφορίας, είναι η αλλαγή της εντροπίας πριν τη κατηγοριοποίηση και μετά. Συνηθέστερη χρήση του Information Gain, γίνεται στη περίπτωση των Decision Trees. Ωστόσο, το Information Gain μπορεί να μην είναι πάντοτε σωστή μετρική. Στη περίπτωση που ένα γνώρισμα μπορεί να λάβει πολύ μεγάλο αριθμό διαφορετικών τιμών (για παράδειγμα ένα γνώρισμα που αφορά τον αριθμό της πιστωτικής κάρτας), είναι προτιμότερο να χρησιμοποιείται η μετρική Information Gain Ratio.

3.5 Τρόπος και μετρικές αξιολόγησης

Ο πιο ενδεδειγμένος και ευρέως γνωστός, τρόπος αξιολόγησης ενός αλγορίθμου στην Εξόρυξη Δεδομένων, είναι ο «Πίνακας Σύγχυσης», ή κοινώς «Confusion Matrix». Αποτελεί μια οπτικοποίηση της απόδοσης ενός αλγορίθμου.

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

Οι γραμμές αποτελούν τις κλάσεις που προέβλεψε ο αλγόριθμος, ενώ οι στήλες είναι οι πραγματικές κλάσεις στις οποίες ανήκουν τα στοιχεία. Πιο συγκεκριμένα:

- TP-True Positive: Το στοιχείο ανήκει στη κλάση Positives και κατηγοριοποιήθηκε σωστά
- FP-False Positive: Το στοιχείο ανήκει στη κλάση Negatives και ο αλγόριθμος το κατηγοριοποίησε στην κλάση Positives
- FN-False Negative: Το στοιχείο ανήκει στη κλάση Positives και ο αλγόριθμος το κατηγοριοποίησε ως Negatives
- TN-True Negative: Το στοιχείο ανήκει στη κλάση Negatives και ο αλγόριθμος το κατηγοριοποίησε σωστά

Μέσω αυτού του πίνακα, μπορούν να υπολογιστούν κάποιες βασικές μετρικές αξιολόγησης. Μερικές από αυτές είναι:

- Accuracy-Ευστοχία: η επιτυχία ενός αλγορίθμου.

$$\frac{TP+TN}{TP+TN+FP+FN}$$

- Precision-Ακρίβεια: ποσοστό στοιχείων που ανήκουν πραγματικά στη κλάση Positives και ο αλγόριθμος τα έχει κατηγοριοποίηση σε αυτήν.

$$\frac{TP}{TP+FP}$$

- Recall-Ανάκληση: το ποσοστό των στοιχείων που ανήκουν στη κλάση Positive και κατάφερε να βρει ο αλγόριθμος.

$$\frac{TP}{TP+FN}$$

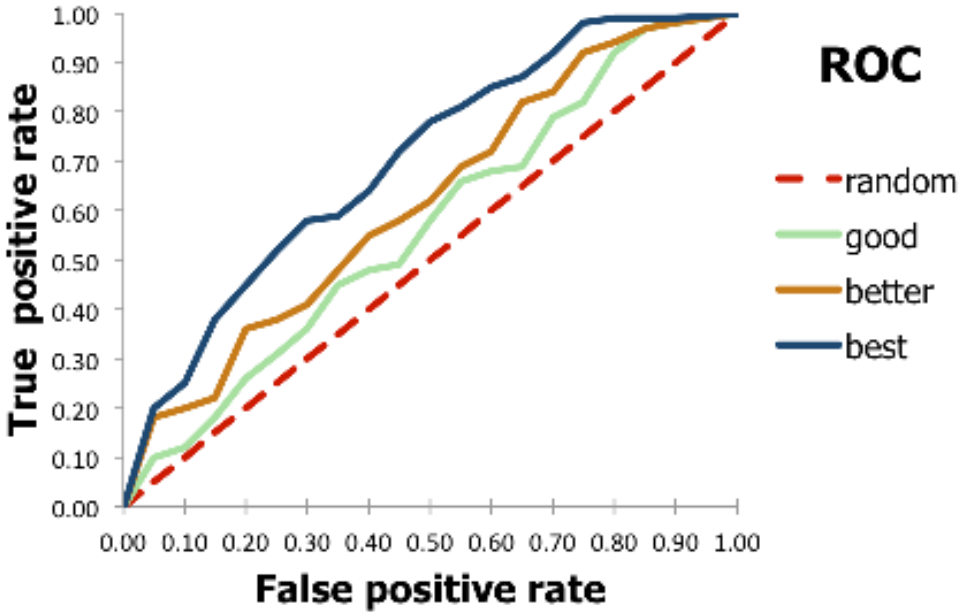
- F-Measure-Αρμονικός Μέσος: αποτελεί το μέσο όρο του precision και του recall.

$$F_1 = \frac{2}{1/r + 1/p}$$

❖ Καμπύλη ROC: καμπύλη μεταξύ αξόνων TP (στο κάθετο άξονα) και FP (στον οριζόντιο άξονα). Απεικονίζει το ποσοστό των TP και FP, τα οποία υπολογίζονται ως εξής:

$$\text{FPR} = \frac{FP}{TN + FP} \qquad \text{TPR} = \frac{TP}{TP + FN}$$

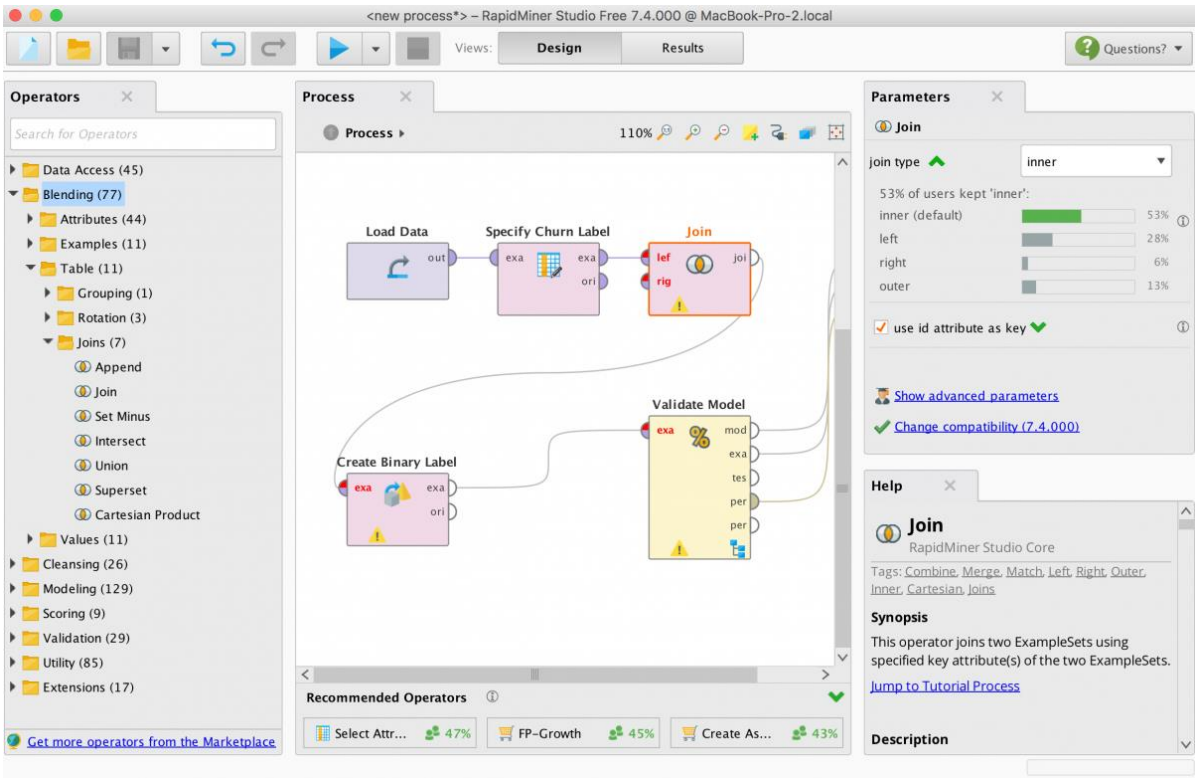
Το σημαντικότερο στοιχείο της καμπύλης ROC είναι η περιοχή που βρίσκεται μεταξύ αυτής και του οριζόντιου άξονα. Η περιοχή αυτή, μας δίνει το accuracy του αλγορίθμου οπότε όσο πιο μεγάλη αυτή, τόσο καλύτερος ο αλγόριθμος.



3.6 Εργαλείο υλοποίησης: RapidMiner

Το RapidMiner είναι μια πλατφόρμα που παρέχει ένα πλήρες οπτικοποιημένο περιβάλλον για μια ολοκληρωμένη διαδικασία ανάλυσης δεδομένων, από το στάδιο της προετοιμασίας αυτών μέχρι και την αξιολόγηση. Το αρχικό όνομα της πλατφόρμας, που ξεκίνησε να σχεδιάζεται το 2001, ήταν YALE (Yet Another Learning Environment). Ωστόσο, το 2013 το εργαλείο μετονομάστηκε στο όνομα με το οποίο είναι γνωστό έως σήμερα, RapidMiner.

Το RapidMiner είναι γραμμένο σε Java. Η διαδικασία της ανάλυσης δεδομένων, μπορεί να γίνει εύκολα μέσω του γραφικού περιβάλλοντος της πλατφόρμας στο οποίο ο χρήστης «σχεδιάζει» όλη τη πορεία της ανάλυσης. Το εργαλείο μπορεί, επίσης, είτε να κληθεί από άλλα προγράμματα είτε να χρησιμοποιηθεί ως API. Επιπλέον, υπάρχει η δυνατότητα χρήσης των αλγορίθμων μέσω της γραμμής εντολών, όπως και η επέκταση αυτών μέσω scripts σε R και Python.



Κεφάλαιο 4: Ανάλυση των δεδομένων και υλοποίηση

Σε αυτό το σημείο, θα αναλύσουμε πλήρως τα δεδομένα που χρησιμοποιήθηκαν, αλλά και όλη τη διαδικασία της υλοποίησης. Θα δούμε τα στάδια τόσο της προ επεξεργασίας των δεδομένων, καθαρισμού και τυχόν μετασχηματισμών, όσο και της υλοποίησης μέσω την εφαρμογή των αλγορίθμων που επιλέχθηκαν.

4.1 Περιγραφή δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην εν λόγω εργασία, είχαν συλλεχθεί από τους Fernandes et al., όπως αναφέρθηκε και στο κεφάλαιο 2 και αφορούσαν πάνω από 39.000 άρθρα τα οποία είχαν δημοσιευθεί στο Mashable.com σε διάστημα 2 χρόνων, από 7.1.2013 μέχρι και 7.1.2015. Κάθε άρθρο είχε τα εξής ενδεικτικά χαρακτηριστικά:

- διάστημα μεταξύ της ημέρας δημοσίευσης και της ημέρας ανάκτησης
- αριθμός λέξεων στο τίτλο και το περιεχόμενο
- αριθμός links στο κείμενο
- αριθμός links που προέρχονται από το Mashable.com
- αριθμός εικόνων και βίντεο
- μέσο μέγεθος λέξεων
- αριθμός keywords (λέξεις-κλειδιά)
- θεματολογία κειμένου

- μέγιστος, ελάχιστος και μέσος όρος αριθμός άλλων άρθρων τα οποία αναφέρονται στο εν λόγω άρθρο
- ημέρα δημοσίευσης
- εάν δημοσιεύθηκε σαββατοκύριακο ή όχι
- υποκειμενικότητα τίτλου και κειμένου
- συναίσθημα τίτλου και κειμένου
- ποσοστό θετικών (πχ φιλία, αυθεντικότητα κλπ.) και αρνητικών λέξεων (πχ κακός, ληστής, χείριστος κλπ.) στο κείμενο
- αριθμός shares (κοινοποιήσεις)
- κ.α.

Η στήλη shares, είναι αυτή η οποία μπορεί να φανερώσει με μια πρώτη ματιά και αυτή στην οποία θα βασιστούμε, ώστε να ελέγξουμε εάν κάποιο άρθρο έγινε δημοφιλές ή όχι. Το σύνολο αυτών των κοινοποιήσεων, προέρχεται από τον αριθμό των shares του άρθρου που πραγματοποιήθηκαν σε Facebook, Twitter, Google+, Linkedin, Pinterest και Stumble-Upon.

Το σύνολο αυτών των χαρακτηριστικών, ανερχόταν στα 61, αριθμός μεγάλος για μια ανάλυση. Γι' αυτό το λόγο, χρειάστηκε να προβούμε σε επεξεργασία αυτών.

4.2 Προεπεξεργασία των δεδομένων

Η προεπεξεργασία των δεδομένων είναι πάντοτε ένα πολύ σημαντικό σκέλος σε οποιαδήποτε ανάλυση προτού εφαρμοστούν οι αλγόριθμοι. Αυτό ισχύει, καθώς τα δεδομένα τα οποία καλείται κάποιος να χειριστεί, δεν είναι πάντοτε «καθαρά». Αυτό πρακτικά μπορεί να σημαίνει, πως είτε λείπουν κάποιες τιμές, είτε υπάρχουν λάθος καταχωρήσεις.

Για παράδειγμα, θα μπορούσε σε πεδίο που πρέπει να υπάρχει ακέραια τιμή, να έχει καταχωρηθεί κείμενο. Είναι, λοιπόν, εξαιρετικά αναγκαίο να ελεγχθούν τα δεδομένα και να περάσουν το στάδιο της επεξεργασίας, προκειμένου το αποτέλεσμα ολόκληρης της ανάλυσης να είναι σωστό και έγκυρο.

4.2.1 Καθαρισμός και μετασχηματισμός δεδομένων

Το πρώτο πράγμα που ελέγχθηκε στο εν λόγω dataset, ήταν η έλλειψη ή μη κάποιων τιμών. Ευτυχώς, δεν υπήρχαν εγγραφές, στις οποίες να λείπουν κάποιες τιμές. Ωστόσο, διαπιστώθηκε ότι υπήρχαν άρθρα τα οποία στο πεδίο του αριθμού των λέξεων του κειμένου, είχε καταγραφεί ο αριθμός 0. Κάτι τέτοιο σαφώς, δε θα μπορούσε να είναι πιθανό και γι’ αυτό το λόγο, αυτές οι εγγραφές διεγράφησαν. Έτσι, το dataset μας από τα 39.797 άρθρα, μειώθηκε στα 38.464.

Μετά από αυτή τη διαγραφή, ερχόμαστε στη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα. Στη περίπτωσή μας, τόσο τα θέματα όσο και οι μέρες δημοσίευσης, δεν καθορίζονται σε μία μονάχα στήλη. Αντιθέτως, υπάρχουν 7 διαφορετικές στήλες, μια για κάθε ημέρα, και 6 στήλες αντίστοιχα για κάθε πιθανό θέμα. Έτσι, κάθε άρθρο έχει τη τιμή 1 στην αντίστοιχη ημέρα δημοσίευσης και στο αντίστοιχο θέμα το οποία πραγματεύεται. Σε όλες τις άλλες στήλες, παίρνει την τιμή 0. Για να μειώσουμε αυτές τις στήλες, θα δημιουργήσουμε μια στήλη με τίτλο ημέρα και μια με θέμα και θα καταχωρηθούν με κείμενο οι επιθυμητές ημέρες και θέματα για κάθε εγγραφή. Με αυτό το τρόπο, τα γνωρίσματα από 61 γίνονται 50.

Ύστερα από τη δημιουργία των νέων χαρακτηριστικών, πραγματοποιήθηκε έλεγχος για την πιθανότητα έλλειψης ή ανακρίβειας των νέων τιμών. Πράγματι, παρατηρήσαμε ότι υπήρχαν περίπου 5.500 εγγραφές οι οποίες δεν είχαν συμπληρωμένη τιμή στη στήλη με τα θέματα.

Γι’ αυτό το λόγο, προχωρήσαμε στη διαγραφή αυτών των εγγραφών, κι έτσι το νέο σύνολό είναι 32.972 άρθρα από 38.464.

Σαν τελευταίο βήμα στο σκέλος του καθαρισμού και μετασχηματισμού του dataset, έπρεπε να οριστεί με κάποιο τρόπο η έννοια της δημοτικότητας. Δηλαδή, ποιο άρθρο θα έπρεπε να θεωρηθεί δημοφιλές και ποιο όχι. Αυτό, όπως αναφέρθηκε και παραπάνω, θα γινόταν με βάση τον αριθμό των κοινοποιήσεων του κάθε άρθρου.

Χρειαζόταν, λοιπόν, να οριστεί ένα «όριο» ή αλλιώς κατώφλι, με βάση το οποίο ο εκάστοτε αλγόριθμος, να μπορεί να αποφασίζει αν ένα άρθρο θα κατηγοριοποιηθεί ως δημοφιλές ή όχι. Δεδομένου του τεράστιο εύρους των κοινοποιήσεων (από 1 κοινοποίηση έως 843.300) ορίσαμε ως κατώφλι, το νούμερο των 1.350 κοινοποιήσεων, στο οποίο ολόκληρο το dataset να κατανέμεται ιδανικά. Έτσι, το σύνολο διαμορφώθηκε ως εξής:

- 16.293 δημοφιλή άρθρα
- 16.0949 μη δημοφιλή άρθρα

Μετά από αυτή τη κατηγοριοποίηση, δημιουργήθηκε νέα στήλη (shares) στην οποία, με βάση αυτό το κατώφλι, κάθε άρθρο έπαιρνε την τιμή «popular» ή «unpopular». Σε αυτή, ακριβώς τη στήλη, θα βασίζονταν και οι αλγόριθμοι αργότερα για να διαπιστώσουν αν υπάρχει τρόπος να προβλεφθεί η δημοτικότητα ενός άρθρου.

4.3 Δημιουργία μοντέλων

Το γεγονός ότι το dataset είχε τόσα πολλά γνωρίσματα, θα είχε ως αποτέλεσμα τη μη άμεση ολοκλήρωση των αλγορίθμων σε γρήγορο χρόνο. Επίσης, είναι πιθανό τα αποτελέσματα να μην ήταν πλήρως σωστά. Γι' αυτό το λόγο, πραγματοποιήσαμε 3 διαφορετικά μοντέλα. Βασική τους διαφορά, ήταν ο τρόπος με τον οποίο μειώναμε τα γνωρίσματα. Στη μια περίπτωση αυτό έγινε με την εφαρμογή της μεθόδου PCA, ενώ στην άλλη με βάση το Information Gain. Επίσης, το 1^{ο} μοντέλο ήταν μια απλή εφαρμογή των αλγορίθμων. Και στις 3 περιπτώσεις, οι classifiers που χρησιμοποιήθηκαν ήταν οι εξής:

1. kNN
2. Naïve Bayes
3. Neural Nets

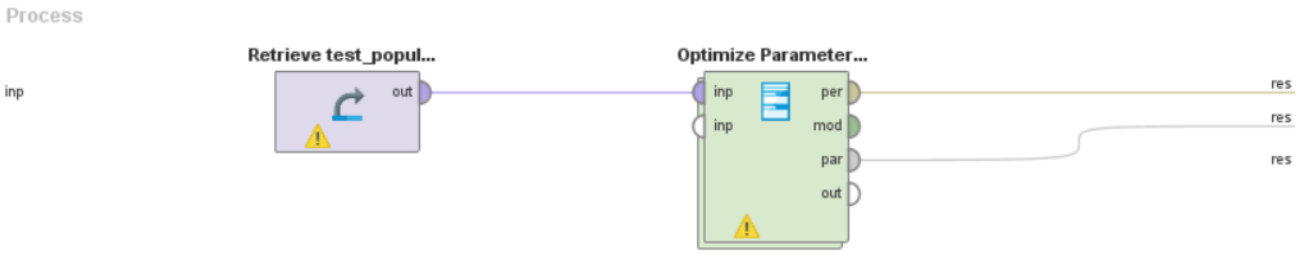
Επιπλέον, και στα 3 μοντέλα εφαρμόστηκε η διεργασία του Optimize Parameters. Αυτό έγινε καθώς κάθε αλγόριθμος έχει έναν αριθμό παραμέτρων. Προκειμένου, λοιπόν, να έχουμε πιο σωστά αποτελέσματα θα έπρεπε κανονικά να ορίζουμε με το χέρι τις τιμές αυτών των παραμέτρων. Κάτι τέτοιο σαφώς δε θα μπορούσε να πραγματοποιηθεί. Έτσι με τη χρήση αυτής της διεργασίες, πραγματοποιείται εξαντλητική μέθοδος πριν την εφαρμογή κάθε αλγορίθμου, με στόχο την εύρεση των πιο σωστών παραμέτρων. Τα όρια που είχε κάθε αλγόριθμος μέσα στα οποία έπρεπε να βρεθούν οι καταλληλότεροι παράμετροι, ήταν:

- ❖ kNN: k<=100
- ❖ Naïve Bayes: δεν έχει παραμέτρους
- ❖ Neural Nets: training cycles <=100

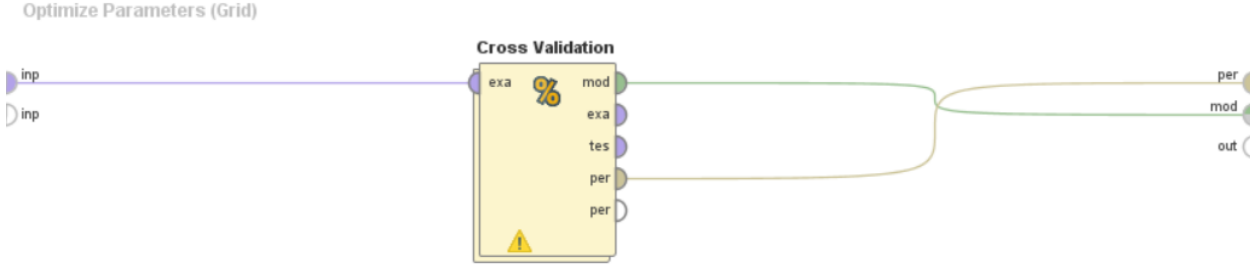
Παρακάτω θα δούμε αναλυτικά και τα 3 μοντέλα που δημιουργήθηκαν.

4.3.1 Εφαρμογή αλγορίθμων σε όλο το dataset

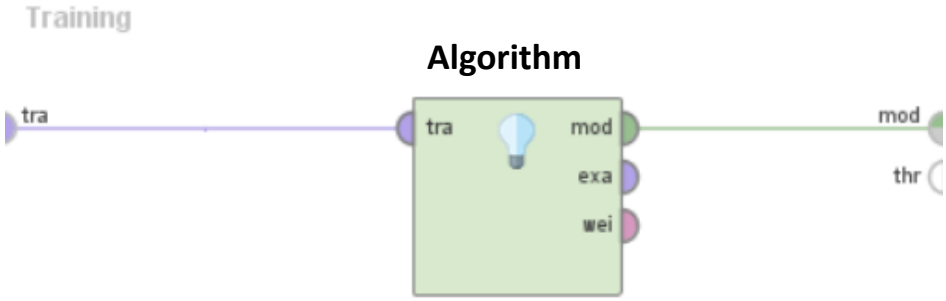
Το 1ο μοντέλο που δημιουργήθηκε ήταν μια απλή εφαρμογή των αλγορίθμων, χωρίς επιλογή συγκεκριμένων δεδομένων. Στο πρώτο μέρος του μοντέλου, πραγματοποιήθηκε η επιλογή των καταλληλότερων παραμέτρων για κάθε αλγόριθμο μέσω της διεργασίας Optimize Parameters.



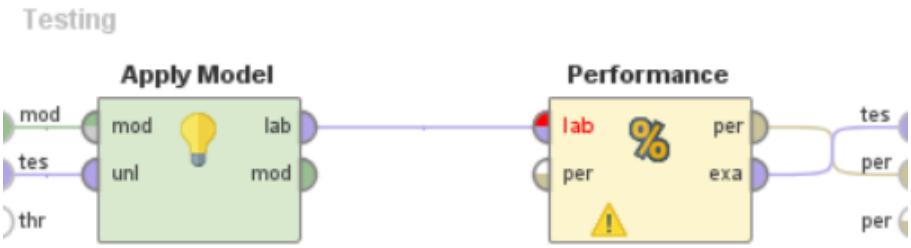
Η διεργασία αυτή, είναι μια εμφωλευμένη μέθοδος. Γι' αυτό το λόγο, στο επόμενο στάδιο προσθέσαμε τη λειτουργία του Cross Validation ώστε να αξιολογήσουμε το μοντέλο. Ο Cross Validation, επιλέγει από το σύνολο των δεδομένων, k ομάδες από τις οποίες η μια ομάδα εγγραφών θα χρησιμοποιηθεί στη διαδικασία του testing, ενώ όλες οι υπόλοιπες (k-1) θα χρησιμεύσουν για την εκπαίδευση του μοντέλου. Αυτή η επιλογή γίνεται μέσω της παραμέτρου της διεργασίας στην οποία ορίζεται ο αριθμός των ομάδων k, στις οποίες θα χωριστεί το σύνολο. Στη προκειμένη περίπτωση, επιλέξαμε τον αριθμό 10 καθώς είναι η πιο συνήθης επιλογή.



Η λειτουργία του Cross Validation χωρίζεται σε 2 μέρη: training και testing. Κατά τη διάρκεια του training πραγματοποιείται η απλή επιλογή του εκάστοτε αλγορίθμου, οι παράμετροι του οποίου καθορίζονται μέσω της διεργασίας Optimize Parameters, όπως αναφέρθηκε παραπάνω.

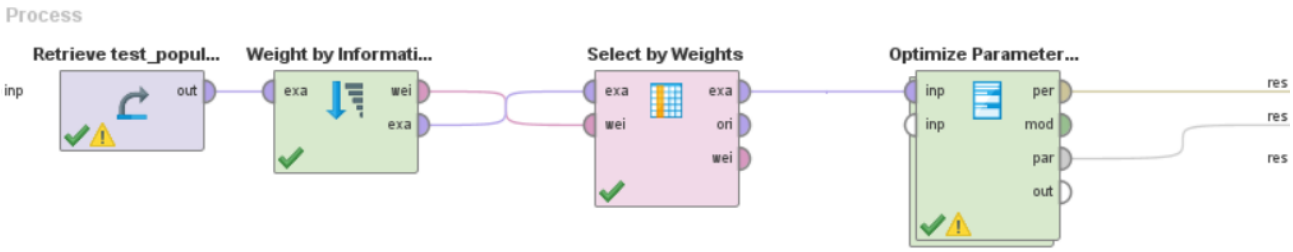


Στο σκέλος του testing, πραγματοποιείται η εφαρμογή του αλγορίθμου και η τελική αξιολόγηση του μοντέλου. Στη περίπτωση της αξιολόγησης, οι μετρικές που επιλέχθηκαν είναι το Accuracy, το Recall, το Precision, το F-Measure και η καμπύλη ROC.

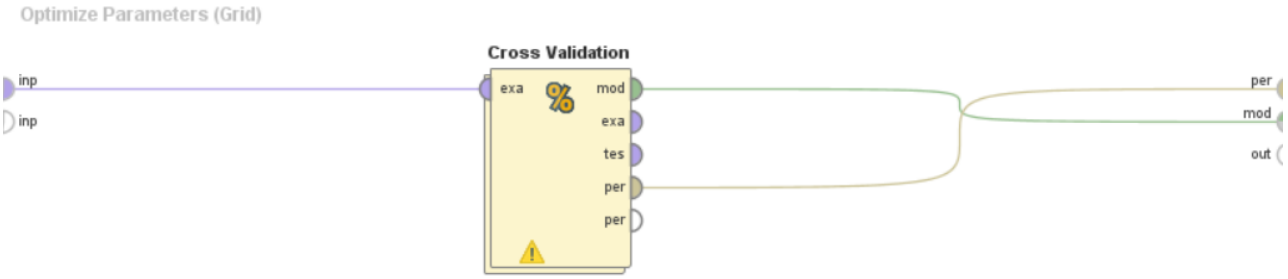


4.3.2 Εφαρμογή Information Gain

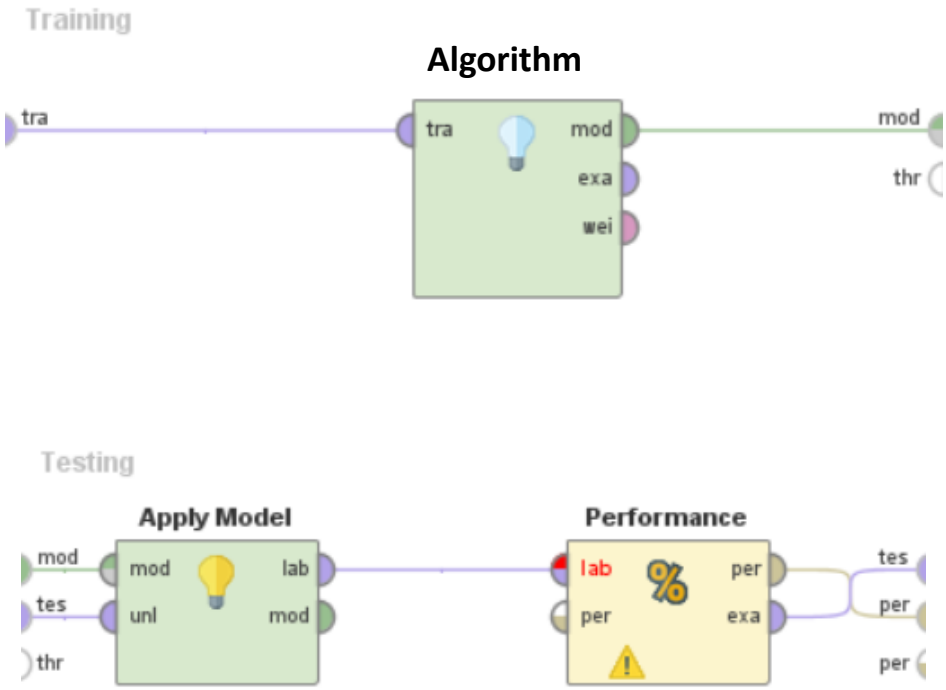
Το 2ο μοντέλο που δημιουργήθηκε, χρησιμοποίησε ως τρόπο μείωσης των γνωρισμάτων, το Information Gain. Μέσω αυτής της μεθόδου, μπορούμε να δούμε τα «βάρη» του κάθε γνωρίσματος σχετικά με το σύνολο των δεδομένων. Έπειτα, επιλέξαμε να δοκιμάσουμε το μοντέλο με τα 10 καλύτερα γνωρίσματα, όπου λέγοντας «καλύτερα» εννοούμε αυτά με το την υψηλότερη βαρύτητα. Στη συνέχεια, όπως και στο προηγούμενο μοντέλο, εφαρμόστηκε η διεργασία Optimize Parameters.



Όπως αναφέρθηκε και παραπάνω, η διεργασία Optimize Parameters, είναι εμφωλευμένη με αποτέλεσμα μέσα σε αυτήν να υπάρχει ο operator, Cross Validation. Οι παράμετροι αυτού του operator, είναι οι ίδιοι όπως και στο 1ο μοντέλο, δηλαδή χωρίσαμε κι εδώ το dataset σε 10 μέρη.

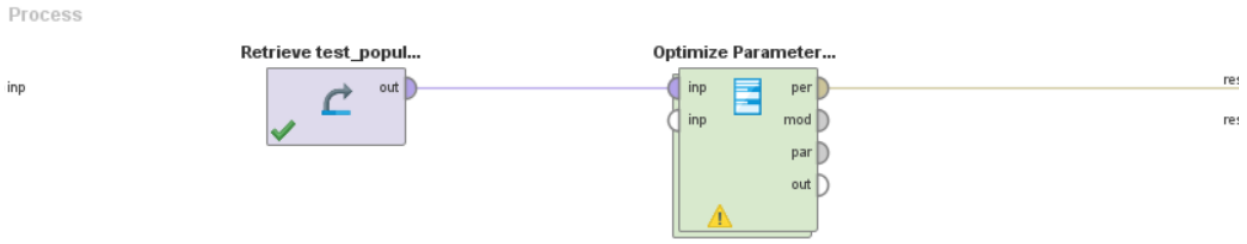


Στην συνέχεια, ακολουθεί η διαδικασία της εκπαίδευσης και του testing του μοντέλου. Όπως και στο προηγούμενο μοντέλο, τα πράγματα είναι πολύ πιο απλά και στα 2 αυτά σκέλη. Έτσι, στο σκέλος της εκπαίδευσης υπάρχει μονάχα ο αλγόριθμος που εφαρμόζεται κάθε φορά και στο σκέλος του testing, υπάρχει η εφαρμογή του αλγορίθμου και ο έλεγχος απόδοσής του.

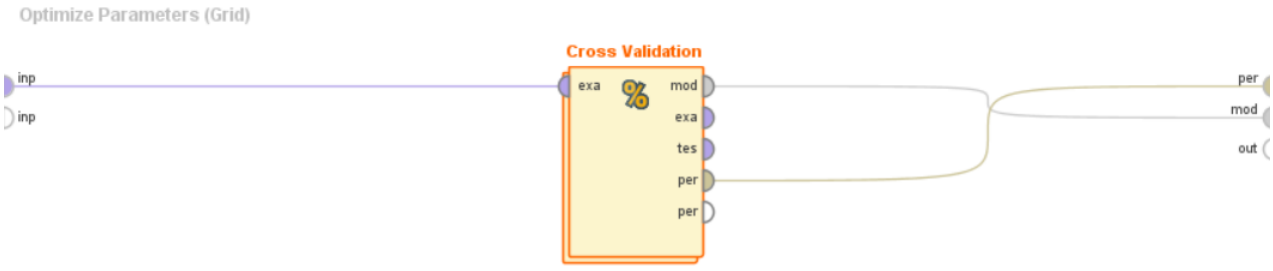


4.3.3 Εφαρμογή PCA

Το τελευταίο μοντέλο που υλοποιήσαμε είχε ως στόχο της εφαρμογή της μεθόδου PCA στο σύνολο των δεδομένων με σκοπό τη μείωση των γνωρισμάτων. Το πρώτο σκέλος του μοντέλου αφορά την εισαγωγή των δεδομένων και τη χρήση του operator, Optimize Parameters, όπως και στα προηγούμενα μοντέλα.



Το επόμενο βήμα είναι το ίδιο με τα υπόλοιπα μοντέλα. Πραγματοποιείται η εφαρμογή του Cross Validation.



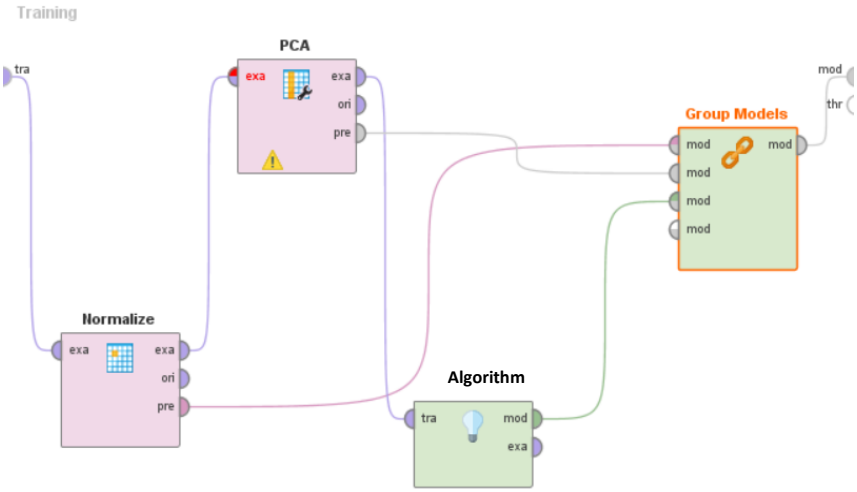
Περνώντας από το στάδιο του Cross Validation, μπαίνουμε στο τελευταίο σκέλος του μοντέλου, την εφαρμογή του αλγορίθμου και την τελική αξιολόγηση αυτού. Αυτό το σκέλος, χωρίζεται σε 2 υποεργασίες: την εκπαίδευση και το testing. Και τα 2 αυτά σκέλη, στο συγκεκριμένο μοντέλο, είναι πιο πολύπλοκα κι έτσι κρίθηκε σκόπιμο να αναλυθούν λίγο παραπάνω.

4.3.3.1 Εκπαίδευση

Όπως αναφέραμε και πριν, στο εν λόγω μοντέλο χρησιμοποιήθηκε η Ανάλυση Πρωταρχικών Συνιστωσών ή αλλιώς PCA. Η εφαρμογή της PCA είχε ως στόχο τη μείωση των διαστάσεων του συνόλου δεδομένων, με σκοπό τη καλύτερη και πιο γρήγορη εφαρμογή των αλγορίθμων. Ο αριθμός των διαστάσεων ήταν 10 ώστε να μπορέσει να υπάρξει μια καλύτερη σύγκριση με το 2^{ο} μοντέλο στο οποίο επιλέχθηκαν τα 10 γνωρίσματα με το καλύτερο Information Gain. Για να μπορέσει, όμως, να εφαρμοσθεί η PCA σε ένα dataset, χρειάζεται να γίνει κανονικοποίηση των τιμών, δηλαδή όλες οι τιμές των γνωρισμάτων να προσαρμοσθούν ώστε να ανήκουν σε ένα συγκεκριμένο εύρος τιμών.

Μετά την εφαρμογή της κανονικοποίησης και του PCA, είναι η ώρα για την επιλογή του αλγορίθμου. Όπως αναφέραμε και πριν, κάθε αλγόριθμος έχει ένα πλήθος παραμέτρων. Αυτές οι παράμετροι, ρυθμίζονται κάθε φορά από τη διεργασία Optimize Parameter. Τέλος, κάθε μια από αυτές τις διεργασίες (κανονικοποίηση, PCA, αλγόριθμος) δίνουν ένα συγκεκριμένο σύνολο δεδομένων.

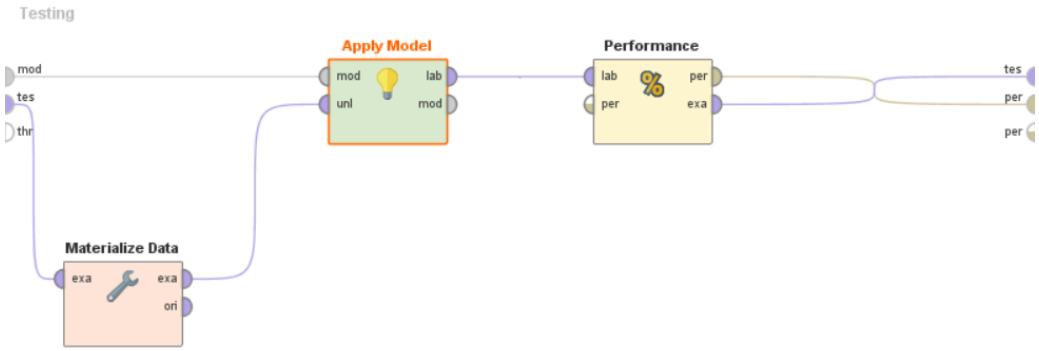
Για να μπορέσουν αυτά τα 3 σύνολα να συγχωνευτούν σε ένα, χρησιμοποιείται η διεργασία Group Models, η οποία μας δίνει το τελικό σύνολο της εκπαίδευσης.



4.3.3.2 Testing

Μπαίνοντας στη διαδικασία του testing, το πρώτο πράγμα που γίνεται, είναι η δημιουργία ενός αντιγράφου του dataset πριν την εφαρμογή του αλγορίθμου. Αυτό είναι χρήσιμο ώστε στο τέλος της όλης διαδικασίας να υπάρχει ένα σύνολο δεδομένων και πριν το αλγόριθμο αλλά και μετά. Αυτό το αντίγραφο στο RapidMiner, δημιουργείται με τη διεργασία Materialize Data. Στη συνέχεια, πραγματοποιείται η εφαρμογή του αλγορίθμου μέσω της διεργασίας Apply Model.

Τέλος, το dataset που δημιουργείται από την εφαρμογή του αλγορίθμου, περνάει από την αξιολόγηση ή αλλιώς Performance. Δεδομένου, των όσων αναφέραμε στο κεφάλαιο 3 σχετικά με τις μετρικές αξιολόγησης των μοντέλων, επιλέξαμε να αξιολογήσουμε τα μοντέλα μας με βάση το Accuracy, την καμπύλη ROC, το Precision, το Recall και το F-Measure.


Testing
mod
tes
thr
Apply Model
mod
lab
unl
mod
Performance
lab
per
per
exa
tes
per
per
Materialize Data
exa
exa
ori

Κεφάλαιο 5: Αποτελέσματα και Συμπεράσματα

Στο τελευταίο κεφάλαιο, θα αναφερθούν τα αποτελέσματα και τα συμπεράσματα που προέκυψαν και από τα 3 μοντέλα με τη χρήση των μετρικών:

1. Accuracy
2. Recall
3. Precision
4. F-Measure
5. Καμπύλη ROC

Στο 1^{ο} μοντέλο, οι αλγόριθμοι εφαρμόστηκαν σε όλο το dataset χωρίς κάποιο περιορισμό στα χαρακτηριστικά. Τα αποτελέσματα στα οποία κατέληξε κάθε αλγόριθμος, είναι τα ακόλουθα:

	Accuracy	Recall	Precision	F-Measure	Καμπύλη ROC
kNN	59%	54%	59%	56%	62%
Naïve Bayes	**77%**	**98%**	**69%**	**81%**	**93%**
Neural Nets	64%	63%	63%	63%	69%

Όπως φαίνεται, ο αλγόριθμος Naïve Bayes, φαίνεται να είναι ο καταλληλότερος καθώς έχει 5/5 μετρικές το υψηλότερο ποσοστό. Δεύτερος καλύτερος, έρχονται τα Neural Nets και τελευταίος ο kNN.

Στη περίπτωση του 2ου μοντέλου, εφαρμόστηκε η μέθοδος μείωσης των γνωρισμάτων με βάση το Information Gain. Τα γνωρίσματα κατατάχθηκαν ανάλογα με το κέρδος πληροφορίας του και έπειτα, χρησιμοποιήθηκαν τα top 10 χαρακτηριστικά. Τα αποτελέσματα τα οποία προέκυψαν είναι τα εξής:

	Accuracy	Recall	Precision	F-Measure	Καμπύλη ROC
kNN	58%	56%	57%	57%	61%
Naïve Bayes	**88%**	**96%**	**82%**	**89%**	**97%**
Neural Nets	66%	65%	67%	65%	74%

Και σε αυτό το μοντέλο, ο καλύτερος κατηγοριοποιητής φαίνεται να είναι ο Naïve Bayes. Αυτό που μπορεί να επισημανθεί, είναι ότι οι μετρικές του Accuracy, του Precision και του F-Measure έχουν αυξηθεί τόσο για το Naïve Bayes όσο και για τα Neural Nets τα οποία είναι κι εδώ ο δεύτερος καλύτερος αλγόριθμος.

Στο τελευταίο μοντέλο, επιλέχθηκε η μέθοδος του PCA (με 10 principal components) για τη μείωση των γνωρισμάτων. Στον ακόλουθο πίνακα μπορούμε να δούμε τη σχετική αξιολόγηση των αλγορίθμων.

	Accuracy	Recall	Precision	F-Measure	Καμπύλη ROC
kNN	61%	58%	61%	59%	66%
Naïve Bayes	52%	**92%**	50%	**65%**	60%
Neural Nets	61%	61%	61%	60%	66%

Κι εδώ, φαίνεται ο Naïve Bayes να πηγαίνει καλύτερα σε σχέση με τους υπόλοιπους αλγορίθμους καθώς καταφέρνει να πετύχει 92% Recall.

Συγκρίνοντας τα νούμερα όλων των μοντέλων, παρατηρείται ότι ο kNN έχει καλύτερα αποτελέσματα όταν γίνεται χρήση της μεθόδου PCA. Αντίθετα, ο Naïve Bayes και τα Neural Nets, έχουν πολύ καλύτερη αξιολόγηση με την εφαρμογή του Information Gain.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. *ICWSM*, *12*, 26-33.

Arapakis, I., Cambazoglu, B. B., & Lalmas, M. (2014, November). On the feasibility of predicting news popularity at cold start. In *International Conference on Social Informatics* (pp. 290-299). Springer, Cham.

Fernandes, K., Vinagre, P., & Cortez, P. (2015, September). A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence* (pp. 535-546). Springer, Cham.