ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΣΩΠΟΠΟΙΗΜΕΝΗ ΣΥΣΤΑΣΗ ΛΟΓΑΡΙΑΣΜΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

Κωνσταντινίδης Κωνσταντίνος

Επιβλέπουσα: Τουσίδου Ελένη

Σεπτέμβρης 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΣΩΠΟΠΟΙΗΜΕΝΗ ΣΥΣΤΑΣΗ ΛΟΓΑΡΙΑΣΜΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διπλωματική Εργασία

Κωνσταντινίδης Κωνσταντίνος

Επιβλέπουσα: Τουσίδου Ελένη

Σεπτέμβρης 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

PERSONALIZED RECOMMENDATION OF SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING ALGORITHMS

Diploma Thesis

Konstantinidis Konstantinos

Supervisor: Tousidou Eleni

September 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα **Τουσίδου Ελένη**
Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Βασιλακόπουλος Μιχαήλ**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Φεύγας Αθανάσιος**
Ε.ΔΙ.Π, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Η εργασία εκπονήθηκε με τη βοήθεια της επιβλέπουσας κυρίας Ελένης Τουσίδου, την οποία ευχαριστώ για τις συμβουλές που μου έδωσε για τη διεκπεραίωση της παρούσας εργασίας. Ευχαριστώ επίσης,τον κύριο Βασιλακόπουλο και τον κύριο Φεύγα για τη συμμετοχή τους ως μέλη της επιτροπής. Τέλος, η εργασία αυτή και η ολοκλήρωση των σπουδών μου δε θα ήταν δυνατή χωρίς τη συμβολή της οικογένειας και των φίλων μου που με στήριξαν σε όλη τη διάρκεια της φοίτησής μου.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

Κωνσταντινίδης Κωνσταντίνος

Διπλωματική Εργασία

ΠΡΟΣΩΠΟΠΟΙΗΜΕΝΗ ΣΥΣΤΑΣΗ ΛΟΓΑΡΙΑΣΜΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Κωνσταντινίδης Κωνσταντίνος

Περίληψη

Η χρήση των μέσων κοινωνικής δικτύωσης έχει διαδοθεί στον παγκόσμιο πληθυσμό, ενώ καταλαμβάνει ένα μεγάλο μέρος των ημερήσιων δραστηριοτήτων των ανθρώπων. Η ταχεία υιοθέτηση αυτών των τεχνολογιών αλλάζει τον τρόπο με τον οποίο επικοινωνούμε με τους φίλους μας, τον τρόπο πρόσβασης σε πληροφορίες, τις πηγές των ειδήσεών μας. Οι λογαριασμοί που ο κάθε χρήστης ακολουθεί αποτελούν σημαντικό μέρος της εμπειρίας του στη πλατφόρμα που επιλέγει να χρησιμοποιήσει.

Τα περισσότερα μέσα κοινωνικής δικτύωσης προτείνουν λογαριασμούς σε κάθε χρήστη, με βάση αυτήν του τη δραστηριότητα. Στη παρούσα διπλωματική εργασία θα χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης, για τη δημιουργία ενός συστήματος με προσωποποιημένες συστάσεις για κάθε χρήστη μιας πλατφόρμας κοινωνικής δικτύωσης. Τα δεδομένα που θα χρησιμοποιήσουμε συνθέτουν ένα κατευθυνόμενο γράφημα με λογαριασμούς του Facebook, που δείχνει ποιος χρήστης ακολουθεί ποιον, είναι ανώνυμα και παρέχονται από τη πλατφόρμα του Facebook.

Λέξεις-κλειδιά:

Μέσα κοινωνικής δικτύωσης, Μηχανική Μάθηση, Συστήματα Συστάσεων

Diploma Thesis

PERSONALIZED RECOMMENDATION OF SOCIAL MEDIA ACCOUNTS USING MACHINE LEARNING ALGORITHMS

Konstantinidis Konstantinos

Abstract

The use of social media has spread to a huge part of the world's population, taking up a large part of many people's day. The rapid adoption of these technologies is changing the way we communicate with our friends, the way we access information, our news sources. The accounts that each user chooses to follow are an important part of his experience when using the platform he or she chooses.

Most social media recommend accounts to their users, based on the aforementioned experience. In this thesis, machine learning algorithms will be used to create and compare recommendation systems offering personalized recommendations for the users of a social media platform. The data that will be used compose a directed graph of Facebook accounts that shows which user follows whom, it is anonymous and is provided by the Facebook platform.

Keywords:

Social Networks, Machine Learning, Recommender Systems

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

κ.α.	και άλλα
WCC	Weakly Connected Components

Κεφάλαιο 1

Εισαγωγή

Η πρόβλεψη συνδέσμων είναι ένα θεμελιώδες πρόβλημα σε μεγάλα και σύνθετα δίκτυα. Σε αυτού του είδους τα προβλήματα, μας δίνεται ένα στιγμιότυπο ενός δικτύου και βάση αυτού συμπεραίνουμε ποιες αλληλεπιδράσεις μεταξύ των υπαρχόντων κόμβων είναι πιθανό να προκύψουν στο εγγύς μέλλον.

Η πρόβλεψη συνδέσμων έχει πολλές εφαρμογές. Μια από αυτές είναι η αναδόμηση δικτύων, η οποία έχει ως στόχο όταν αφαιρεθεί μια ακμή από ένα γράφημα να μπορέσει να προβλέψει ποια είναι η ακμή που αφαιρέθηκε. Ακόμα, οι προβλέψεις συνδέσμων χρησιμοποιούνται για την αναγνώριση των ανεπιθύμητων αλληλογραφιών, που είναι αρκετά επιβλαβείς σε ένα σύστημα καθώς καταναλώνουν μεγάλο εύρος ζώνης και μνήμη από το δίκτυο. Επίσης, συμβάλλουν στην αναγνώριση των αναφορών που μπορεί να λείπουν από μια επιστημονική δημοσίευση, καθώς εντοπίζοντας αγνοούμενες αναφορές βοηθάει στην αποφυγή λογοκλοπής. Στην παρούσα εργασία τα δίκτυα που θα ασχοληθούμε είναι αυτά των μέσων κοινωνικής δικτύωσης. Πιο συγκεκριμένα, θα γίνει χρήση των μέσων δικτύωσης για την πρόβλεψη και σύσταση λογαριασμών προς ακολούθηση στους χρήστες της εκάστοτε πλατφόρμας.

Τα μέσα κοινωνικής δικτύωσης επικεντρώνονται κυρίως στη δημιουργία και την επέκταση των κοινωνικών αλληλεπιδράσεων μεταξύ των χρηστών. Για να εντείνουμε τις αλληλεπιδράσεις βασιζόμαστε κυρίως σε κοινά ενδιαφέροντα, κοινούς φίλους κ.λπ. Αυτό οδηγεί αναπόφευκτα, στο δίκτυο να επεκταθεί στο χρόνο, να εγγραφούν νέοι χρήστες και να προστεθούν συνδέσεις μεταξύ παλιών και νέων χρηστών. Με βάση το τρέχον δίκτυο θέλουμε να μπορούμε να προβλέψουμε τις επερχόμενες αλλαγές του και να κάνουμε τις ανάλογες συστάσεις.

Οι χρήστες που ακολουθεί ο κάθε λογαριασμός στις πλατφόρμες κοινωνικών μέσων θα επηρεάσουν σημαντικά την εμπειρία του. Εξάλλου, οι αναρτήσεις από τους λογαριασμούς που ακολουθεί κανείς, είναι αυτές που συμπληρώνουν τη ροή της κάθε πλατφόρμας κοινωνικών μέσων. Αποτελεί λοιπόν, σημαντικό παράγοντα της πορείας της κάθε κοινωνικής πλατφόρμας οι συστάσεις που θα κάνει η ίδια.

Το Facebook παρέχει ένα μεγάλο δείγμα κοινωνικού δικτύου σε μια χρονική περίοδο και, βάσει αυτού, θα προβλέψουμε τους μελλοντικούς πιθανούς συνδέσμους. Ακόμα θα προτείνουμε νέες συνδέσεις μεταξύ χρηστών που θα βασιστούν στην έως τώρα μορφή του δικτύου. Σε αυτήν την κατεύθυνση, σημαντικό μέρος της εργασίας θα αποτελέσει η εξαγωγή χαρακτηριστικών (feature extraction) από το δίκτυο μας, καθώς τα δεδομένα που μας παρέχονται είναι αποκλειστικά οι συνδέσεις μεταξύ των χρηστών.

1.1 Αντικείμενο της διπλωματικής

Ο στόχος της διπλωματικής εργασίας είναι η πρόβλεψη αλλά και η σύσταση λογαριασμών μέσων κοινωνικής δικτύωσης προς ακολούθηση σε χρήστες της εκάστοτε πλατφόρμας. Στη συγκεκριμένη περίπτωση, χρησιμοποιούνται δεδομένα που παρέχει το Facebook ενός κοινωνικού δικτύου [2], σε μια χρονική περίοδο της ύπαρξης του. Τα δεδομένα αποτελούνται από ενώσεις μεταξύ λογαριασμών, που δημιουργούν μεταξύ τους ένα κατευθυνόμενο γράφημα. Η πρόβλεψη αλλά και η σύσταση των λογαριασμών θα γίνει με αλγορίθμους μηχανικής μάθησης αλλά και νευρωνικών δικτύων. Στα μοντέλα μηχανικής μάθησης χρησιμοποιήθηκαν boosting αλγόριθμοι αλλά και ο Random Forest αλγόριθμος. Πιο συγκεκριμένα, χρησιμοποιούνται ο XGBoost αλλά και ο LightGBM της Microsoft.

1.2 Οργάνωση του τόμου

Η εργασία αποτελείται από έξι κεφάλαια στο σύνολό της. Στο πρώτο αναφέρονται η εισαγωγή και ο σκοπός της εργασίας. Στο δεύτερο κεφάλαιο περιλαμβάνονται εν συντομία εργασίες παρόμοιας θεματολογίας με την παρούσα. Στο τρίτο κεφάλαιο γίνεται επεξήγηση των όρων και του θεωρητικού υπόβαθρου που χρειάζεται να υπάρχει για να κατανοηθεί η εργασία. Ακόμα, στο τέταρτο κεφάλαιο αναλύονται τα δεδομένα και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας. Στο πέμπτο κεφάλαιο αναλύονται και επε-

ξηγούνται σε βάθος τα μοντέλα που υλοποιήθηκαν για τη διαδικασία της πρόβλεψης και της σύστασης. Τέλος, στο έκτο κεφάλαιο συνοψίζονται οι αποδόσεις των μοντέλων και γίνεται επεξήγηση των αποτελεσμάτων που αποκομίσαμε.

Κεφάλαιο 2

Πρόβλεψη Συνδέσμων και Συναφείς Εργασίες

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιαστούν προηγούμενες μελέτες με τεχνικές και στρατηγικές που έχουν χρησιμοποιήσει την πρόβλεψη συνδέσμων για συστάσεις σε μέσα κοινωνικής δικτύωσης.

2.2 Πρόβλεψη συνδέσμων με επιβλεπόμενη μάθηση

Στο πρόβλημα πρόβλεψης συνδέσμων με επίβλεψη, το αρχικό πρόβλημα μετατρέπεται σε ένα πρόβλημα κατηγοριοποίησης. Αυτή είναι και η τακτική που εφαρμόστηκε και στη παρούσα εργασία. Αυτό το πρόβλημα δυαδικής κατηγοριοποίησης, λύνεται συνήθως με αλγορίθμους όπως δέντρα αποφάσεων (decision trees), SVM (Support Vector Machines), νευρωνικά δίκτυα και άλλα που θα δούμε στη συνέχεια της εργασίας.

Μια σύνοψη όλων των μεθόδων που μπορούν να χρησιμοποιηθούν για τη πρόβλεψη συνδέσμων στα μέσα κοινωνικής δικτύωσης, παρουσιάζουν οι Wang Peng, Xu Baowen, Wu Yurong και Zhou Xiaoyu [1]. Οι δύο μέθοδοι που μπορούν να χρησιμοποιηθούν είναι αυτοί της ομοιότητας χρηστών και των μεθόδων εκμάθησης. Στη μέθοδο εκμάθησης γίνεται χρήση και των μεθόδων ομοιότητας, ως χαρακτηριστικά, σε μοντέλα μηχανικής μάθησης που χρησιμοποιούνται για κατηγοριοποίηση.

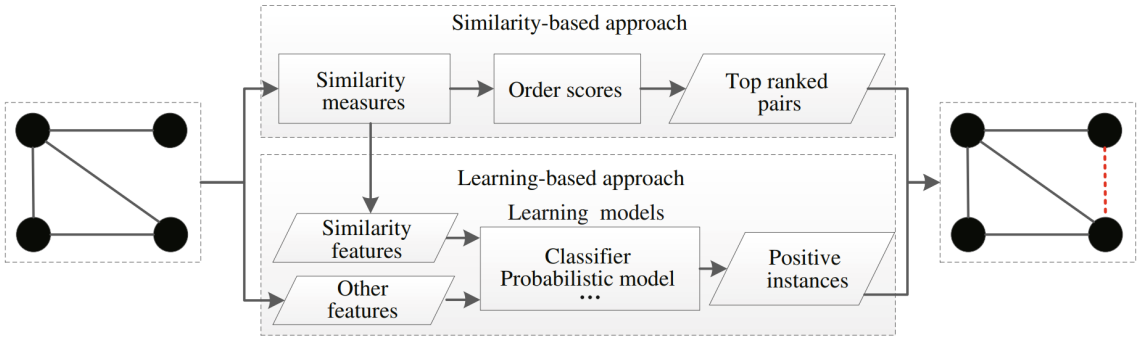
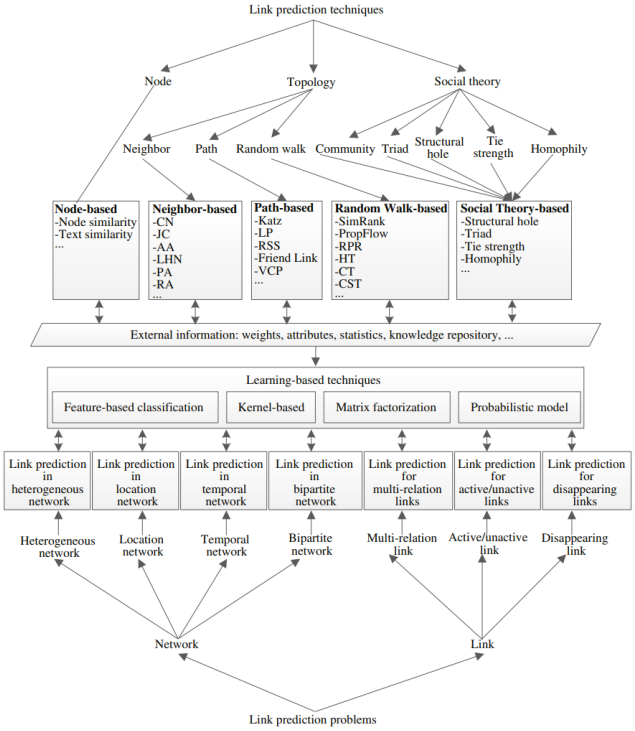


Figure 7 The generic link prediction framework.

Σχήμα 2.1: Μεθοδολογία πρόβλεψης κόμβων (Πηγή: [1])

Οι τεχνικές εκμάθησης μπορούν να παρουσιαστούν σε διάφορες μορφές αναλόγως τον τύπο του γραφήματος που μας παρέχεται και των χαρακτηριστικών που μπορούμε να εξάγουμε από αυτό. Τα χαρακτηριστικά αυτά μπορεί να είναι, όπως προαναφέρθηκε, ομοιότητας, κοινών γειτόνων κ.α. Στο σχήμα παρακάτω, παρουσιάζονται αναλυτικά οι τεχνικές που μπορούν να χρησιμοποιηθούν για τις προβλέψεις συνδέσμων.



Σχήμα 2.2: Σύνοψη όλων των τεχνικών πρόβλεψης κόμβων

Η πηγή του σχήματος βρίσκεται στη δημοσίευση [1]

Μία από τις κύριες προκλήσεις του συστήματος αυτού είναι η εξαγωγή του κατάλληλου σετ χαρακτηριστικών. Οι David Liben-Nowell και Jon Kleinberg στο [3] παρουσιάζουν την πρόβλεψη ακμών σε ένα δίκτυο μέσων κοινωνικής δικτύωσης. Στη δημοσίευση γίνεται μια εκτενής επεξήγηση των μεθόδων που μπορούν να χρησιμοποιηθούν για την εξαγωγή του σωστού συνόλου δεδομένων που μπορεί να χρησιμοποιηθεί για την δυαδική κατηγοριοποίηση.

Οι William Cukierski, Benjamin Hamner, Bo Yang [4] χρησιμοποίησαν επιβλεπόμενη μάθηση στο σύνολο δεδομένων του Flickr, για τις συσχετίσεις μεταξύ εικόνων, που περιλαμβάνει ένα κατευθυνόμενο γράφημα με 7.237.983 ακμές και 1.133.547 κόμβους. Στη δημοσίευση δείχνουν την απόδοση της κάθε μεθόδου μη επιβλεπόμενης μάθησης και την επίδοση των αλγορίθμων μηχανικής μάθησης. Στην υλοποίηση τους, χρησιμοποίησαν τόσο Local Neighbor, όσο και Global Neighbor αλγορίθμους ως χαρακτηριστικά για τους αλγορίθμους μηχανικής μάθησης που χρησιμοποίησαν. Αυτή τη κατεύθυνση θα έχει και η παρούσα εργασία, κάνοντας εξαγωγή χαρακτηριστικών από το γράφημα με σκοπό να χρησιμοποιηθούν για την εκπαίδευση του συστήματός μας.

Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

3.1 Εισαγωγή

Στο κεφάλαιο αυτό περιγράφονται οι αλγόριθμοι που χρησιμοποιήθηκαν για την εξαγωγή των χαρακτηριστικών από τα δεδομένα, οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν για τις προβλέψεις μας και οι θεωρητικές έννοιες που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας.

3.2 Σύνδεση χρηστών στα μέσα κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης έχουν διάφορες εκφάνσεις στον τρόπο λειτουργίας τους. Στην εργασία αυτή μας ενδιαφέρει ο τρόπος που συνδέονται οι χρήστες μεταξύ τους. Υπάρχουν δύο κύριοι τρόποι σύνδεσης:

- σύνδεση των χρηστών με ακολούθηση.
- σύνδεση των χρηστών με φιλία.

Στη σχέση ακολούθησης, που θα ασχοληθούμε σε αυτή την εργασία, υπάρχουν οι ορολογίες **follower** και **followee**. Σε μια σύνδεση μεταξύ δύο χρηστών, follower ορίζεται ο χρήστης που ακολουθεί τον άλλο, ενώ followee ορίζεται εκείνος που ακολουθείται. Ένας follower μπορεί ακολουθώντας έναν άλλο λογαριασμό να μένει ενήμερος για τη δραστηριότητα του followee. Ωστόσο, αυτή η σχέση δεν καθιστά υποχρεωτική και την αντίστροφη συνθήκη, δηλαδή ο followee να ακολουθεί τον follower. Αυτή η βασική αρχή είναι που καθιστά το δίκτυο χρηστών που δημιουργείται να είναι και κατευθυνόμενο.

Στη σύνδεση φιλίας, που δεν θα μας απασχολήσει σε αυτή την εργασία, ο κάθε χρήστης εκτελεί ένα αίτημα φιλίας. Η δυνατότητα δηλαδή ενός χρήστη, μέσω του αιτήματος, να ζητήσει την σύνδεσή του με έναν άλλο λογαριασμό. Σε αυτήν την περίπτωση, εφόσον ο έτερος χρήστης συμφωνήσει, η ένωση που δημιουργείται είναι αμφίδρομη. Έτσι, στο δίκτυο ενώσεων, δημιουργείται ένα γράφημα μη κατευθυνόμενο.

3.3 Είδη πρόβλεψης συνδέσμων

3.3.1 Πρόβλεψη χωρίς επίβλεψη

Σε μια πρόβλεψη χωρίς επίβλεψη δίνεται ένα στιγμιότυπο ενός ομοιογενούς δικτύου G = (V, E). Οι μέθοδοι πρόβλεψης χωρίς επίβλεψη στοχεύουν στο συμπέρασμα των πιθανών συνδέσμων που θα δημιουργηθούν στο μέλλον. Συνήθως, τα μοντέλα πρόβλεψης συνδέσμων χωρίς επίβλεψη υπολογίζουν ορισμένες βαθμολογίες για τους συνδέσμους, οι οποίες θα δείξουν την πιθανότητα μελλοντικής σύνδεσης δύο κόμβων.

Local Neighbor Αλγόριθμοι

Οι Local Neighbor [5] αλγόριθμοι βασίζονται στις τοπικές κοινωνικές πληροφορίες των χρηστών τοπικών κοινωνικών δικτύων, όπως κοινοί γείτονες δύο χρηστών. Μερικοί αλγόριθμοι με αυτή τη προσέγγιση είναι ο Preferential Attachment, ο Common Neighbor και ο Adam/Adar index.

- Ο πρώτος βασίζεται στη λογική πως αν και οι δύο χρήστες έχουν πολλούς ακολούθους, θα θέλουν να επεκτείνουν το δίκτυο τους και έτσι είναι πιο πιθανό να υπάρχει μελλοντική ένωση.
- Ο δεύτερος υπολογίζει τον αριθμό των κοινών ακολούθων μεταξύ δύο χρηστών.
- Ο τρίτος υπολογίζει τη βαθμολογία ομοιότητας μεταξύ δύο ιστοσελίδων και βασίζεται στα κοινόχρηστα χαρακτηριστικά.

Global Neighbor Αλγόριθμοι

Εκτός από τους αλγόριθμους που βασίζονται σε τοπικές πληροφορίες, πολλές άλλες μέθοδοι βασισμένες σε χαρακτηριστικά όλου του δικτύου έχουν επίσης προταθεί για μέτρηση της βαρύτητας μεταξύ των χρηστών. Μερικά παραδείγματα είναι:

- Ο αλγόριθμος Shortest Path (συντομότερο μονοπάτι), που υπολογίζει το γρηγορότερο μονοπάτι για μετάβαση από έναν αρχικό σε έναν τελικό κόμβο.
- Ο αλγόριθμος Katz, ένα μέτρο υπολογισμού της βαρύτητας ενός κόμβου σε ένα δίκτυο. Χρησιμοποιείται για να υπολογίζει τον σχετικό βαθμό επιρροής ενός κόμβου σε ένα κοινωνικό δίκτυο.
- Ο αλγόριθμος PageRank, σχεδιάστηκε με σκοπό τη βαθμολογία ιστοσελίδων με βάση τη βαρύτητα τους στο δίκτυο.

3.3.2 Πρόβλεψη με επίβλεψη

Στο πρόβλημα πρόβλεψης συνδέσμων με επίβλεψη το αρχικό πρόβλημα μετατρέπεται σε ένα πρόβλημα κατηγοριοποίησης. Αυτό γίνεται με την ανάθεση κλάσεων με 1 για τις δεδομένες ακμές του γραφήματος και 0 για μη διαθέσιμες ακμές. Το σύνολο ακμών που θα δημιουργηθεί πρέπει να είναι ισορροπημένο και με τις δύο κλάσεις. Ο σκοπός είναι από τις ακμές με κλάση 0 (δηλαδή σε συνδέσμους που λείπουν από το δίκτυο), να προβλέψουμε ποιες έχουν τη μεγαλύτερη πιθανότητα να μετατραπούν σε 1 στο υπάρχον γράφημα στο μέλλον.

Αυτό το πρόβλημα δυαδικής κατηγοριοποίησης, λύνεται συνήθως με αλγορίθμους όπως δέντρα αποφάσεων, SVM(Support Vector Machines), νευρωνικά δίκτυα και άλλα που θα δούμε στη συνέχεια της εργασίας.

Μείωση διαστάσεων

Στη μηχανική μάθηση, το πρόβλημα των πολλαπλών διαστάσεων είναι πολύ συχνό. Για τη λύση αυτού του προβλήματος έχουν υλοποιηθεί αλγόριθμοι που προσπαθούν να μειώσουν τις πολλαπλές διαστάσεις για την ευκολότερη ανάλυση των δεδομένων και την απλοποίηση του προβλήματος. Η μείωση των διαστάσεων ωστόσο, σε προβλήματα πρόβλεψης ακμών δεν είναι ιδιαίτερα ανεπτυγμένη. Μια τακτική που χρησιμοποιείται είναι η πρόβλεψη των ακμών να μετατρέπεται σε πρόβλημα επιβλεπόμενης μάθησης και η μείωση των διαστάσεων να γίνεται με απλούς αλγορίθμους αυτής της χρήσης [6].

Ένας από τους πιο διαδεδομένους αλγορίθμους είναι ο PCA(Principal Component Analysis) [7]. Ο PCA μπορεί να χωριστεί σε πέντε στάδια:

- Ομαλοποίηση του εύρους των τιμών των μεταβλητών έτσι ώστε η κάθε μεταβλητή να συμβάλει εξίσου στην ανάλυση.

- Υπολογισμός του πίνακα συνδιακύμανσης (covariance matrix).
- Υπολογισμός των ιδιοδιανυσμάτων και των ιδιοτιμών του πίνακα συνδιακύμανσης. Το βήμα αυτό χρησιμοποιείται για τον υπολογισμό των κυρίων συνιστωσών (Principal Components) των δεδομένων.
- Επιλογή των διανυσμάτων με τις ιδιοτιμές που μας δίνουν τις καλύτερες δυνατές κύριες συνιστώσες για τα δεδομένα.
- Μετατροπή των αρχικών ομαλοποιημένων δεδομένων στο τελικό σύνολο δεδομένων με τη χρήση των διανυσμάτων ιδιοτιμών που επιλέχθηκαν στο προηγούμενο βήμα.

3.3.3 Πρόβλεψη συνδέσμων σε χρονικά δίκτυα

Ένα ακόμα είδος δικτύου που σχετίζεται με τα μέσα κοινωνικής δικτύωσης είναι τα χρονικά δίκτυα. Σε πολλά μέσα όπως Facebook ή Twitter προστίθενται νέοι σύνδεσμοι σε πολύ συχνό χρονικό περιθώριο. Σε αυτά τα δίκτυα είναι χρήσιμο λοιπόν, να είναι και ο χρόνος παράγοντας στις προβλέψεις των συνδέσμων.

$$\mathrm{A}((i,j,T) = \begin{cases} 1, \text{αν υπάρχει σύνδεσμος από το i στο j σε χρόνο T} \\ 0 \end{cases}$$

Ο αριθμός των δημοσιεύσεων σε αυτόν το τομέα δεν είναι τόσο μεγάλος σχετικά με άλλους τομείς, ωστόσο έχουν υπάρξει προσπάθειες πάνω στο αντικείμενο. Σκοπός της πρόβλεψης είναι σε ένα χρόνο t, όπου το t παίρνει τιμές από 1 ως T, να προβλέψουμε τους συνδέσμους που θα προκύψουν σε χρόνο T+1.

Ο Purnamrita Sarkar στο [8] εισήγαγε μια μη παραμετρική μέθοδο για την πρόβλεψη συνδέσμων σε χρονικά δίκτυα όπου ο χρόνος χρησιμοποιείται σε υποακολουθίες στιγμιότυπων του γραφήματος. Αυτή η προσέγγιση προβλέπει συνδέσμους με βάση τοπολογικά χαρακτηριστικά, όπως η διαδρομή που χρειάζεται να διασχίσεις μέσα σε ένα δίκτυο για να πας από ένα κόμβο σε ένα άλλο, και κοινούς γείτονες μεταξύ των κόμβων του γραφήματος.

3.4 Κατηγοριοποίηση

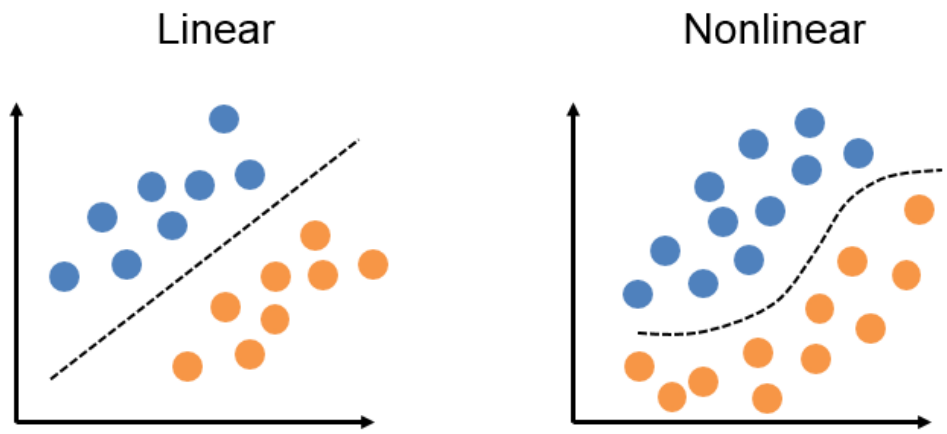
Μια συχνή χρήση των αλγορίθμων μηχανικής μάθησης είναι η κατηγοριοποίηση. Αυτή η διαδικασία βοηθάει στο διαχωρισμό μεγάλων ποσοτήτων δεδομένων σε διακριτές τιμές. Η

κατηγοριοποίηση μπορεί να γίνει σε πολλές μορφές.

Στη δυαδική κατηγοριοποίηση διαχωρίζονται δύο κλάσεις μεταξύ τους. Η διαδικασία αυτή είναι χρήσιμη σε εφαρμογές όπως εντοπισμός email τύπου spam, ιατρικές διαγνώσεις κ.α. Στη κατηγοριοποίηση σε πολλές κλάσεις, υπάρχουν εξίσου πολλές εφαρμογές όπως η αναγνώριση εικόνων.

Η κατηγοριοποίηση πολλαπλών κλάσεων χρησιμοποιείται για τον διαχωρισμό μεγαλύτερου αριθμού κλάσεων. Κάθε δείγμα μπορεί να χαρακτηριστεί μόνο ως μία κατηγορία. Για παράδειγμα, στη κατηγοριοποίηση εικόνων από οχήματα, υπάρχουν πολλοί τύποι οχημάτων όπως μηχανή, φορτηγό κ.α.

Υπάρχουν αρκετοί αλγόριθμοι που μπορούν να συνεισφέρουν στη κατηγοριοποίηση των δεδομένων. Αλγόριθμοι που διαχωρίζουν τα δεδομένα με γραμμική μορφή είναι ο Logistic Regression ή και ο SVM(Support Vector Machines). Αντίθετα, αλγόριθμοι με μη γραμμικό διαχωρισμό είναι οι αλγόριθμοι δένδρων αποφάσεων, νευρωνικών δικτύων, naive bayes κ.α. Ακόμα υπάρχουν και οι αλγόριθμοι ενίσχυσης, που συνδυάζουν πολλά μοντέλα προσπαθώντας να χρησιμοποιήσουν τα λάθη τον προηγούμενων αλγορίθμων για να εκπαιδευτεί καλύτερα το τελικό μοντέλο.



Σχήμα 3.1: Παράδειγμα κατηγοριοποίησης

Η πηγή του σχήματος βρίσκεται στη δημοσίευση [9]

3.5 Μέθοδοι γραφημάτων

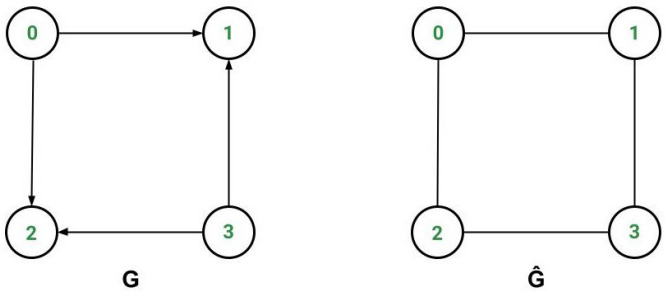
3.5.1 Συντομότερο Μονοπάτι

Στη θεωρία γραφημάτων, το πρόβλημα της συντομότερης διαδρομής [10] είναι το πρόβλημα της εύρεσης μιας διαδρομής μεταξύ δύο κορυφών (ή κόμβων) σε ένα γράφημα έτσι ώστε το άθροισμα των βαρών των ακμών που το αποτελούν να ελαχιστοποιείται.

Το κοινωνικό δίκτυο μοντελοποιείται ως γράφημα και το αντίστοιχο πρόβλημα συντομότερης διαδρομής είναι ο μικρότερος αριθμός χρηστών του δικτύου που χρειάζεται, έτσι ώστε να φθάσουμε από τον ένα χρήστη στον άλλο.

3.5.2 Ασθενώς Συνδεδεμένο Στοιχείο

Ένα Ασθενώς συνδεδεμένο στοιχείο (Weakly Connected Components) [11] ενός απλού κατευθυνόμενου γραφήματος, είναι το μέγιστο υπογράφημα έτσι ώστε για κάθε ζεύγος διακριτών κορυφών u, v στον υπογράφο, να υπάρχει μια μη κατευθυνόμενη διαδρομή από το u στο v.



Σχήμα 3.2: WCC

Πηγή: [12]

3.5.3 Adamic/Adar δείκτης

Ο δείκτης Adamic–Adar [10] είναι ένα μέτρο που εισήχθη το 2003 από τους Lada Adamic και Eytan Adar για την πρόβλεψη συνδέσμων σε ένα κοινωνικό δίκτυο, σύμφωνα με τον αριθμό των κοινόχρηστων συνδέσεων μεταξύ δύο κόμβων. Ορίζεται ως το άθροισμα της κεντρικότητας του αντίστροφου λογαριθμικού βαθμού των γειτόνων που μοιράζονται οι δύο κόμβοι

$A(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$

3.5.4 Preferential Attachment

Η μέθοδος Preferential Attachment [10], βασίζεται στην λογική πως στα μέσα κοινωνικής δικτύωσης οι λογαριασμοί με πολλούς ακολούθους τείνουν να αποκτήσουν περισσότερους σε μελλοντική χρήση της πλατφόρμας. Ο αλγόριθμος αυτός υπολογίζει πόσο έντονη είναι η δραστηριότητα δύο λογαριασμών πολλαπλασιάζοντας των αριθμό των ακολούθων των δύο χρηστών.

$$preferential - attachment - score(u, v) = |\Gamma(u)||\Gamma(v)|$$

3.5.5 Βάρος Ακμής

Η μέθοδος υπολογισμού του Βάρους Ακμής (Edge Weight) [4] έχει ως σκοπό της να θέσει βάρη σε κάθε κόμβο για τους followers και τους followees της κάθε ακμής. Το βάρος ενός κόμβου μειώνεται καθώς αυξάνεται ο αριθμός των ακολούθων του αντίστοιχου χρήστη. Αυτό συμβαίνει με βάση την παρατήρηση ότι σε ένα μέσο κοινωνικής δικτύωσης με πολλά δημόσια πρόσωπα, βλέπουμε να υπάρχουν πολλοί ακόλουθοι. Το πιθανότερο είναι ότι οι περισσότεροι από τους ακολούθους δεν γνωρίζονται ούτε μεταξύ τους αλλά ούτε και με τη διασημότητα. Ο τύπος υπολογισμού του βάρους είναι ο παρακάτω:

$$W = \frac{1}{\sqrt{1+|X|}}$$

Σε κάθε ακμή ενός μέσου κοινωνικής δικτύωσης έχουμε τον follower και τον followee. Με αυτόν το τρόπο, υπολογίζουμε το βάρος για τους χρήστες που ακολουθεί ο follower και το βάρος για τους χρήστες που ακολουθούν τον followee. Ονομάζουμε το πρώτο εξερχόμενο βάρος (outgoing weight), ενώ το δεύτερο εισερχόμενο βάρος (incoming weight). Με αυτό τον τρόπο, μπορούμε να υπολογίσουμε τις 4 εκδοχές του βάρους ακμής που είναι οι παρακάτω:

εισερχόμενο βάρος + εξερχόμενο βάρος
εισερχόμενο βάρος * εξερχόμενο βάρος
2*εισερχόμενο βάρος + εξερχόμενο βάρος
εισερχόμενο βάρος + 2*εξερχόμενο βάρος

3.6 Δείκτες σημαντικότητας ιστοσελίδων

3.6.1 Εισαγωγή

Οι αλγόριθμοι που έχουν υλοποιηθεί για τον υπολογισμό της αξίας μιας ιστοσελίδας μπορεί να φανούν αρκετά χρήσιμοι σε αυτή την εργασία. Οι αλγόριθμοι βασίζονται στις διασυνδέσεις που έχουν οι σελίδες του διαδικτύου μεταξύ τους. Για παράδειγμα, μια σελίδα μπορεί να χρησιμοποιεί ως πηγή της μια άλλη. Το σύνολο των ιστοσελίδων αυτών σχηματίζουν μεταξύ τους ένα δίκτυο, παρόμοιο με αυτό των μέσων κοινωνικής δικτύωσης.

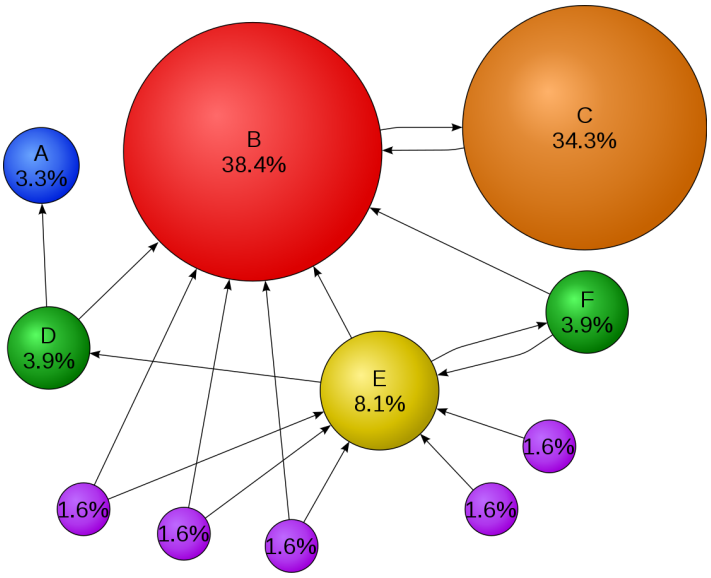
3.6.2 Hits algorithm (Hubs και Authorities)

Ο Hyperlink-Induced Topic Search [13] (HITS) είναι ένας αλγόριθμος ανάλυσης συνδέσμων που βαθμολογεί ιστοσελίδες. Η ιδέα πίσω από τα Hubs και Authorities προέκυψε καθώς ιστοσελίδες αναπαρήγαγαν πληροφορία που είτε δεν ήταν δική τους ή την παρερμήνευαν και οδηγούσαν τους χρήστες να κατευθύνονται σε άλλες έγκυρες σελίδες. Έτσι με τον αλγόριθμο hits, ένα καλό hub αντιπροσωπεύει μια σελίδα που δείχνει σε πολλές άλλες σελίδες, ενώ ένα καλό authority είναι μια σελίδα που συνδέεται με πολλούς διαφορετικούς κόμβους.

Αντίστοιχα, στα μέσα κοινωνικής δικτύωσης ένας καλός hub είναι ένας χρήστης που ακολουθεί πολλούς λογαριασμούς σχετικά με τον μέσο όρο. Ένα καλό Authority είναι ο χρήστης που τον ακολουθούν πολλοί λογαριασμοί. Με αυτό τον τρόπο μπορούμε να χρησιμοποιήσουμε αυτές τις πληροφορίες για τις προβλέψεις συνδέσμων.

3.6.3 Page Rank

Ο PageRank [13] είναι ένας αλγόριθμος που δημιουργήθηκε από την Google για την κατάταξη ιστοσελίδων στα αποτελέσματα των μηχανών αναζήτησής τους. Ο PageRank είναι ένας τρόπος μέτρησης της σημαντικότητας των ιστοσελίδων. Η μέτρηση αυτή βασίζεται στον αριθμό και την ποιότητα των συνδέσμων μιας σελίδας για να καθορίσει μια κατά προσέγγιση εκτίμηση του πόσο σημαντικός είναι ο ιστότοπος.



Σχήμα 3.3: Page Rank Παράδειγμα

Πηγή: [14]

3.7 Μετρικές Αξιολόγησης και δείκτες ομοιότητας

Το πρόβλημα πρόβλεψης συνδέσμων αντιμετωπίζεται ως πρόβλημα δυαδικής κατηγοριοποίησης, για αυτόν τον λόγο θα αξιοποιηθούν στην παρούσα εργασία οι ακόλουθες μετρικές [15].

- True Positive (TP): Το θετικό στοιχείο δεδομένων(ύπαρξη ακμής), προβλέφθηκε ως θετικό.
- True Negative (TN): Το αρνητικό στοιχείο δεδομένων (μη ύπαρξη ακμής) προβλέφθηκε ως αρνητικό.
- False Positive (FP): Το αρνητικό στοιχείο δεδομένων (μη ύπαρξη ακμής) προβλέφθηκε ως θετικό.
- False Negative (FN): Το θετικό στοιχείο δεδομένων (ύπαρξη ακμής) προβλέφθηκε ως αρνητικό .

Με τη χρήση αυτών των τιμών μπορούν να υπολογιστούν μετρικές όπως Accuracy, Precision, Recall, F1-score και ROCAUC.

Μετρικές Αξιολόγησης		
Μετρική	Τύπος	Ορισμός
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	ποσοστό των σωστών προβλέψεων προς το σύνολο των δειγμάτων
Precision	$\frac{TP}{TP+FP}$	ποσοστό των δειγμάτων που προβλέφθηκαν στη θετική τάξη και ανήκουν στην θετική τάξη
Recall	$\frac{TP}{TP+FN}$	ποσοστό των σωστών θετικών προβλέψεων προς όλα τα θετικά παραδείγματα των δεδομένων
F1-score	$\frac{2}{\mathrm{Recall}^{-1}+\mathrm{Precision}^{-1}}$	ενιαία επίδοση που εξισορροπεί την επίδοση του Precision όσο και του Recall

Πίνακας 3.1: Επιδόσεις όλων των μοντέλων με 7 χαρακτηριστικά

Η μετρική AUROC (Area under the receiver operating characteristics curve) υπολογίζει την απόδοση σε προβλήματα κατηγοριοποίησης. ROC ονομάζεται η καμπύλη πιθανότητας και AUC το μέτρο του διαχωρισμού μεταξύ των κλάσεων. Η μετρική δείχνει κατά πόσο το μοντέλο μπορεί να διαχωρίσει τις κλάσεις μεταξύ τους.

Jaccard similarity

Ο δείκτης ομοιότητας Jaccard [16] συγκρίνει την ομοιότητα δύο σύνολων, ενώ η τιμή της κυμαίνεται από μηδέν ως ένα. Ορίζεται ως το μέτρο της τομής των δύο συνόλων, διαιρούμενη με το μέτρο της ένωσης των δύο συνόλων.

$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$

Ομοιότητα συνημιτόνου/Otsuka–Ochiai coefficient

Ο δείκτης ομοιότητας συνημιτόνου [10] συγκρίνει την ομοιότητα δύο σύνολων, ενώ η τιμή της κυμαίνεται από μηδέν ως ένα. Στη παρούσα εργασία θα χρησιμοποιηθεί ο τελεστής Otsuka–Ochiai coefficient, που είναι μια παρόμοια εκδοχή του τελεστή συνημιτόνου. Αυτή ορίζεται ως το μέτρο της τομή των δύο συνόλων, διαιρούμενο με τη ρίζα του γινομένου των μέτρων των δύο συνόλων.

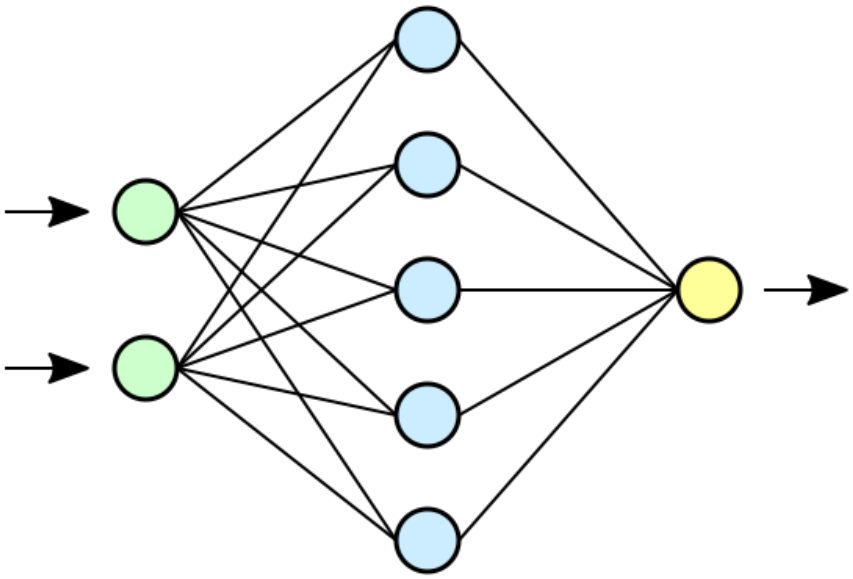
$$K = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$$

3.8 Μοντέλα μηχανικής μάθησης

Για την εκπαίδευση του συστήματος συστάσεων χρησιμοποιήθηκαν διάφοροι αλγόριθμοι μηχανικής μάθησης. Στη συνέχεια θα αναλύσουμε το θεωρητικό υπόβαθρο του κάθε αλγορίθμου.

3.8.1 Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα [17], στον τομέα της Τεχνητής Νοημοσύνης έχουν σχεδιαστεί με βάση το δίκτυο των νευρώνων ενός ανθρώπινου εγκεφάλου. Τα νευρωνικά δίκτυα σχεδιάζονται ως μαθηματικά μοντέλα με σκοπό τους την αναγνώριση προτύπων. Τα νευρωνικά δίκτυα δομούνται από τρία βασικά είδη νευρώνων. Αυτά περιέχουν ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Κάθε κόμβος, ή τεχνητός νευρώνας, συνδέεται με έναν άλλο και έχει ένα σχετικό βάρος, μια σταθερά (bias) που προστίθεται στο τέλος και ένα κατώφλι ενεργοποίησης. Κάθε είσοδος πολλαπλασιάζεται με το βάρος του κάθε νευρώνα και προστίθεται με τη σταθερά, ενώ μετά περνάει από μια συνάρτηση ενεργοποίησης για να υπολογιστεί η τελική του έξοδος.



Σχήμα 3.4: Νευρωνικό δίκτυο

Πηγή:

`https://commons.wikimedia.org/wiki/File:Neural_network.svg`

Για την καλύτερη απόδοση των νευρωνικών δικτύων σκοπός είναι η μείωση της συνάρτησης κόστους στο ελάχιστο δυνατό. Καθώς το μοντέλο προσαρμόζει τα βάρη και το bias

του, χρησιμοποιεί τη συνάρτηση κόστους και την ενίσχυση της εκμάθησης για να συγκλίνει η τιμή κόστους στο ελάχιστο. Η διαδικασία κατά την οποία ο αλγόριθμος προσαρμόζει τα βάρη του είναι μέσω της συνάρτησης gradient descent, οδηγώντας το μοντέλο στη κατεύθυνση που πρέπει να ακολουθήσει για να μειώσει τα σφάλματα του. Με κάθε επανάληψη της εκπαίδευσης με ένα καινούργιο παράδειγμα, οι παράμετροι του μοντέλου προσαρμόζονται ώστε να συγκλίνουν σταδιακά στη καλύτερη δυνατή τιμή.

Συνάρτηση Ενεργοποίησης

Η επιλογή της συνάρτησης ενεργοποίησης έχει μεγάλο αντίκτυπο στην απόδοση του νευρωνικού νευρωνικού δικτύου, ενώ διαφορετικές συναρτήσεις μπορούν να χρησιμοποιηθούν σε διαφορετικά layers του μοντέλου. Στα hidden layers τρεις είναι οι κυριότερες συναρτήσεις ενεργοποίησης:

- Logistic (Sigmoid)
- Hyperbolic Tangent (Tanh)
- Rectified Linear Activation (ReLU)

Η **Sigmoid** ονομάζεται επίσης και Logistic επειδή είναι η ίδια συνάρτηση που χρησιμοποιεί ο αλγόριθμος Logistic Regression. Η συνάρτηση έχει ως είσοδο πραγματικές τιμές και ως έξοδο στην περιοχή από 0 έως 1. Υπολογίζεται με τον τύπο:

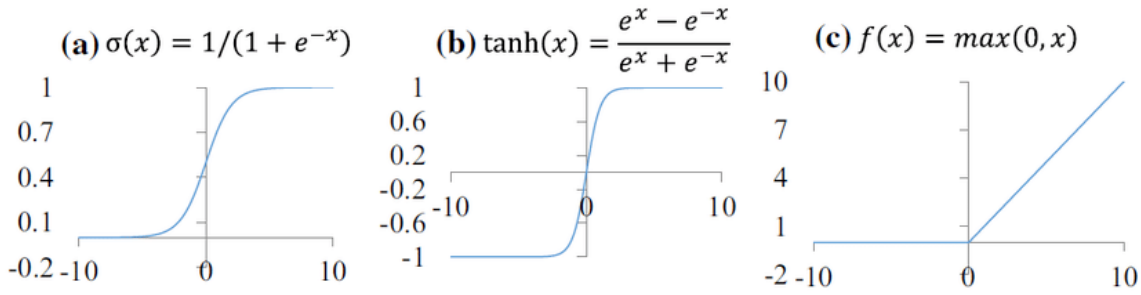
$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x).$$

Η **Tanh**, είναι παρόμοια της συνάρτησης Sigmoid καθώς και αυτή έχει έξοδο παρόμοια με το αγγλικό γράμμα S. Η συνάρτηση έχει ως είσοδο πραγματική τιμή και τιμές εξόδου στην περιοχή -1 έως 1. Η συνάρτηση υπολογίζεται με τον τύπο :

$$\tanh(x) \doteq \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Η **ReLu** συνηθίζεται να χρησιμοποιείται διότι είναι αποτελεσματική ενώ ταυτόχρονα πολύ απλή στον υπολογισμό της. Ακόμα, ξεπερνάει αρκετούς από τους περιορισμούς άλλων συναρτήσεων ενεργοποίησης. Η συνάρτηση υπολογίζεται με τον τύπο :

$$\max(0, x)$$



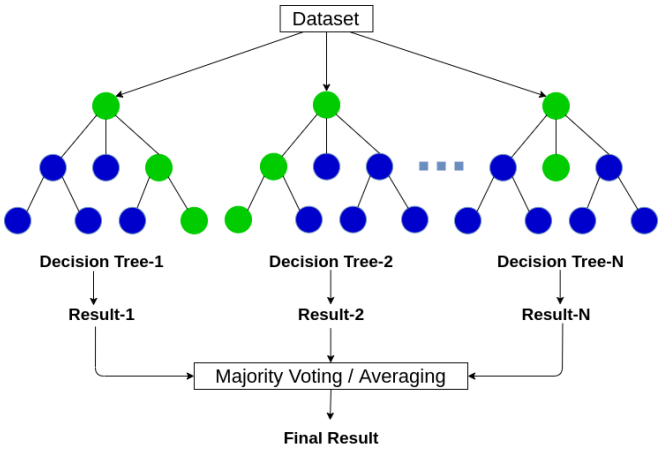
Σχήμα 3.5: Συναρτήσεις ενεργοποίησης

Πηγή: [9]

3.8.2 Random Forest

Ο αλγόριθμος Random Forest [18] αποτελεί ένα υποσύνολο των δέντρων αποφάσεων. Τα δέντρα αποφάσεων είναι αλγόριθμοι χωρίς παραμέτρους, και χρησιμοποιούνται για κατηγοριοποίηση, οπισθοδρόμηση (regression) ή και συσταδοποίηση. Έχει μια ιεραρχική, δενδρική δομή, η οποία αποτελείται από έναν κόμβο ρίζα , εσωτερικούς κόμβους που δέχονται μια εισερχόμενη ακμή, δύο ή παραπάνω εξερχόμενες καθώς και τερματικούς κόμβους στους οποίους βρίσκεται η τελική πρόβλεψη του αλγορίθμου.

Ο αλγόριθμος Random Forest χρησιμοποιεί ένα μεγάλο αριθμό μεμονωμένων, μη κλαδευμένων δέντρων αποφάσεων που δημιουργούνται με ένα τυχαίο διαχωρισμό σε κάθε κόμβο του δέντρου αποφάσεων. Κάθε δέντρο είναι πιθανό να είναι λιγότερο ακριβές παρά ένα δέντρο που δημιουργήθηκε με τις ακριβείς διασπάσεις. Όμως, συνδυάζοντας πολλά από αυτά τα «κατά προσέγγιση» δέντρα σε ένα σύνολο, βελτιώνεται η ακρίβεια και ο Random Forest συχνά τα καταφέρνει καλύτερα από ένα μόνο δέντρο με ακριβείς διασπάσεις.



Σχήμα 3.6: Random Forest Παράδειγμα

Πηγή: `https://bit.ly/3U0WUvY`

3.8.3 XGBoost

Το XGBoost [19] είναι ένας αλγόριθμος μηχανικής μάθησης που βασίζεται σε δέντρα αποφάσεων που χρησιμοποιεί ένα framework gradient boosting.

Η μέθοδος Boosting είναι μια τεχνική μοντελοποίησης που επιχειρεί να δημιουργήσει έναν ισχυρό κατηγοριοποιητή συνδυάζοντας μαζί πολλούς αδύναμους σε σειρά.

- Πρώτον, δημιουργείται ένα μοντέλο από τα δεδομένα εκπαίδευσης.
- Μετά κατασκευάζεται το δεύτερο μοντέλο το οποίο προσπαθεί να διορθώσει τα σφάλματα που υπάρχουν στο πρώτο μοντέλο.
- Η διαδικασία συνεχίζεται και προστίθενται μοντέλα μέχρι η απόδοση της πρόβλεψης να είναι απόλυτα ακριβής ή να συμπληρωθεί ο μέγιστος αριθμός μοντέλων.

Το XGBoost είναι μια υλοποίηση δέντρων αποφάσεων με Gradient Boosting. Σε αυτόν τον αλγόριθμο, τα δέντρα αποφάσεων δημιουργούνται σε διαδοχική μορφή. Τα βάρη παίζουν σημαντικό ρόλο στο XGBoost. Τα βάρη εκχωρούνται σε όλες τις ανεξάρτητες μεταβλητές οι οποίες στη συνέχεια τροφοδοτούνται στο δέντρο αποφάσεων που προβλέπει τα αποτελέσματα. Το βάρος των μεταβλητών που προβλέπονται λανθασμένα από το δέντρο αυξάνεται και αυτές οι μεταβλητές τροφοδοτούνται στη συνέχεια στο δεύτερο δέντρο απόφασης. Αυτοί οι μεμονωμένοι ταξινομητές στη συνέχεια συνδυάζονται για να δώσουν ένα ισχυρό και πιο ακριβές μοντέλο. Μπορεί να λειτουργήσει σε προβλήματα οπισθοδρόμησης (regression), κατηγοριοποίησης και συσταδοποίησης.

3.8.4 LightGBM

Το framework LightGBM [20] υποστηρίζει πολλούς διαφορετικούς αλγόριθμους, συμπεριλαμβανομένων των Gradient Boosted δένδρων αποφάσεων και άλλων πολλών. Το LightGBM έχει πολλά από τα πλεονεκτήματα του XGBoost αλγόριθμου, όπως της παράλληλης εκπαίδευσης, πολλαπλών συναρτήσεων απώλειας. Μια σημαντική διαφορά μεταξύ των δύο έγκειται στην κατασκευή των δέντρων. Επιπλέον, το LightGBM δεν χρησιμοποιεί τον ευρέως χρησιμοποιούμενο αλγόριθμο εκμάθησης δέντρων αποφάσεων βάσει κατηγοριοποίησης, ο οποίος αναζητά το καλύτερο σημείο διαχωρισμού σε ταξινομημένες τιμές χαρακτηριστικών, όπως ο XGBoost. Αντίθετα, το LightGBM εφαρμόζει έναν εξαιρετικά βελτιστοποιημένο αλγόριθμο εκμάθησης δένδρων αποφάσεων που βασίζεται σε ιστόγραμμα, ο οποίος αποφέρει

μεγάλα πλεονεκτήματα τόσο στην απόδοση όσο και στην κατανάλωση μνήμης. Ο αλγόριθμος LightGBM χρησιμοποιεί δύο νέες τεχνικές που ονομάζονται Gradient-Based One-Side Sampling (GOSS) και Exclusive Feature Bundling (EFB) που επιτρέπουν στον αλγόριθμο να εκτελείται πιο γρήγορα διατηρώντας παράλληλα υψηλό επίπεδο ακρίβειας.

Κεφάλαιο 4

Δεδομένα και εργαλεία

4.1 Python

Όλες οι μέθοδοι εκπαίδευσης, η διαχείριση του δικτύου και οι συστάσεις υλοποιήθηκαν με τη χρήση της γλώσσας προγραμματισμού Python. Η Python είναι από τις πιο εύχρηστες γλώσσες στην επιστήμη δεδομένων ενώ παρέχει στο προγραμματιστή πολλαπλές δυνατότητες μέσω των βιβλιοθηκών της. Η γλώσσα χρησιμοποιείται για υψηλού επιπέδου προγραμματισμό, είναι δυναμική γλώσσα και χρησιμοποιεί διερμηνευτή για την εκτέλεση της.

Πιο αναλυτικά, για την υλοποίηση της εργασίας και συγκεκριμένα για την διαχείριση των δεδομένων και την εκπαίδευση των μοντέλων σύστασης, χρησιμοποιήθηκαν οι βιβλιοθήκες pandas, tensorflow, sklearn . Ακόμα, για τη διαχείριση του δικτύου και την εξαγωγή των χαρακτηριστικών για την εκπαίδευση των μοντέλων χρησιμοποιήθηκε η βιβλιοθήκη networkx. Τέλος, για την γραφική αναπαράσταση των αποτελεσμάτων χρησιμοποιήθηκε η βιβλιοθήκη matplotlib.

4.2 Jupyter Notebook

Το Jupyter Notebook είναι ένα διαδραστικό υπολογιστικό περιβάλλον με χρήση στο διαδίκτυο. Η υλοποίηση του βασίζεται σε βιβλιοθήκες ανοιχτού κώδικα. Η χρήση του είναι κυρίως στην διαχείριση δεδομένων σε συνδυασμό με το διαχωρισμό του κώδικα σε διαφορετικά κελιά για πιο εύκολη χρήση.

Το Jupyter Notebook υποστηρίζει πολλές γλώσσες προγραμματισμού. Οι κυριότερες από αυτές είναι είναι η Python, η R και η Julia. Το Project Jupyter έχει σχεδιαστεί έτσι ώστε να

υπολογίζει και άλλες υπηρεσίες όπως το Jupyter Lab αλλά και το Jupyter Hub.

Ένα Jupyter Notebook αρχείο είναι στην βάση του ένα έγγραφο JSON. Αυτό μπορεί να περιλαμβάνει δύο τύπου κελιά. Το πρώτο είναι κελί εισόδου και εξόδου κώδικα, ενώ το δεύτερο είναι κελί κειμένου με τη χρήση των Markdowns. Τα Markdowns, εκτός από κείμενο, μπορούν να περιλαμβάνουν μαθηματικά σύμβολα, γραφικές παραστάσεις και αρχεία φωτογραφιών.

4.3 Δεδομένα

Το Facebook έχει παραχωρήσει ένα σύνολο δεδομένων τα οποία είναι διαθέσιμα από την πλατφόρμα του Kaggle [2]. Τα δεδομένα περιλαμβάνουν δύο στήλες, μια πηγή και έναν προορισμό, για κάθε ακμή του μέσου κοινωνικής δικτύωσης. Αυτές συμβολίζουν τον follower (source_node) και τον followee (destination_node) της κάθε σύνδεσης των χρηστών του γραφήματος που παρέχει Facebook.

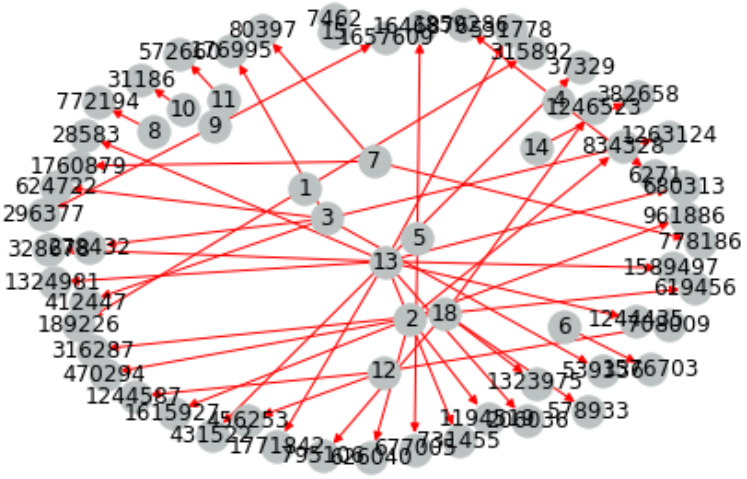
	source_node	destination_node
0	1	690569
1	1	315892
2	1	189226
3	2	834328
4	2	1615927
...	...	...
9437514	1862219	1187308
9437515	1862219	563943
9437516	1862219	1044046
9437517	1862219	1022613
9437518	1862220	1748794

Σχήμα 4.1: Αρχικά δεδομένα

Ο αριθμός των κόμβων του δικτύου είναι 1.862.220 ενώ των ακμών 9.437.519. Αυτό σημαίνει πως η πλατφόρμα περιλαμβάνει 1.862.220 ξεχωριστούς χρήστες με 9.437.519 ακολουθήσεις συνολικά.

4.3.1 Οπτικοποίηση των Δεδομένων

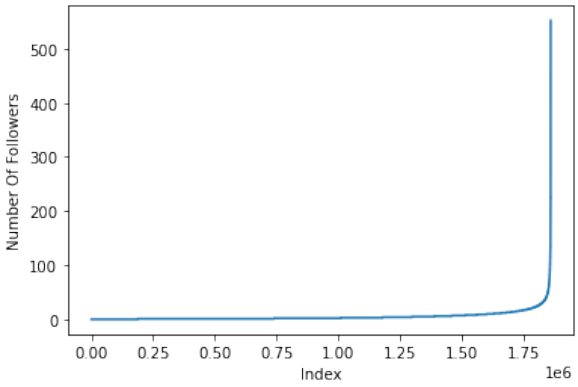
Μια απεικόνιση του δικτύου που μας παρέχουν τα δεδομένα της πλατφόρμας του Facebook είναι η παρακάτω. Όπως φαίνεται από τα βέλη του γραφήματος το δίκτυο είναι κατευθυνόμενο. Τα ονόματα των χρηστών έχουν αντικατασταθεί με μοναδικούς αριθμούς για να διατηρηθεί η προστασία των προσωπικών τους δεδομένων.



Σχήμα 4.2: Οπτικοποίηση δείγματος γραφήματος

4.3.2 Ανάλυση των Δεδομένων

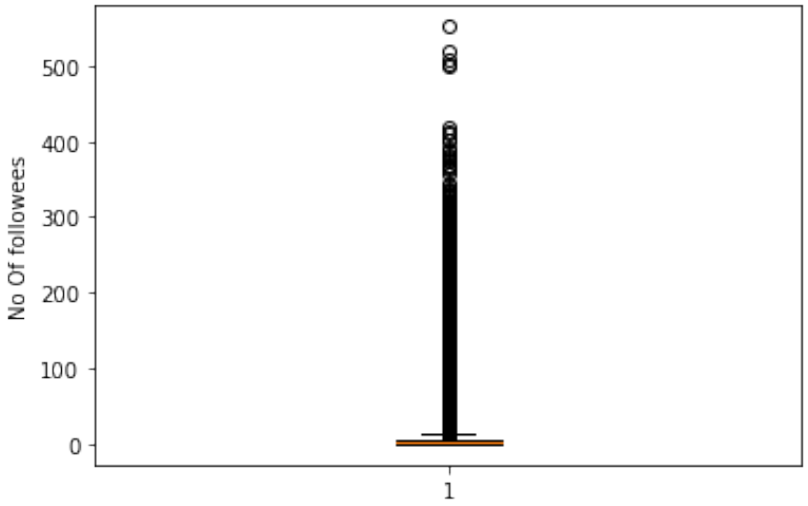
Οι followers κάθε χρήστη της πλατφόρμας είναι κατά μέσο όρο 5. Αυτό συμβαίνει καθώς σε τέτοιες πλατφόρμες δημιουργούνται πολλοί λογαριασμοί που μετά από κάποιο χρόνο δε χρησιμοποιούνται. Για αυτό το λόγο είναι σημαντικό στο στάδιο της δειγματοληψίας να γίνει επιλογή των χρηστών με μεγάλο συγκριτικά αριθμό followers.



Σχήμα 4.3: Followers κάθε ανθρώπου

Στο παραπάνω γράφημα φαίνεται ο αριθμός των ακολούθων ανά χρήστη της πλατφόρμας. Από το γράφημα αυτό μπορούμε να καταλάβουμε πώς αυξάνεται το πλήθος των ακολούθων στο σύνολο των χρηστών.

Στη συνέχεια βλέπουμε ένα barplot, με τον αριθμό των followees του κάθε λογαριασμού της πλατφόρμας.



Σχήμα 4.4: Followees κάθε ανθρώπου

Η αναλογία του αριθμού των followers και των followees είναι παρόμοια, όπως θα περίμενε κανείς.

Οι λογαριασμοί που έχουν παραπάνω από 50 followers και 50 followees είναι κοντά στους 10.000 που είναι αρκετοί έτσι ώστε να μπορέσουμε να εκπαιδεύσουμε τα μοντέλα μας. Ακόμα είναι χρήσιμη και η χρήση λογαριασμών με μικρότερο αριθμό ακολούθων καθώς υπάρχουν αρκετοί χρήστες που δεν επιλέγουν να ακολουθούν πολλούς λογαριασμούς.

Με τη συνάρτηση weakly_connected_components της βιβλιοθήκης networkxx, βλέπουμε πως υπάρχουν 45.558 ασθενώς συνδεδεμένες συνιστώσες (weakly connected components) κόμβων στο γράφημα με 32.195 από αυτές να είναι μεταξύ δύο κόμβων.

Τα ασθενώς συνδεδεμένα στοιχεία μας δείχνουν πως υπάρχουν υπογραφήματα για κάθε ζεύγος διακριτών κορυφών u, v στον υπογράφο και υπάρχει μια **μη κατευθυνόμενη** διαδρομή από το u στο v στα οποία όλοι οι χρήστες ενώνονται με όλους. Στα μέσα κοινωνικής δικτύωσης αυτό μπορεί να μεταφραστεί λέγοντας πως υπάρχουν χρήστες που ακολουθούν ο ένας τον άλλο, σχηματίζοντας έτσι μια κλειστή ομάδα.

Όμοια, με τη συνάρτηση strongly_connected_components της βιβλιοθήκης networkxx, βλέπουμε πως υπάρχουν 527.748 ισχυρά συνδεδεμένες συνιστώσες (strongly connected components) κόμβων στο γράφημα με 33.105 από αυτές να είναι μεταξύ δύο κόμβων.

Η διαφορά των ισχυρά συνδεδεμένων με των ασθενώς συνδεδεμένων ενώσεων, είναι πως υπάρχουν υπογραφήματα για κάθε ζεύγος διακριτών κορυφών u, v στον υπογράφο και υπάρχει μια **κατευθυνόμενη** διαδρομή από το u στο v στα οποία όλοι οι χρήστες ενώνονται με όλους. Οπότε, είναι λογικό οι ισχυρές συνιστώσες να είναι περισσότερες των ασθενών, καθώς οι διαδρομές περιορίζονται και οι συνιστώσες γίνονται μικρότερες, άρα και περισσότερες. Στα μέσα κοινωνικής δικτύωσης, αυτό συνεπάγεται σε περισσότερες και ολιγάριθμες ομάδες χρηστών.

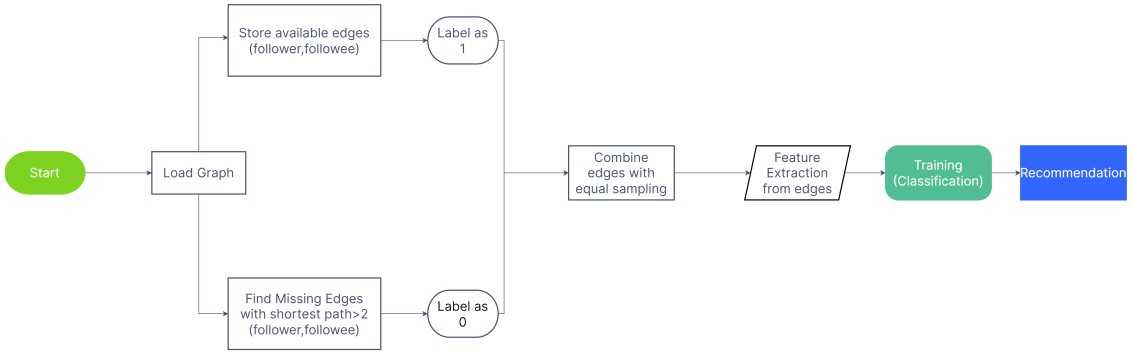
Κεφάλαιο 5

Σχεδιασμός και υλοποίηση μοντέλων σύστασης

5.1 Εισαγωγή

Σε αυτή την εργασία γίνεται σύσταση μέσω επιβλεπόμενης μάθησης. Σκοπός του συστήματος είναι η μετατροπή του προβλήματος από πρόβλεψη συνδέσμων, σε πρόβλημα **κατηγοριοποίησης**. Σε παρόμοια προβλήματα πρόβλεψης συνδέσμων, εφαρμόζονται και αλγόριθμοι μάθησης χωρίς επίβλεψη. Τέτοιες μέθοδοι βασίζονται συνήθως σε αλγορίθμους που υπολογίζουν την ομοιότητα μεταξύ των κόμβων. Ωστόσο, με την επιβλεπόμενη μάθηση, μας δίνεται η δυνατότητα συνδυασμού πολλών παρόμοιων χαρακτηριστικών και όχι μόνο δεικτών ομοιότητας.

Η παρούσα εργασία θα χωριστεί σε τέσσερα βήματα. Παρακάτω παρουσιάζονται τα βήματα για την υλοποίηση του συστήματος σύστασης.



Σχήμα 5.1: Σχέδιο υλοποίησης του συστήματος σύστασης

Οι αλγόριθμοι μηχανικής μάθησης που θα χρησιμοποιηθούν είναι οι **Random Forest, XGBoost, LightGBM** και τα **νευρωνικά δίκτυα**.

5.2 Δειγματοληψία

Για τη δειγματοληψία, θα επιλέξουμε συνδυασμούς χρηστών ανά δύο του γραφήματος μας, ορίζοντας τον έναν ως αρχικό και τον άλλο ως τελικό. Αυτό το στάδιο είναι απαραίτητο καθώς το γράφημά μας είναι κατευθυνόμενο και με αυτό τον τρόπο ορίζουμε τη κατεύθυνση της ένωσης. Αρχίζουμε τη δειγματοληψία μας βασιζόμενοι στη δημιουργία ενός συνόλου δεδομένων με συνδέσμους χρηστών ανά δύο, με τους μισούς συνδυασμούς να είναι υπαρκτοί από τα αρχικά μας δεδομένα και τους άλλους μισούς να είναι σύνδεσμοι που δεν υπάρχουν σε αυτά. Το χαρακτηριστικό connected υποδεικνύει την ύπαρξη (1) ή μη (0) σύνδεσης μεταξύ των δύο χρηστών. Ένα δείγμα του συνόλου αυτού είναι το ακόλουθο.

	source	destination	connected
0	1	690569	1
1	170	358857	1
2	171	1784396	1
3	320	1264771	1
50043	333324	444707	0
50044	820415	1203346	0
50045	155117	97140	0
50046	82597	1410277	0

Σχήμα 5.2: Δειγματοληψία

Η επιλογή των κόμβων που δεν υπάρχει ένωση έγινε με τυχαία επιλογή, από κόμβους που έχουν απόσταση μεταξύ τους μεγαλύτερη των 2 κόμβων. Η υλοποίηση έγινε με τη συνάρτηση shortest_path της βιβλιοθήκης networkx και έγινε βάσει του αρχικού μας γραφήματος.

Ο δυαδικός διαχωρισμός αυτός γίνεται για να εκπαιδεύσουμε τα μοντέλα μας σε ένα σύστημα δυαδικής κατηγοριοποίησης. Αποτέλεσμα αυτού, θα είναι να μπορούμε να προβλέψουμε σε ένα δίκτυο ποιοι κόμβοι πιστεύουμε πως θα έπρεπε να συνδέονται. Αυτή η

διαδικασία θα γίνει παίρνοντας τους κόμβους που δεν είναι συνδεδεμένοι ανά δύο και προβλέποντας βάσει των χαρακτηριστικών που έχουν εξαχθεί από το δίκτυο, αν θα έπρεπε να υπάρχει σύνδεση των δύο κόμβων. Στα μέσα κοινωνικής δικτύωσης, αυτό σημαίνει πως θα προβλέψουμε εάν κάποιος χρήστης θα έπρεπε να ακολουθεί τον άλλο, βάσει του μοντέλου μηχανικής μάθησης.

5.3 Εξαγωγή χαρακτηριστικών

Στη συνέχεια, υπολογίζονται τα χαρακτηριστικά όλων των ζευγαριών που δημιουργήθηκαν για να προχωρήσουμε στη διαδικασία της μηχανικής μάθησης. Τα χαρακτηριστικά αυτά μπορούν να χωριστούν σε πέντε κυρίως κατηγορίες :

- Web page Importance, υπολογισμός της σημαντικότητας κάθε σελίδας αναλόγως του αριθμού των followers και των followees της (Hits algorithm, Hubs, Authorities, Page Rank).
- Χαρακτηριστικά δικτύων, υπολογίζει διάφορα χαρακτηριστικά όπως η βαρύτητα του κάθε κόμβου σε ένα δίκτυο, ομοιότητα δύο κόμβων βάσει των ακμών από και προς κάθε κόμβο (adam/adar index, Katz algorithm).
- Πιθανότητα ομοιότητας, υπολογισμός ομοιότητας κόμβων (ομοιότητα συνημιτόνου, jaccard similarity).
- Γενικά χαρακτηριστικά κοινωνικών δικτύων, όπως κοινό αριθμό followers, αριθμό followers και followees κάθε κόμβου, εάν υπάρχει ήδη ακολούθηση από τον followee προς τον follower.
- Προσεγγίσεις για μέσα κοινωνικής δικτύωσης, τάση του πλούσιου να γίνει πλουσιότερος, πιθανότητα ύπαρξης δημοσίου προσώπου που δε θα γνωρίζει τους περισσότερους ακολούθους του, μικρότερη διαδρομή από ένα χρήστη σε έναν άλλο (Preferential Attachment, Weight of Edges, shortest path).

Τα χαρακτηριστικά που εξήχθησαν από το γράφημα εξηγούνται αναλυτικά στο τρίτο κεφάλαιο και είναι τα παρακάτω:

Δείκτες ομοιότητας	Χαρακτηριστικά δικτύων
1. ομοιότητα συνημιτόνου για followers	1. katz για follower
2. ομοιότητα συνημιτόνου για followees	2. katz για followee
3. jaccard simillarity για followers	3. adar index
4. jaccard simillarity για followees	

Δείκτες σημαντικότητας ιστοσελίδων	Γενικά χαρακτηριστικά κοινωνικών δικτύων	Προσεγγίσεις για μέσα κοινωνικής δικτύωσης
1. page rank του follower	1. followers του follower	1. εσωτερικό weight
2. page rank του followee	2. followers του followee	2. εξωτερικό weight
3. hubs του follower	3. followees του follower	3. weight version 1
4. hubs του followee	4. followees του followee	4. weight version 2
5. authorities του follower	5. κοινοί followers	5. weight version 3
6. authorities του followee	6. κοινοί followees	6. weight version 4
	7. follow back	7. Prefer. Attachment για followers
		8. Prefer. Attachment για followees
		9. συντομότερο μονοπάτι
		10. κοινός υπογράφος (WCC)

Πίνακας 5.1: Τα 30 χαρακτηριστικά για εκπαίδευση

Πιο συγκεκριμένα, τα χαρακτηριστικά ανά κατηγορία υπολογίζουν τις παρακάτω τιμές:

Στους **δείκτες ομοιότητας**, υπολογίζονται οι ομοιότητες, με τιμές από 0 ως 1, των followers και των followees των δύο χρηστών, την σύνδεση των οποίων προβλέπουμε, με τις μετρικές jaccard και συνημιτόνου.

Στις **προβλέψεις δικτύων**, υπολογίζεται η κεντρικότητα (centrality) katz των δύο χρηστών μέσα στο δίκτυο όπου παίρνει τιμές από 0 ως 1. Το Adar index κάνει σύσταση του συνδέσμου μεταξύ των 2 χρηστών, υπολογίζοντας τους κοινούς γείτονες τους.

Στους **δείκτες σημαντικότητας ιστοσελίδων**, υπολογίζονται η βαρύτητα των follower και followee μέσα στο δίκτυο, βάσει του αλγορίθμου Page Rank. Επίσης, υπολογίζει το πόσους followers και followees έχουν οι δύο χρήστες, συγκριτικά με ολόκληρο το δίκτυο, βάσει των αλγορίθμων authorities και hubs αντίστοιχα.

Στα **γενικά χαρακτηριστικά κοινωνικών δικτύων**, υπολογίζονται οι followers και οι followees των δύο χρηστών, ο αριθμός των κοινών followers και followees τους και η ύπαρξη σύνδεσης με αντίθετη κατεύθυνση από αυτή που ψάχνουμε.

Στις **Προσεγγίσεις για μέσα κοινωνικής δικτύωσης**, υπολογίζεται το βάρος που έχει ο follower στις ενώσεις που ακολουθεί εκείνος και το βάρος που έχει ο followee όταν τον ακολουθούν. Το βάρος αυτό είναι αντιστρόφως ανάλογο με τον αριθμό των followers και followees τους αντίστοιχα. Επίσης υπολογίζονται τέσσερις διαφορετικοί συνδυασμοί από τα δύο βάρη που προαναφέρθηκαν. Πιο συγκεκριμένη αναφορά σε αυτά γίνεται στο 3ο κεφάλαιο. Ακόμα, υπολογίζεται το Preferential attachment για τους followers και followees των 2 χρηστών. Ένα χαρακτηριστικό είναι και το συντομότερο μονοπάτι από τον ένα χρήστη στον άλλο. Τελευταίο χαρακτηριστικό, αποτελεί η σύνδεση των δύο χρηστών σε κοινό υπογράφημα.

Χρησιμοποιώντας τα προαναφερθέντα χαρακτηριστικά, δημιουργείται ένα σύνολο δεδομένων με 100.000 ενώσεις μεταξύ κόμβων. Όταν τελειώσει η διαδικασία, τα δεδομένα βρίσκονται στην εξής μορφή:

	source	destination	connected	jaccard_followers	jaccard_followees	cosine_followers	cosine_followees	num_followers_s	num_followers_d
0	1	690569	1	0	0.090909	0.039817	0.251976	3	29
1	18	1003537	1	0	0.058824	0.092450	0.149071	13	3
2	25	992602	1	0	0.242424	0.138866	0.440386	21	11
3	61	1408376	1	0	0.315789	0.151523	0.514496	8	7
4	62	402932	1	0	0.000000	0.149071	0.000000	5	3
5	98	472174	1	0	0.000000	0.000000	0.000000	3	10
6	109	1003858	1	0	0.000000	0.000000	0.000000	1	1
7	113	547644	1	0	0.000000	0.000000	0.000000	0	2
8	135	717660	1	0	0.277778	0.117688	0.466478	45	19
9	186	311232	1	0	0.171429	0.053333	0.321634	9	25
10	191	1157784	1	0	0.000000	0.000000	0.000000	8	24
11	216	530173	1	0	0.105263	0.089803	0.223607	31	2
12	226	1775565	1	0	0.000000	0.000000	0.000000	1	6

Σχήμα 5.3: Τελικό σύνολο δεδομένων

5.4 Εκπαίδευση μοντέλων

Στην εκπαίδευση των μοντέλων έγιναν δοκιμές με σκοπό να εντοπιστούν ποια χαρακτηριστικά είναι πιο χρήσιμα σε κάθε αλγόριθμο. Με αυτό τον τρόπο, θα γίνει κατανοητή η λογική της σύστασης τους στα επόμενα στάδια. Ακόμα, θα δούμε το ποσοστό μείωσης της απόδοσης ανά χαρακτηριστικό, για να αποφασίσουμε το καλύτερο δυνατό μοντέλο. Επίσης, θα γίνουν πειράματα με τις παραμέτρους του κάθε αλγορίθμου, για να αποφασιστεί ποιες τιμές λειτουργούν καλύτερα. Μας ενδιαφέρει κυρίως η απόδοση των συστημάτων σχετικά με τη μετρική F1-score, καθώς ασχολείται με την αρμονική μέση τιμή των συναρτήσεων precision και recall. Αυτές οι μετρικές δίνουν στατιστικά αποτελέσματα των θετικών τιμών που θα μας δείξουν τις συστάσεις που θα κάνουμε. Ο έλεγχος της απόδοσης έγινε με 5 τμήματα cross validation.

5.4.1 Χαρακτηριστικά

Στην παρούσα παράγραφο θα γίνει εμφανές μέσα από μια σειρά δοκιμών πως υπάρχουν δύο ομάδες χαρακτηριστικών που αποδίδουν καλύτερα στην εκπαίδευση όλων των μοντέλων. Η μία αποτελείται από 5 και η άλλη από 13 χαρακτηριστικά. Θα γίνει ακόμα σύγκριση

των δύο αυτών ομάδων με 10 και 7 χαρακτηριστικά. Η μέθοδος με την οποία αποφασίστηκαν τα συγκεκριμένα χαρακτηριστικά βασίστηκε σε πρώτο βαθμό σε συναφείς εργασίες [4], που σύγκριναν τις αποδόσεις των χαρακτηριστικών σε παρόμοια γραφήματα καθώς και σε πειράματα που έγιναν στην παρούσα εργασία.

Ο λόγος μείωσης του αριθμού των χαρακτηριστικών είναι τόσο η βελτίωση του χρόνου εκπαίδευσης, όσο και του χρόνου υπολογισμού της σύστασης που θα εκτελεί το σύστημα. Επιπρόσθετα, θα είναι πιο ξεκάθαρη η λογική με την οποία γίνεται η σύσταση.

Τα χαρακτηριστικά που έδειξαν πως έχουν τη καλύτερη απόδοση στα μοντέλα με 5 χαρακτηριστικά είναι τα:

1. same weakly connected component
2. follow back
3. weight version 4
4. Preferential Attachement για followers
5. shortest path from source to destination

Πίνακας 5.2: Τα 5 χαρακτηριστικά για εκπαίδευση

Ενώ τα χαρακτηριστικά που έδειξαν πως έχουν τη καλύτερη απόδοση στα μοντέλα με 13 χαρακτηριστικά είναι τα:

1. same weakly connected component
2. follow back
3. weight version 4
4. Preferential Attachement για followers
5. shortest path from source to destination
6. Preferential Attachement για followees
7. weight version 2
8. ομοιότητα συνημιτόνου για followees
9. hubs για followees
10. hits importance για followers
11. ομοιότητα συνημιτόνου για followers
12. katz simillarity destination
13. authorities για followees

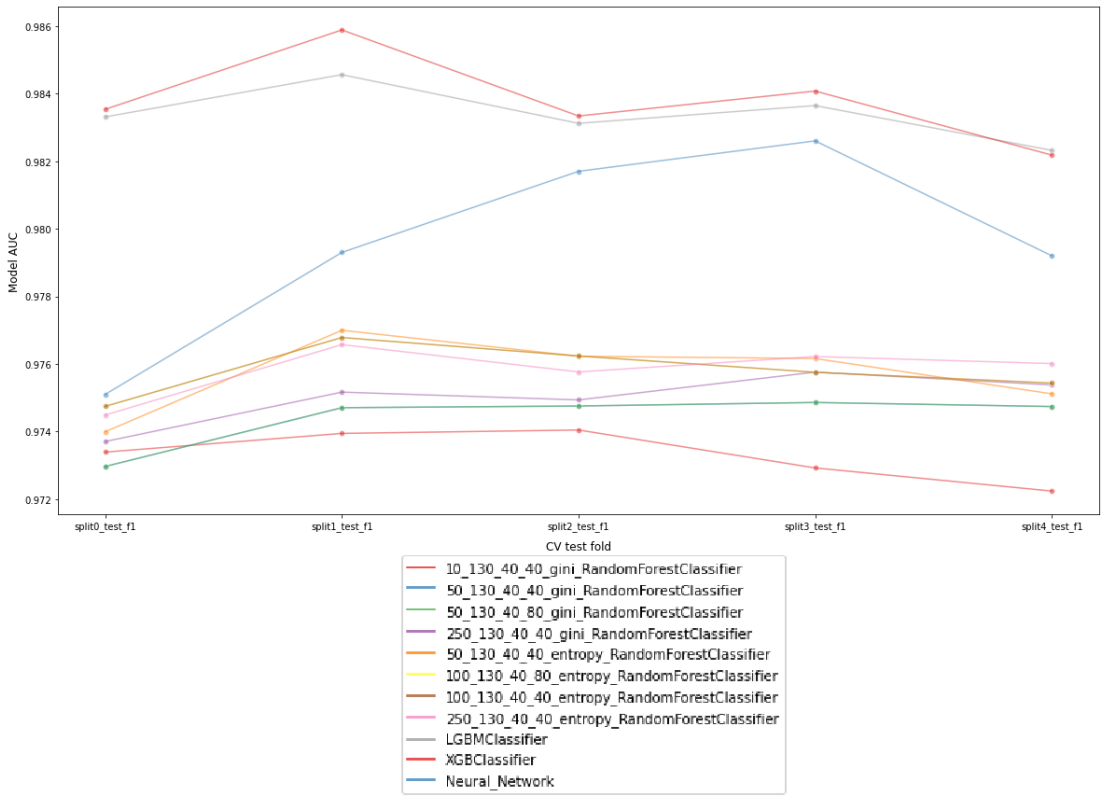
Πίνακας 5.3: Τα 13 χαρακτηριστικά για εκπαίδευση

5.4.2 Επίδοση όλων των αλγορίθμων

Με τη χρήση όλων των δεδομένων που έχουμε εξάγει από το δίκτυο, όλοι οι αλγόριθμοι έχουν πολύ καλά αποτελέσματα και στα 5 τμήματα που έχουν χωριστεί τα δεδομένα. Στη συνέχεια θα δούμε τη γραφική αναπαράσταση της απόδοσης των μοντέλων με τη μετρική F1-score, όπως και τον μέσο όρο των 5 διαχωρισμών σε όλες τις μετρικές των αλγορίθμων που χρησιμοποιήθηκαν. Πιο συγκεκριμένα οι μετρικές που χρησιμοποιήθηκαν ήταν οι: **Precision, Recall, F1-score, Accuracy** και **Roc AUC**. Όπως αναφέρεται και στη εισαγωγή του κεφαλαίου, οι αλγόριθμοι μηχανικής μάθησης που θα χρησιμοποιηθούν είναι οι: **Random Forest, XGBoost, LightGBM** και τα **νευρωνικά δίκτυα**.

Επίδοση εκπαίδευσης με 30 χαρακτηριστικά

Αρχικά βλέπουμε την επίδοση των μοντέλων με τη χρήση και των 30 διαθέσιμων χαρακτηριστικών. Στο σχήμα 5.4 αναπαριστάται γραφικά η επίδοση των αλγορίθμων στα 5 τμήματα των δεδομένων με τη μετρική F1-score.



Σχήμα 5.4: F1 επίδοση όλων των αλγορίθμων με 30 χαρακτηριστικά

Στον πίνακα 5.4, βλέπουμε επίσης την αριθμητική επίδοση όλων των αλγορίθμων, τις μετρικές οι οποίες χρησιμοποιούνται για την επίδοση των μοντέλων σε μέσο όρο των 5 τμημάτων των δεδομένων.

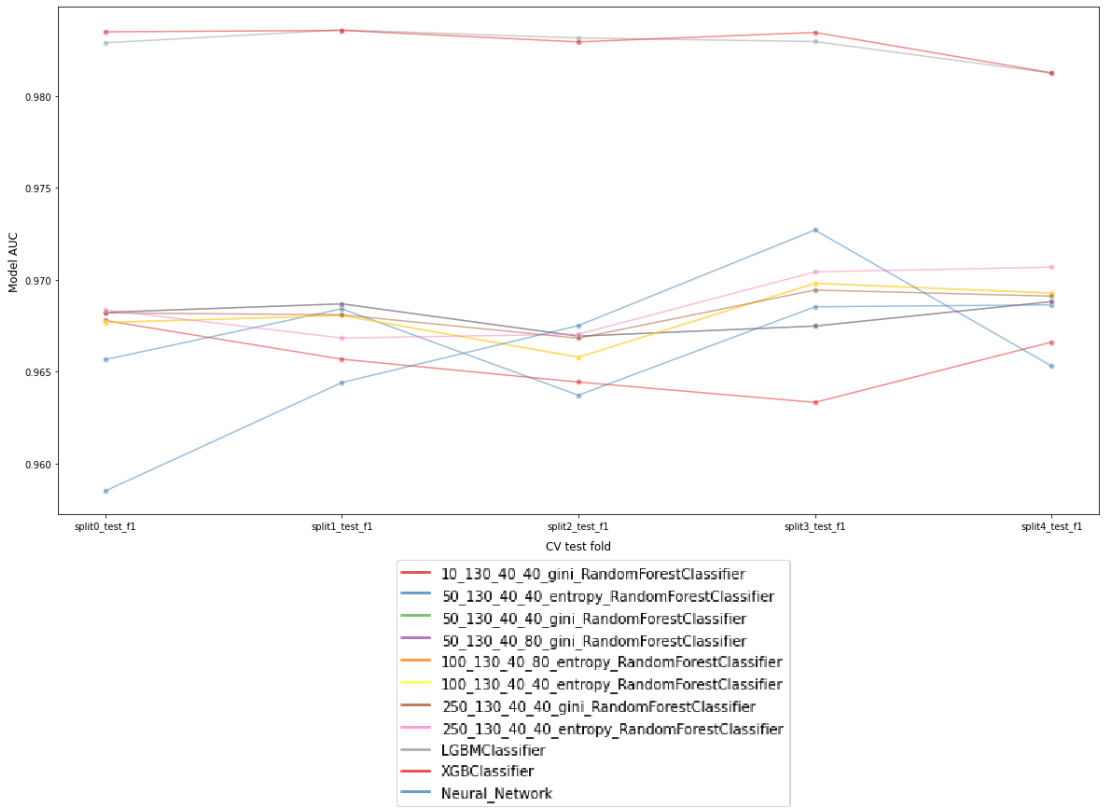
Εκπαίδευση μοντέλων με 30 χαρακτηριστικά					
Μοντέλα	Precision	Recall	F1-score	Roc AUC	Accuracy
250_80_gini_RF	0.989323	0.962626	0.975790	0.997037	0.976149
250_40_entr_RF	0.989323	0.962687	0.975811	0.997081	0.976169
LGBMClassifier	0.990537	0.976348	0.983391	0.998776	0.983538
XGBClassifier	0.989866	0.977813	0.983802	0.998815	0.983932
Neural_Network	0.989580	0.966690	0.979341	0.997988	0.980169

Πίνακας 5.4: Επιδόσεις όλων των μοντέλων με 30 χαρακτηριστικά

Ωστόσο η χρήση 30 διαφορετικών χαρακτηριστικών και ο υπολογισμός τους είναι πολύ χρονοβόρος. Για αυτό το λόγο επιλέχθηκαν οι 4 εναλλακτικές των 13, 10, 7 και 5 χαρακτηριστικών, ανάλογα με την ισχύ του συστήματός μας.

Επίδοση εκπαίδευσης με 13 χαρακτηριστικά

Στη συνέχεια, βλέπουμε την επίδοση των μοντέλων με τη χρήση 13 διαθέσιμων χαρακτηριστικών. Στο σχήμα 5.5 αναπαριστάται γραφικά η επίδοση των αλγορίθμων στα 5 τμήματα των δεδομένων με τη μετρική F1-score.



Σχήμα 5.5: F1 επίδοση όλων των αλγορίθμων με 13 χαρακτηριστικά

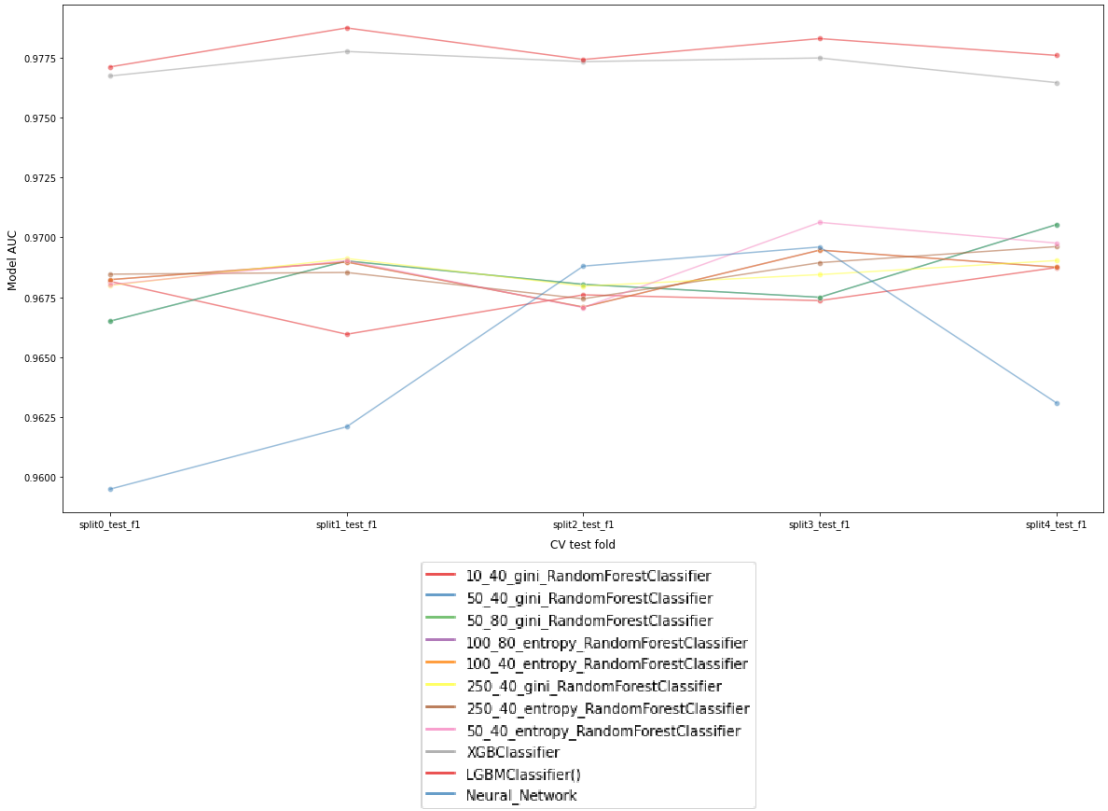
Στον πίνακα 5.5, αναπαριστάται η αριθμητική επίδοση όλων των αλγορίθμων όπως υπολογίστηκαν και στα 30 χαρακτηριστικά παραπάνω.

Εκπαίδευση μοντέλων με 13 χαρακτηριστικά					
Μοντέλα	Precision	Recall	F1-score	Roc AUC	Accuracy
250_80_gini_RF	0.990351	0.946859	0.968113	0.995191	0.968843
250_40_entr_RF	0.990358	0.947902	0.968663	0.995287	0.969365
LGBMClassifier	0.991066	0.974603	0.982765	0.998517	0.982935
XGBClassifier	0.989371	0.976589	0.982938	0.998604	0.983080
Neural_Network	0.990251	0.946999	0.968015	0.995176	0.968741

Πίνακας 5.5: Επιδόσεις όλων των μοντέλων με 13 χαρακτηριστικά

Επίδοση εκπαίδευσης με 10 χαρακτηριστικά

Παρόμοια υπολογίζεται και η επίδοση των μοντέλων με τη χρήση 10 διαθέσιμων χαρακτηριστικών. Στο σχήμα 5.6 παρουσιάζεται το γράφημα της επίδοσης των αλγορίθμων με τη μετρική F1-score.



Σχήμα 5.6: F1 επίδοση όλων των αλγορίθμων με 10 χαρακτηριστικά

Από το σχήμα, βλέπουμε πως η μείωση της μετρικής F1-score είναι σημαντική σε όλα τα μοντέλα.

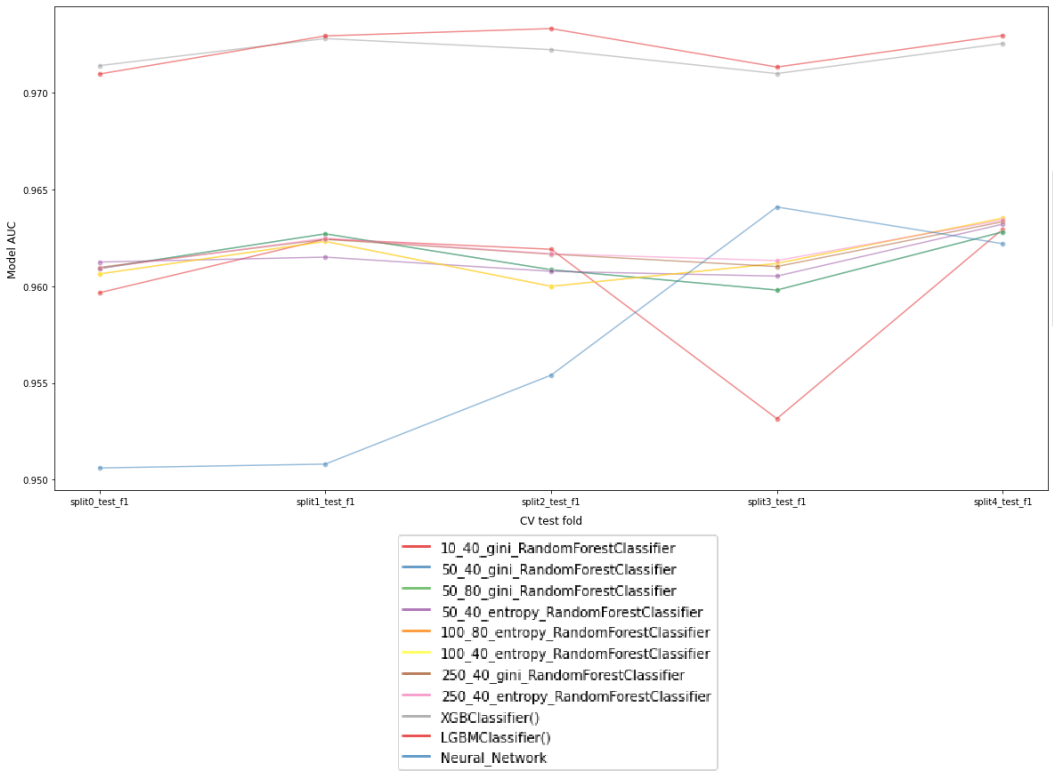
Στον πίνακα 5.6 θα γίνει εμφανέστερο με τις συγκεκριμένες αριθμητικές επιδόσεις όλων των αλγορίθμων, όπως και στις προηγούμενες παραγράφους. Υπολογίζονται επίσης, όλες οι μετρικές οι οποίες χρησιμοποιούνται.

Εκπαίδευση μοντέλων με 10 χαρακτηριστικά					
Μοντέλα	Precision	Recall	F1-score	Roc AUC	Accuracy
250_80_gini_RF	0.987474	0.950269	0.968512	0.994505	0.969143
250_40_entr_RF	0.987375	0.950530	0.968600	0.994625	0.969223
LGBMClassifier	0.988761	0.967140	0.977831	0.997191	0.978107
XGBClassifier	0.986268	0.968203	0.977152	0.997016	0.977402
Neural_Network	0.985146	0.948015	0.963251	0.992999	0.968147

Πίνακας 5.6: Επιδόσεις όλων των μοντέλων με 10 χαρακτηριστικά

Επίδοση εκπαίδευσης με 7 χαρακτηριστικά

Σε αυτό το στάδιο βλέπουμε την επίδοση με τη χρήση 7 χαρακτηριστικών. Το σχήμα 5.7 δείχνει τη μετρική F1-score όλων των αλγορίθμων στα 5 τμήματα των δεδομένων.



Σχήμα 5.7: F1 επίδοση όλων των αλγορίθμων με 7 χαρακτηριστικά

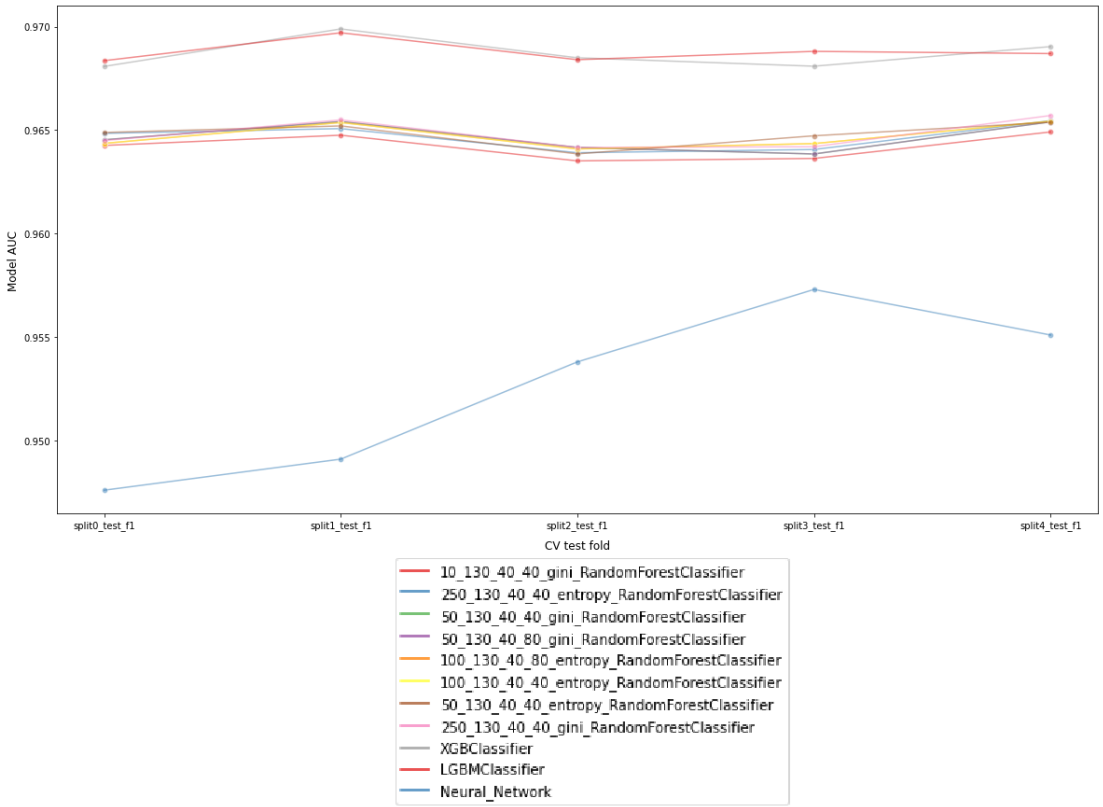
Από το σχήμα, φαίνεται πως η μείωση της μετρικής F1-score συνεχίζει να μειώνεται σε όλα τα μοντέλα. Στον πίνακα 5.7, θα δούμε πιο αναλυτικά τις τιμές των επιδόσεων σε όλες τις μετρικές, όπως και στις προηγούμενες παραγράφους.

Εκπαίδευση μοντέλων με 7 χαρακτηριστικά					
Μοντέλα	Precision	Recall	F1-score	Roc AUC	Accuracy
250_80_gini_RF	0.989412	0.935183	0.961530	0.991785	0.962617
250_40_entr_RF	0.989149	0.936086	0.961884	0.991694	0.962938
LGBMClassifier	0.986920	0.958133	0.972313	0.994947	0.972756
XGBClassifier	0.984567	0.959758	0.972003	0.994649	0.972402
Neural_Network	0.989876	0.933125	0.95341	0.990097	0.962941

Πίνακας 5.7: Επιδόσεις όλων των μοντέλων με 7 χαρακτηριστικά

Επίδοση εκπαίδευσης με 5 χαρακτηριστικά

Τέλος, βλέπουμε την επίδοση των μοντέλων με τη χρήση 5 διαθέσιμων χαρακτηριστικών. Στο σχήμα 5.8 αναπαριστάται γραφικά η επίδοση των αλγορίθμων στα 5 τμήματα των δεδομένων με τη μετρική F1-score.



Σχήμα 5.8: F1 επίδοση όλων των αλγορίθμων με 5 χαρακτηριστικά

Ο πίνακας 5.8, δείχνει την αριθμητική επίδοση όλων των αλγορίθμων, στα 5 τμήματα των δεδομένων σε μέσο όρο, για όλες τις μετρικές που έχουν προαναφερθεί.

Εκπαίδευση μοντέλων με 5 χαρακτηριστικά					
Μοντέλα	Precision	Recall	F1-score	Roc AUC	Accuracy
100_80_gini_RF	0.985056	0.945213	0.964721	0.992972	0.965479
100_40_entr_RF	0.985160	0.945294	0.964813	0.992922	0.965569
LGBMClassifier	0.986527	0.951693	0.968796	0.994046	0.969386
XGBClassifier	0.983303	0.954562	0.968719	0.993908	0.969226
Neural_Network	0.975065	0.936398	0.963463	0.984834	0.953578

Πίνακας 5.8: Επιδόσεις όλων των μοντέλων με 5 χαρακτηριστικά

Παρατηρήσεις στην επίδοση των αλγορίθμων

Οι αλγόριθμοι με τα καλύτερα αποτελέσματα είναι οι boosting αλγόριθμοι, δηλαδή ο XGBoost και ο LighGBM. Η διαφορά της απόδοσης των δύο αυτών αλγορίθμων είναι σχεδόν μηδαμινή. Ωστόσο, τόσο στο χρόνο εκπαίδευσης όσο και στο χρόνο σύστασης ο LightGBM είναι εμφανώς καλύτερος.

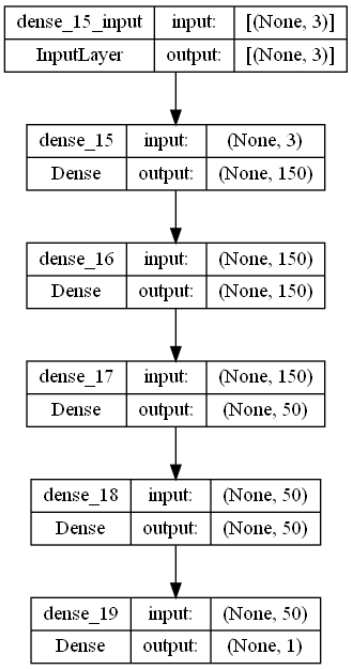
Πιο συγκεκριμένα, για την εκπαίδευση των μοντέλων με 5 χαρακτηριστικά ο LightGBM ήταν 7 φορές πιο ταχύς από τον XGBoost, ενώ με 13 χαρακτηριστικά σχεδόν 14 φορές. Ο LightGBM εφαρμόζει έναν εξαιρετικά βελτιστοποιημένο αλγόριθμο εκμάθησης δένδρων αποφάσεων που βασίζεται σε ιστογράμματα, ο οποίος αποφέρει μεγάλα πλεονεκτήματα τόσο στην απόδοση όσο και στην κατανάλωση μνήμης. Επίσης, μας δίνει την πιλογή δύο νέων τεχνικών εκπαίδευσης που ονομάζονται Gradient-Based One-Side Sampling (GOSS) και Exclusive Feature Bundling (EFB) που επιτρέπουν στον αλγόριθμο να εκτελείται πιο γρήγορα διατηρώντας παράλληλα υψηλό επίπεδο ακρίβειας. Αλγόριθμοι όπως ο XGBoost χρησιμοποιούν αλγόριθμους δέντρων αποφάσεων βάσει ταξινόμησης, οι οποίοι αναζητούν το καλύτερο σημείο διαχωρισμού σε ταξινομημένες τιμές χαρακτηριστικών, που αν και αποδοτικός είναι αρκετά χρονοβόρος.

Επίσης, βλέπουμε πως με τη χρήση 13 χαρακτηριστικών η απόδοση των μοντέλων είναι πολύ κοντά με αυτή που χρησιμοποιούνται 30 χαρακτηριστικά. Από εκείνο το στάδιο και μετά η πτώση είναι εμφανής, ενώ ο μικρότερος αριθμός χαρακτηριστικών που παρέχει αποτελέσματα μεγαλύτερα του 95% είναι εκείνος των 5 χαρακτηριστικών. Για αυτό τον λόγο, οι αναλύσεις των αλγορίθμων μηχανικής μάθησης θα γίνουν με 13 και 5 χαρακτηριστικά.

Παρακάτω ακολουθεί η περιγραφή της παραμετροποίησής του κάθε ενός αλγορίθμου καθώς και η επιμέρους ανάλυση της επίδοσης του.

5.4.3 Ανάλυση επίδοσης Νευρωνικών Δικτύων

Τα Νευρωνικά Δίκτυα χρησιμοποιούνται σε μεγάλο βαθμό στα συστήματα σύστασης. Έχει ενδιαφέρον λοιπόν, να δούμε πως αποδίδουν σε ένα σύστημα δυαδικής ταξινόμησης. Ακόμα, θέλουμε να δούμε ποια χαρακτηριστικά είναι τα πιο χρήσιμα σε αυτόν τον αλγόριθμο. Η εκπαίδευση του μοντέλου έγινε με 30 εποχές καθώς σε αυτή τη τιμή το μοντέλο έδειχνε σύγκλιση. Το μοντέλο που επιλέχθηκε είναι το παρακάτω:



Σχήμα 5.9: Μοντέλο νευρωνικού δικτύου

Το μοντέλο χρησιμοποιεί απλά Dense Layers καθώς είναι η προτεινόμενη δομή [21] για την υλοποίηση των layers σε τέτοιου τύπου προβλήματα. Για το μοντέλο χρησιμοποιείται 1 layer εισόδου, 4 hidden layers και 1 layer εξόδου. Στα 4 hidden layers χρησιμοποιούμε 50 units, καθώς δεν υπάρχει ιδιαίτερη βελτίωση μετά από πειράματα στο συγκεκριμένο σύνολο δεδομένων. Η συνάρτηση ενεργοποίησης των hidden layers είναι η Relu ενώ στο output layer η Sigmoid. Ο βελτιστοποιητής που επιλέχθηκε είναι ο Adam και είναι μια μέθοδος στοχαστικής Gradient Descent. Ο Adam είναι υπολογιστικά αποδοτικός, έχει μικρή απαίτηση μνήμης και είναι κατάλληλος για προβλήματα που είναι μεγάλα από άποψη δεδομένων ή και παραμέτρων.

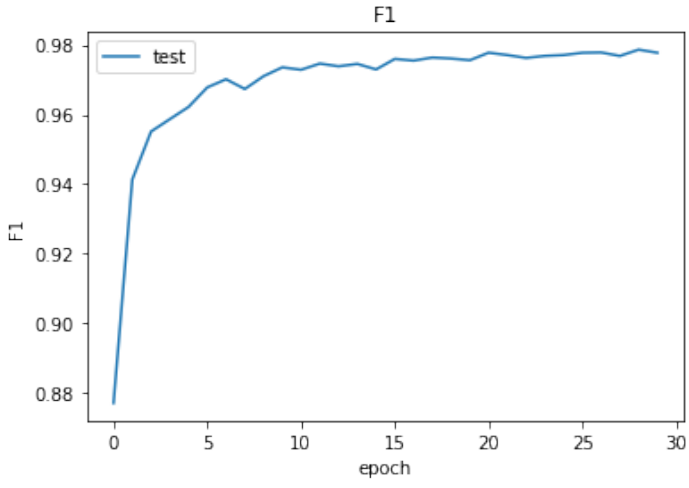
Επιλογή Χαρακτηριστικών

Αρχικά, θα δούμε ποια χαρακτηριστικά χρησιμοποιεί το συγκεκριμένο μοντέλο νευρωνικών δικτύων στη εκπαίδευση του. Αυτή η διαδικασία θα γίνει με τη συνάρτηση permutation_importance της βιβλιοθήκης sklearn, που έχει ως είσοδο ένα εκπαιδευμένο μοντέλο και ως έξοδο τα χαρακτηριστικά με τη μεγαλύτερη βαρύτητα στην εκπαίδευση.

Weight	Feature
0.2110 ± 0.0019	follows_back
0.1193 ± 0.0015	inter_followers
0.1049 ± 0.0014	adar_index
0.0974 ± 0.0023	inter_followees
0.0948 ± 0.0011	shortest_path
0.0948 ± 0.0015	num_followers_d
0.0762 ± 0.0012	num_followees_s
0.0720 ± 0.0013	prefer_Attach_followers
0.0700 ± 0.0013	prefer_Attach_followees
0.0329 ± 0.0012	num_followees_d
0.0270 ± 0.0009	num_followers_s
0.0185 ± 0.0002	same_comp
0.0076 ± 0.0004	cosine_followees
0.0066 ± 0.0003	weight_f3
0.0062 ± 0.0004	weight_f4
0.0048 ± 0.0002	weight_f2
0.0043 ± 0.0005	jaccard_followees
0.0013 ± 0.0004	cosine_followers
0.0010 ± 0.0002	weight_f1
0.0008 ± 0.0002	weight_in
… *10 more* …	

Σχήμα 5.10: Σημαντικότητα χαρακτηριστικών νευρωνικού δικτύου με 30 χαρακτηριστικά

Στη συνέχεια θα δούμε την επίδοση του μοντέλου ανά εποχή, με τη μετρική F1-score.



Σχήμα 5.11: F1 νευρωνικού δικτύου με 30 χαρακτηριστικά

Η απόδοση του F1-score στον μέσο όρο των 5 τμημάτων είναι κοντά στο 98%. Η απόδοση είναι αρκετά ικανοποιητική με τη μικρότερη τιμή του F1-score στα 5 τμήματα να είναι 97.6% και η μεγαλύτερη 98.2%.

Επίδοση με 13 και 5 χαρακτηριστικά

Αρχικά, θα δούμε ποια χαρακτηριστικά χρησιμοποιεί το συγκεκριμένο μοντέλο νευρωνικών δικτύων στη εκπαίδευση του με τη συνάρτηση permutation_importance, με χρήση 13 και 5 χαρακτηριστικών αντίστοιχα.

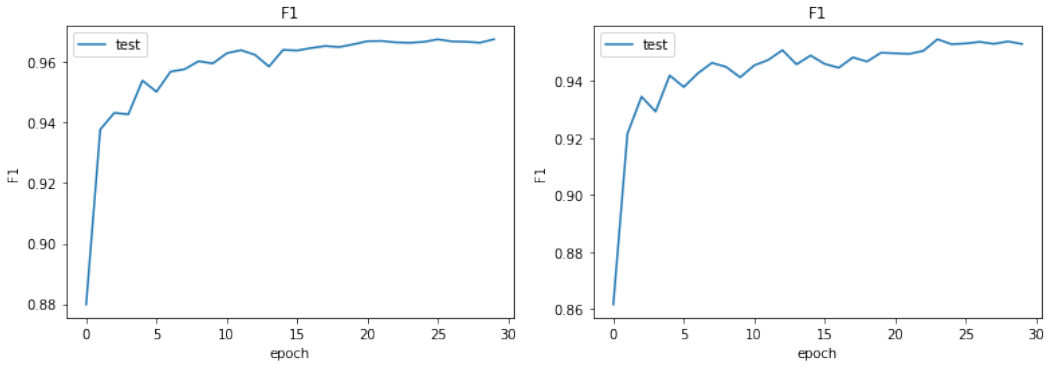
Weight	Feature
0.2098 ± 0.0008	follows_back
0.1774 ± 0.0008	shortest_path
0.0770 ± 0.0013	cosine_followees
0.0669 ± 0.0005	prefer_Attach_followers
0.0290 ± 0.0008	weight_f4
0.0264 ± 0.0007	prefer_Attach_followees
0.0238 ± 0.0010	cosine_followers
0.0157 ± 0.0002	same_comp
0.0103 ± 0.0001	weight_f2
0.0000 ± 0.0000	hubs_d
0.0000 ± 0.0000	authorities_d
0.0000 ± 0.0000	hubs_s
0.0000 ± 0.0000	katz_d

Σχήμα 5.12: Σημαντικότητα χαρακτηριστικών νευρωνικού δικτύου με 13 χαρακτηριστικά

Weight	Feature
0.2069 ± 0.0023	follows_back
0.2034 ± 0.0017	shortest_path
0.0295 ± 0.0006	prefer_Attach_followers
0.0176 ± 0.0005	weight_f4
0.0151 ± 0.0005	same_comp

Σχήμα 5.13: Σημαντικότητα χαρακτηριστικών νευρωνικού δικτύου με 5 χαρακτηριστικά

Είναι εμφανές, πως το μοντέλο χρησιμοποιεί σε μεγάλο βαθμό το χαρακτηριστικό follow back και το shortest path για να κάνει τις προβλέψεις του. Αυτό μπορεί να μεταφραστεί λέγοντας πως ο αλγόριθμος θα κάνει προτάσεις σε χρήστες βάσει της παραμέτρου του συντομότερου μονοπατιού.



Σχήμα 5.14: F1 νευρωνικού με 13 και 5 χαρακτηριστικά αντίστοιχα

Ακόμα, θα δούμε πιο αναλυτικά την επίδοση των δύο αυτών μοντέλων με τη μετρική F1-score ανά εποχή. Παρακάτω, βλέπουμε πως υπάρχει σύγκλιση και των δύο στις 30 εποχές.

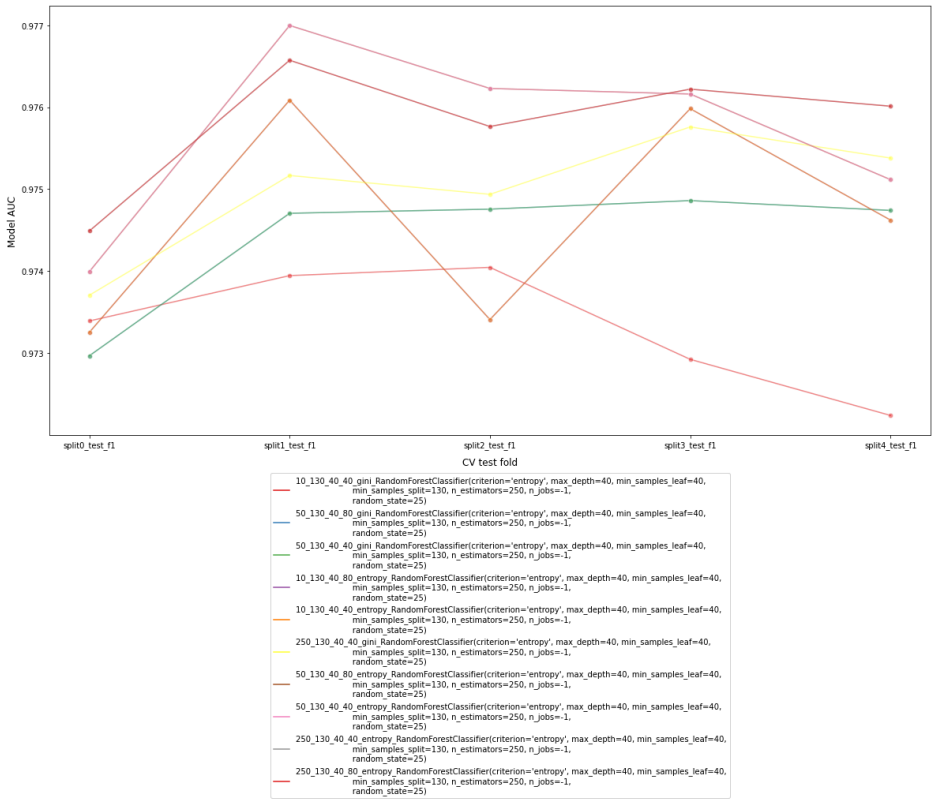
Με τη χρήση των 13 πιο σημαντικών χαρακτηριστικών που επιλέξαμε έχουμε απόδοση 96.6% με 30 εποχές. Όμοια, με τα 5 πιο σημαντικά χαρακτηριστικά έχουμε απόδοση 95.3% με 30 εποχές κατά μέσο όρο στα 5 τμήματα των δεδομένων μας για εκπαίδευση.

5.4.4 Ανάλυση επίδοσης Random Forest

Ο αλγόριθμος Random Forest επιλέχθηκε καθώς σε αρκετές συναφείς εργασίες [4] είχε από τα καλύτερα, αν όχι τα καλύτερα, αποτελέσματα από όλους τους αλγορίθμους. Με τη χρήση του GridsearchCV δοκιμάστηκαν αρκετές παράμετροι για τον αλγόριθμο, έτσι ώστε να έχουμε την καλύτερη αλλά και πιο σταθερή επίδοση στα αποτελέσματά μας. Στο παρακάτω σχήμα φαίνονται τα αποτελέσματα των αλγορίθμων με 5 τμήματα cross validation για την τιμή f1 του αλγορίθμου.

Το μοντέλο με $criterion =' entropy', max\_depth = 40, min\_samples\_leaf = 40,$ $min\_samples\_split = 130, n\_estimators = 250$ είχε τα καλύτερα αποτελέσματα από τα υπόλοιπα. Στη συνέχεια βλέπουμε τις αποδόσεις όλων των μοντέλων με τη μετρική F1-score σε 5 τμήματα.

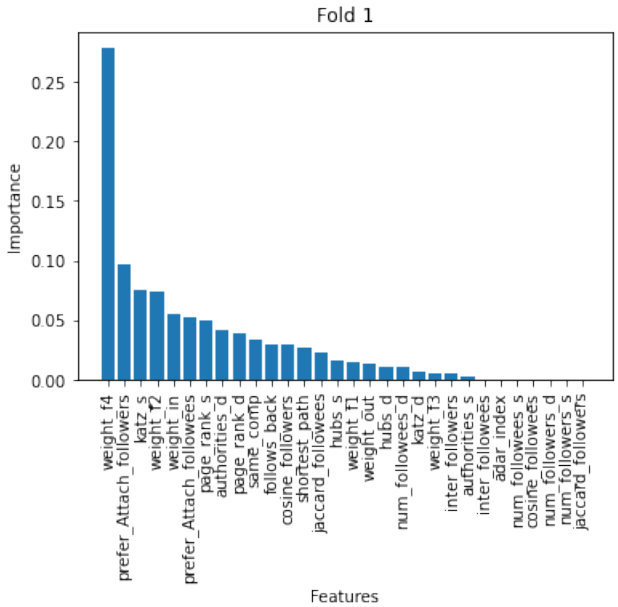
Το μοντέλο είναι πολύ αποδοτικό, φτάνοντας τον μέσο όρο των 5 τμημάτων σε απόδοση κοντά στο 97% στο F1-score.



Σχήμα 5.15: Random Forest επίδοση μοντέλων με διαφορετικές παραμέτρους

Επιλογή Χαρακτηριστικών

Τα χαρακτηριστικά που χρησιμοποιεί το συγκεκριμένο μοντέλο φαίνονται παρακάτω.

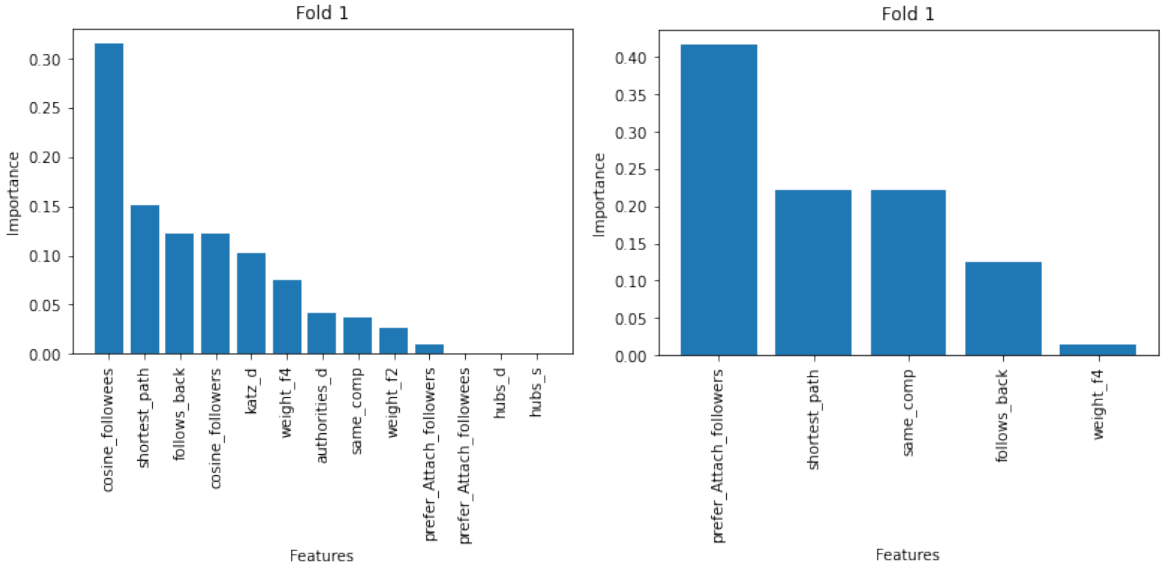


Σχήμα 5.16: Σημαντικότητα χαρακτηριστικών Random Forest με 30 χαρακτηριστικά

Το μοντέλο χρησιμοποιεί κυρίως τη παράμετρο weight_f4 σε μεγάλο βαθμό, σε συνδυασμό με πολλές άλλες σε μικρότερο. Πιο συγκεκριμένα, χρησιμοποιεί τα χαρακτηριστικά Prederential_Attachment για followers, katz για τον follower και weight version 2. Το σχήμα 5.16 αναπαριστά πιο συγκεκριμένα όλα τα χαρακτηριστικά με την επιρροή που έχουν σε αυτό το μοντέλο του Random Forest. Η αξία του κάθε χαρακτηριστικού ωστόσο, αλλάζει σε συνδυασμό με τα υπόλοιπα διαθέσιμα.

Επίδοση με 13 και 5 χαρακτηριστικά

Στη συνέχεια, βλέπουμε τη σημαντικότητα τους με 13 και 5 διαθέσιμα χαρακτηριστικά.



Σχήμα 5.17: Σημαντικότητα Random Forest με 13 και 5 χαρακτηριστικά αντίστοιχα

Το μοντέλο αυτό μας δίνει απόδοση 96.5% στο F1-score, με τη χρήση 5 χαρακτηριστικών και 97.5% με τη χρήση των 13. Η μείωση κατά 1% της απόδοσης, σε συνδυασμό με την αφαίρεση των χαρακτηριστικών που πρέπει να υπολογιστούν είναι αρκετά ικανοποιητική.

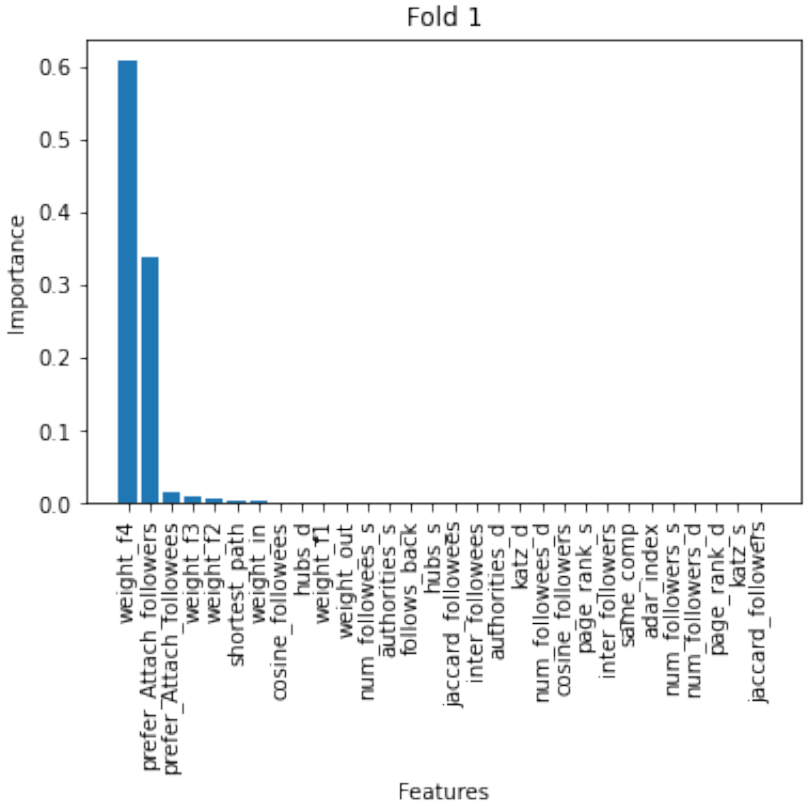
Το μοντέλο με τα 5 χαρακτηριστικά, χρησιμοποιεί σε μεγάλο βαθμό τη συνάρτηση Preferential Attachment για followers. Το μοντέλο με τα 13 χαρακτηριστικά, χρησιμοποιεί σε μεγάλο βαθμό το ομοιότητα συνημιτόνου για followees.

5.4.5 Ανάλυση επίδοσης Xgboost

Το μοντέλο με base_score=0.5, booster='gbtree', learning_rate=0.300000013, max_depth=10, n_estimators=109, n_jobs=4, verbosity=None, είχε τα καλύτερα αποτελέσματα από τα υπόλοιπα. Το μοντέλο είναι πολύ αποδοτικό, φτάνοντας τον μέσο όρο των 5 τμημάτων σε απόδοση κοντά στο 98.5% στο F1-score.

Επιλογή Χαρακτηριστικών

Στη παρακάτω εικόνα βλέπουμε τα χαρακτηριστικά με τη μεγαλύτερη αξία για τα μοντέλα. Τα μοντέλα αυτά χρησιμοποιούν κυρίως τη παράμετρο weight_f4 σε μεγάλο βαθμό, σε συνδυασμό με πολλές άλλες σε μικρότερο βαθμό.



Σχήμα 5.18: Σημαντικότητα χαρακτηριστικών XGBoost με 30 χαρακτηριστικά

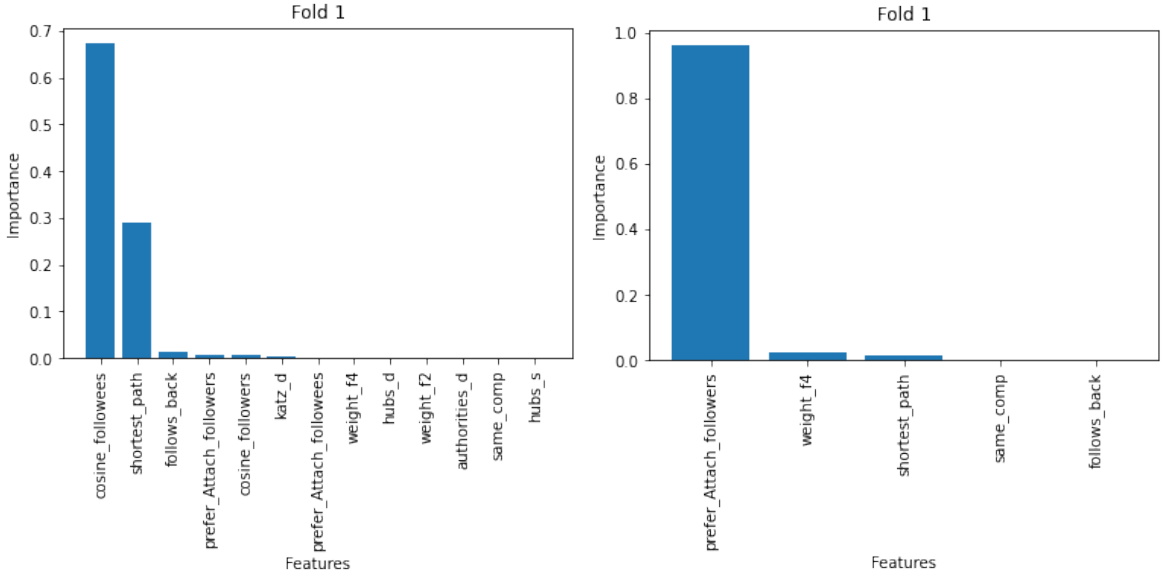
Μετά από αρκετές δοκιμές ωστόσο, βλέπουμε πως αλλάζει η αξία του κάθε χαρακτηριστικού. Έτσι, προσπαθώντας να συνδυάσουμε τη καλύτερη δυνατή απόδοση σε συνδυασμό με τη λιγότερο δυνατή υπολογιστική δύναμη καταλήγουμε στις δύο παρακάτω επιλογές χαρακτηριστικών.

Επίδοση με 13 και 5 χαρακτηριστικά

Στη συνέχεια, έχουμε ένα μοντέλο με 13 χαρακτηριστικά και απόδοση κατά μέσο όρο 98.3% στα 5 τμήματα που το μοντέλο εκπαιδεύτηκε. Το μοντέλο αυτό προτείνεται για συστήματα με μεγάλη υπολογιστική δύναμη και δίνουν σημασία στην υψηλή απόδοση του συστήματος.

Τέλος, το μοντέλο με τα 5 χαρακτηριστικά μας δίνει απόδοση 96.9% στο F1-score. Το μοντέλο αυτό προτείνεται σε πιο αδύναμα συστήματα, με απόδοση που απέχει 2% διαφορά από εκείνη του μοντέλου που χρησιμοποιούσε το σύνολο των δεδομένων.

Παρακάτω βλέπουμε τα χαρακτηριστικά που χρησιμοποιούν κατά κύριο λόγο τα δύο μοντέλα.



Σχήμα 5.19: Σημαντικότητα XGBoost με 13 και 5 χαρακτηριστικά αντίστοιχα

Η μείωση κατά 2% της απόδοσης, σε συνδυασμό με την αφαίρεση των χαρακτηριστικών που πρέπει να υπολογιστούν από 30 σε 5 αποτελεί το καλύτερο αποτέλεσμα που υπολογίστηκε.

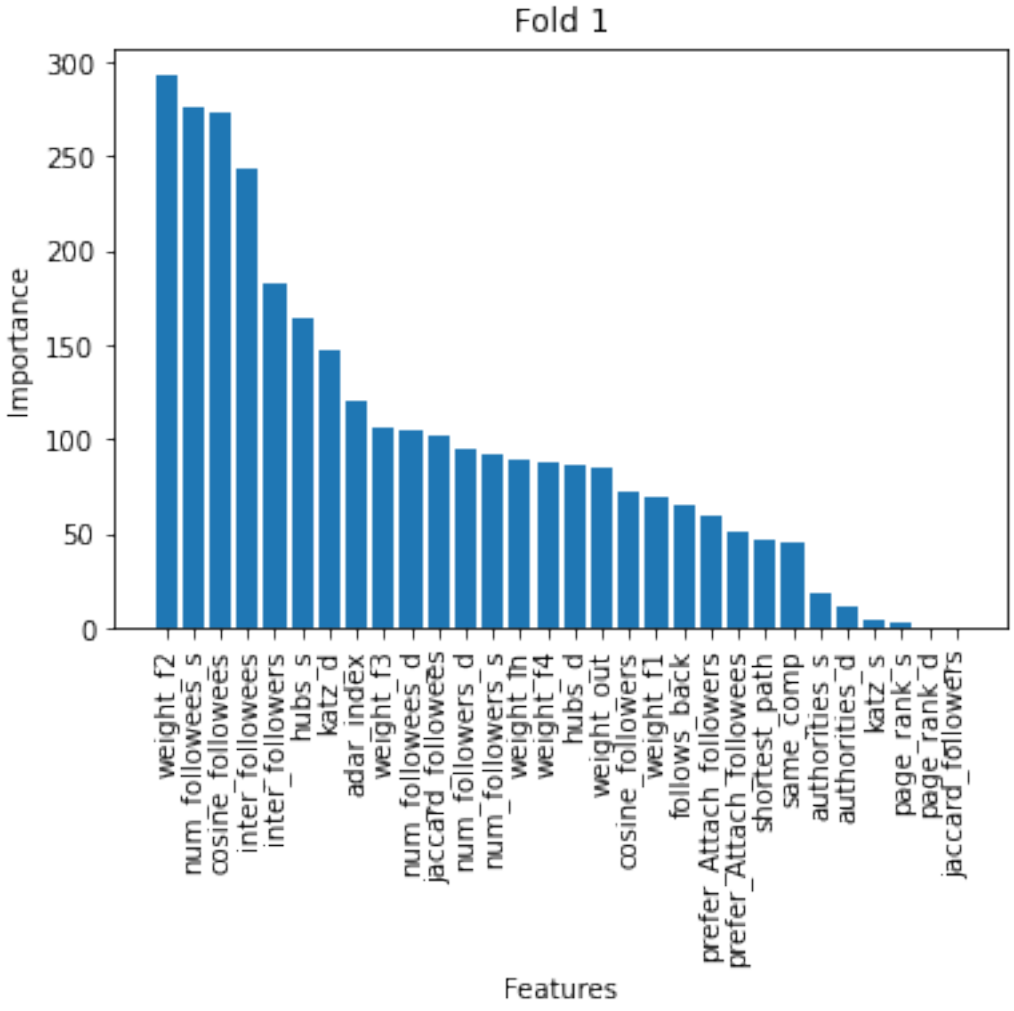
5.4.6 Ανάλυση επίδοσης LightGBM

Το μοντέλο με LightGBM είναι και αυτός ένας boosting αλγόριθμος και έχει δημιουργηθεί από τη microsoft. Ο αλγόριθμος αυτός έχει πλεονεκτήματα την ταχύτερη εκπαίδευση και την υψηλότερη απόδοση σε σχέση με άλλους αλγορίθμους τέτοιου τύπου. Ακόμα, χρησιμοποιεί λιγότερη μνήμη που μας χρησιμεύει όταν έχουμε τόσο μεγάλου όγκου δεδομένα.

Το μοντέλο είναι πολύ αποδοτικό, φτάνοντας τον μέσο όρο των 5 τμημάτων σε απόδοση κοντά στο 98.3% στο F1-score.

Επιλογή Χαρακτηριστικών

Στη παρακάτω εικόνα βλέπουμε τα χαρακτηριστικά με τη μεγαλύτερη αξία για τα μοντέλα. Το μοντέλο αυτό χρησιμοποιεί αρκετές παραμέτρους για την εκπαίδευση του μοντέλου.

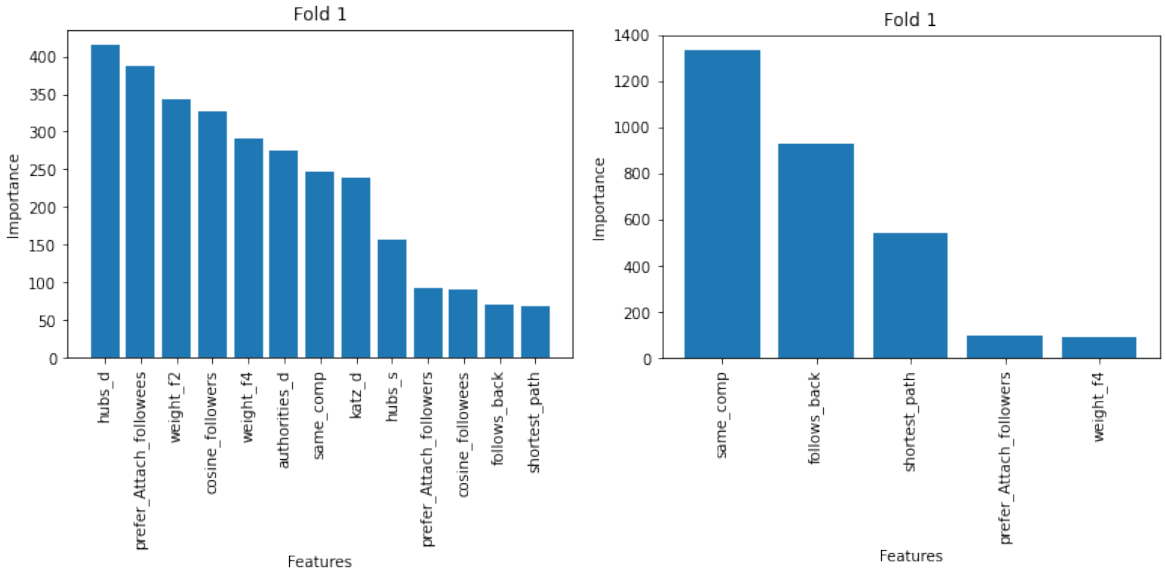


Σχήμα 5.20: Σημαντικότητα χαρακτηριστικών LightGBM με 30 χαρακτηριστικά

Μετά από αρκετές δοκιμές ωστόσο, όπως είδαμε και στους άλλους αλγορίθμους αλλάζει η αξία του κάθε χαρακτηριστικού. Στη συνέχεια θα δούμε τα χαρακτηριστικά με τη μεγαλύτερη σημαντικότητα όταν είναι 13 και 5.

Επίδοση με 13 και 5 χαρακτηριστικά

Παρακάτω φαίνεται η σμαντικότητα των χαρακτηριστικών με 13 και 5 διαθέσιμα χαρακτηριστικά στο μοντέλο LightGBM.



Σχήμα 5.21: Σημαντικότητα LightGBM με 13 και 5 χαρακτηριστικά αντίστοιχα

Το μοντέλο με τα 5 χαρακτηριστικά έχει απόδοση 96.9% στο F1-score. Η απόδοση που απέχει 2% διαφορά από εκείνη του μοντέλου που χρησιμοποιούσε το σύνολο των δεδομένων.

Στο μοντέλο με τα 13 χαρακτηριστικά έχει απόδοση κοντά στο 98.3% στο F1-score. Η απόδοση που απέχει 0.2% διαφορά από εκείνη του μοντέλου που χρησιμοποιούσε το σύνολο των δεδομένων.

5.5 Σύσταση

Μετά από την εκπαίδευση των μοντέλων μας σειρά έχει η σύσταση. Η σύσταση δε θα γίνει ελέγχοντας τον πιο πιθανό κόμβο προς ακολούθηση, από όλο το γράφημα για κάθε χρήστη. Αντιθέτως, θα γίνει συλλογή όλων των λογαριασμών με τους οποίους ο χρήστης που εξετάζεται έχει τουλάχιστον έναν κοινό ακόλουθο, αλλά ο ίδιος δεν ακολουθεί. Με αυτό τον τρόπο θα εξοικονομηθεί σημαντικός χρόνος υπολογισμού και σύστασης στο σύστημα. Για τη διαδικασία της συλλογής των χρηστών γίνεται χρήση της συνάρτησης ego_graph της βιβλιοθήκης networkx, που μας επιστρέφει όλους τους κόμβους από ένα αρχικό κόμβο, εντός της απόστασης που θα επιλέξουμε.

Στη συνέχεια, γίνεται επιλογή του αριθμού των χαρακτηριστικών που θεωρούμε καλύτερο για το δίκτυο μας να γίνει η σύσταση. Η σύσταση και με τα τριάντα χαρακτηριστικά αν και είναι η καλύτερη σε απόδοση είναι αρκετά αργή και δεν θεωρείται βιώσιμη για ένα δίκτυο με δεκάδες εκατομμύρια κόμβους και ακμές. Για αυτό τον λόγο θα γίνει η σύσταση με δεκατρία και πέντε χαρακτηριστικά για να γίνει σύγκριση των προτάσεών τους.

Για την σύσταση θα χρησιμοποιήσουμε τη συνάρτηση predict_proba των βιβλιοθηκών sklearn και keras για τα μοντέλα μας. Θα υπολογίσουμε έτσι τη προτεραιότητα βάσει της οποίας θα γίνουν οι συστάσεις που θα κάνουμε. Ακόμα θα συγκρίνουμε τα αποτελέσματα των αλγορίθμων μας για να δούμε την ομοιότητα των αποτελεσμάτων τους.

Για να συγκρίνουμε τα μοντέλα μας, επιλέγεται να γίνει σύσταση σε έναν από τους λογαριασμούς με τους περισσότερους ακολούθους στο δίκτυο, για ένα μεγαλύτερο δείγμα κοινών φίλων. Στο σύνολο οι χρήστες που θα προβλεφθεί η πιθανότητα τους να συσταθούν είναι 11.500. Ο αριθμός αυτός προκύπτει συλλέγοντας όλους τους λογαριασμούς με απόσταση 2 κόμβους από τον χρήστη, δηλαδή όλους τους χρήστες με τους οποίους υπάρχει τουλάχιστον ένας κοινός φίλος.

Οι αλγόριθμοι που επιλέχθηκαν είναι ο LightGBM και ο XGBoost καθώς σε όλες τους τις μετρικές είχαν σαφώς καλύτερα αποτελέσματα από τους υπόλοιπους αλγορίθμους. Στο σχήμα 5.22, στη στήλη xgb_1 βλέπουμε τις πιθανότητες με 13 χαρακτηριστικά του XGBoost και στη στήλη xgb_2 με τα 5, ενώ στη στήλη lgb_1 και lgb_2 είναι οι πιθανότητες του LightGBM αντίστοιχα. Στην αρχή των γραμμών ο αριθμός που υπάρχει συμβολίζει τον χρήστη για τον οποίο γίνεται η σύσταση.

Από τους αρχικούς κόμβους που έγινε η πρόβλεψη της σύστασης, περίπου 150 φάνηκαν να έχουν πιθανότητα μεγαλύτερη του 50% και έτσι να συστήνονται στον χρήστη για ακολού-

	xgb_1	xgb_2	lgb_1	lgb_2
5434	0.971792	0.999161	0.386156	0.441912
9662	0.921921	0.996628	0.871827	0.796906
9946	0.916942	0.998431	0.871827	0.825531
4974	0.907467	0.998491	0.833708	0.856817
9765	0.907467	0.998491	0.833708	0.856817
10122	0.904684	0.996628	0.871827	0.796906
10802	0.903406	0.995235	0.833708	0.796906
171	0.903406	0.995947	0.833708	0.314079
3877	0.903406	0.995947	0.833708	0.796906
7210	0.903406	0.995947	0.833708	0.796906
10296	0.903406	0.996628	0.833708	0.744804
3407	0.894853	0.969278	0.833708	0.796906

Σχήμα 5.22: Συστάσεις LightGBM και XGBoost με 5 και 13 χαρακτηριστικά

θηση. Στο σχήμα 5.22, βλέπουμε μερικές από τις συστάσεις των αλγορίθμων XGBoost και LightGBM με τη χρήση 5 και 13 χαρακτηριστικών. Οι χρήστες που συστήνονται φαίνονται στον δείκτη του συνόλου δεδομένων ως αριθμητικές τιμές. Οι συστάσεις των αλγορίθμων αυτών αν και έχουν κατά κύριο λόγο κοινή άποψη για τη σύσταση ή όχι του κάθε λογαριασμού, διαφέρουν στη κατάταξη των χρηστών που συστήνουν.

Στο σχήμα 5.23 φαίνεται πως τα πιο σημαντικά χαρακτηριστικά που χρησιμοποιήθηκαν για την εκπαίδευση έχουν ίδιες τιμές για τους λογαριασμούς που συστήνουν και λίγα από αυτά διαφορετικές.

	same_comp	weight_f4	hubs_d	prefer_Attach_followees	prefer_Attach_followers
5434	1	0.785373	1.387399e-16	652	391
9662	1	0.785373	-0.000000e+00	0	391
9946	1	0.785373	-0.000000e+00	0	391
4974	1	0.785373	-0.000000e+00	0	391
9765	1	0.785373	-0.000000e+00	0	391
10122	1	0.785373	-0.000000e+00	0	391
10802	1	0.785373	-0.000000e+00	0	391
171	1	0.785373	-0.000000e+00	0	391
3877	1	0.785373	-0.000000e+00	0	391
7210	1	0.785373	-0.000000e+00	0	391
10296	1	0.785373	-0.000000e+00	0	391
3407	1	0.785373	-0.000000e+00	0	391

Σχήμα 5.23: Τιμές των πιο σημαντικών χαρακτηριστικών

Το χαρακτηριστικό same_comp που καθορίζει αν ο χρήστης βρίσκεται στο ίδιο WCC μένει σταθερό πάντα στη τιμή 1. Όμοια, τα weight_f4, Prefer_Attach_followers που παίρνουν τιμές από 0.127592 ως 0.785373 και 391 ως 160310 αντίστοιχα, έχουν ίδιες τιμές. Τα χαρακτηριστικά hubs_d και Prefer_Attach_followees ωστόσο, στα οποία το μοντέλο LightGBM δίνει παραπάνω βαρύτητα έχουν διαφορετικές τιμές. Αυτός είναι και ο λόγος που στο σχήμα 5.22 ο αλγόριθμος XGBoost προτείνει τον λογαριασμό 5434 και ο LightGBM όχι.

Η κατάταξη των συστάσεων βασίζεται επίσης, στη βαρύτητα που δίνουν τα μοντέλα στο κάθε χαρακτηριστικό. Ο αλγόριθμος XGBoost χρησιμοποιεί σε μεγάλο βαθμό το χαρακτηριστικό Preferential Attachement για followers ενώ ο LightGBM τα same component και Preferential Attachement για followees. Πλέον, αποτελεί επιλογή του σχεδιαστή του συστήματος σύστασης ποια παράμετρο θεωρεί πιο σημαντική για το μέσο κοινωνικής δικτύωσης του.

Κεφάλαιο 6

Συμπεράσματα

6.1 Εισαγωγή

Έπειτα από την εξαγωγή των χαρακτηριστικών από τον αρχικό γράφημα, την εκπαίδευση των μοντέλων αλλά και το στάδιο της σύστασης προέκυψαν μερικά συμπεράσματα για τα στάδια της εκπαίδευσης, τα χαρακτηριστικά που είχαν τη μεγαλύτερη επίδραση καθώς και τα μοντέλα και την απόδοσή τους. Επίσης, γίνεται μια προσέγγιση των μελλοντικών επεκτάσεων που θα μπορούσαν να γίνουν στην εργασία και των δεδομένων που θα έκαναν τις συστάσεις του συστήματος πιο ακριβείς.

6.2 Σύνοψη των μοντέλων Random Forest

Ο αλγόριθμος Random Forest που χρησιμοποίηθηκε ανήκει στη βιβλιοθήκη sklearn της python, όπου δίνεται η έδινε τη δυνατότητα να αλλάξουμε τις παραμέτρους του μοντέλου. Χρησιμοποιώντας αυτούς τους αλγόριθμους παρατηρήσαμε σημαντική αποτελεσματικότητα στην απόδοση των διάφορων μοντέλων που εξετάστηκαν.Οι παράμετροι που διαφοροποιήθηκαν για την εκπαίδευση των μοντέλων ωστόσο, δεν επηρέασαν εμφανώς τα αποτελέσματα που είχε ο αλγόριθμος. Ακόμα, ο αλγόριθμος δεν ήταν ιδιαίτερα χρονοβόρος στην εκπαίδευση, αλλά ούτε και στη σύστασή του.

Έπειτα, ελέγχθηκε η απόδοση του αλγόριθμου με χρήση διαφορετικού αριθμού χαρακτηριστικών. Το μοντέλο απέδιδε με μικρή μείωση της ακρίβειας του με 5 χαρακτηριστικά αντί για 30 που ήταν τα αρχικά. Ωστόσο, η χρήση των 13 χαρακτηριστικών δεν έδειξε σημαντική βελτίωση.

Έτσι, συμπεραίνουμε πως η χρήση αυτού του αλγόριθμου θα προτεινόταν για χρήση με λίγα χαρακτηριστικά, όπου αποδίδει ικανοποιητικά.

Το μοντέλο έδειξε επίσης να χρησιμοποιεί σε μεγάλο βαθμό για τις προβλέψεις του το χαρακτηριστικό Preferential attachment. Αυτό μας δείχνει πως βασίζεται στη λογική που αναφέρει πως όταν δυο χρήστες έχουν αρκετούς ακολούθους είναι πιο πιθανό να υπάρχει ένωση μεταξύ τους. Ακόμα, δίνει έμφαση στα χαρακτηριστικά shortest path και same component, που δείχνουν την ελάχιστη απόσταση του follower από τον followee καθώς και το αν ανήκουν στον ίδιο κύκλο ακολούθων.

6.3 Σύνοψη του μοντέλου XGBoost

Στα μοντέλα με boosting αλγορίθμους παρατηρήθηκε η καλύτερη απόδοση σε σχέση με όλα τα μοντέλα. Συγκεκριμένα ο XGBoost έδειξε σταθερότητα στις αποδόσεις του σε όλα τα τμήματα των δεδομένων που έγιναν για εκπαίδευση και τον έλεγχο της απόδοσης. Ο αλγόριθμος ωστόσο, αποδείχθηκε σχετικά χρονοβόρος στην εκπαίδευση.

Έπειτα, ελέγχθηκε η απόδοση του αλγόριθμου με χρήση διαφορετικού αριθμού χαρακτηριστικών. Η μείωση της απόδοσης στα 13 χαρακτηριστικά ήταν της τάξης του 0.2%. Επίσης, με 5 χαρακτηριστικά η απόδοση του μειώθηκε στο 96,8%, ενώ ήταν και πάλι ίσως ο αποδοτικότερος από τους αλγορίθμους με χρήση 5 χαρακτηριστικών.

Συνολικά, συμπεραίνουμε πως ο αλγόριθμος θα συστηνόταν σε συστήματα που έχουν μεγάλη υπολογιστική δύναμη για να μπορέσουν να διαχειριστούν γραφήματα με πολλές ακμές και κόμβους.

Το μοντέλο έδειξε επίσης να χρησιμοποιεί σε μεγάλο βαθμό για τις προβλέψεις του με τα πέντε χαρακτηριστικά αυτό του Preferential attachment. Αυτό μας δείχνει,όπως και στον Random Forest αλγόριθμο, πως βασίζεται στη λογική που αναφέρει πως όταν δυο χρήστες έχουν αρκετούς ακολούθους είναι πιο πιθανό να υπάρχει ένωση μεταξύ τους. Ακόμα, εκμεταλλεύεται σε μικρότερο βαθμό τα χαρακτηριστικά weight version 4 και shortest path, που χρησιμοποιούν το βάρος που είναι αντιστρόφως ανάλογο των ακολούθων των δύο χρηστών καθώς και την ελάχιστη απόσταση του follower από τον followee.

Με τη χρήση των δεκατριών χαρακτηριστικών, ο αλγόριθμος βασίζεται στο γνώρισμα της ομοιότητας συνημιτόνου των followees, στην ομοιότητα δηλαδή των χρηστών που ακολουθούν οι δυο λογαριασμοί που θέλουμε να προβλέψουμε. Ακόμα, εκμεταλλεύεται σε μι-

κρότερο βαθμό τα χαρακτηριστικά shortest path και follow back, που δείχνουν την ελάχιστη απόσταση του follower από τον followee καθώς και εάν υπάρχει δηλαδή και η αντίστροφη σύνδεση μεταξύ τους.

6.4 Σύνοψη του μοντέλου LightGBM

Όπως αναφέρθηκε και στη προηγούμενη παράγραφο, τα μοντέλα με boosting αλγορίθμους έδειξαν τη καλύτερη απόδοση. Αυτό ισχύει και για τον αλγόριθμο LightGBM της Microsoft, ο οποίος είχε παρόμοια απόδοση με τον XGBoost. Το προτέρημα του ωστόσο είναι η ταχύτητα του, έχοντας σχεδιαστεί με αυτόν το σκοπό εξαρχής.

Όμοια με τον XGBoost, οι αποδόσεις του LightGBM ήταν αντίστοιχα καλές, με πολύ μικρή πτώση απόδοσης στα 13 χαρακτηριστικά και όμοιες αποδόσεις στα 5.

Συνοψίζοντας, ο αλγόριθμος θα εκπλήρωνε τις ανάγκες της πλειονότητας των μέσων κοινωνικής δικτύωσης, ανεξάρτητα με το είδος του γραφήματος που θα διαχειριζόταν.

Το μοντέλο έδειξε επίσης να χρησιμοποιεί σε μεγάλο βαθμό για τις προβλέψεις του με τα πέντε χαρακτηριστικά το γνώρισμα του same component. Αυτό μας δείχνει, αν οι δύο χρήστες ανήκουν στον ίδιο κύκλο ακολούθων. Επίσης, εκμεταλλεύεται τα χαρακτηριστικά follow back και shortest path, που δείχνουν όπως προαναφέραμε, εάν ο followee ακολουθεί τον follower ήδη και την ελάχιστη απόσταση του follower από τον followee.

Όταν χρησιμοποιούνται 13 χαρακτηριστικά ο αλγόριθμος βασίζεται σε αρκετά από αυτά με πρώτο το γνώρισμα που προκύπτει από τη συνάρτηση hubs για τον followee. Το χαρακτηριστικό αυτό βασίζεται στην ιδιότητα του followee να ακολουθεί πολλούς χρήστες. Ακόμα, εκμεταλλεύεται και τα χαρακτηριστικά Preferential Attachment για τους followees, weight version 2, ομοιότητα συνημιτόνου των followers, και άλλα που αναφέρθηκαν προηγουμένως.

Μια σημαντική διαφορά που έχει ο αλγόριθμος LightGBM είναι η χρήση σε μεγάλο βαθμό όλων των χαρακτηριστικών που δίνονται στην εκπαίδευση του μοντέλου. Η θετική σκοπιά σε αυτό είναι μια πιο ολοκληρωμένη σύσταση, που δίνει έμφαση σε πολλά χαρακτηριστικά μαζί. Αντίθετα, μια άλλη οπτική θα ήταν πως η χρήση λιγότερων και καλύτερων χαρακτηριστικών θα βελτίωνε τις συστάσεις του αλγορίθμου. Έτσι, εξαρτάται από τον διαχειριστή του συστήματος σύστασης να καθορίσει αν η χρήση αρκετών χαρακτηριστικών είναι καλή για την εκάστοτε πλατφόρμα κοινωνικού δικτύου.

6.5 Σύνοψη των Νευρωνικών Δικτύων

Το μοντέλο που χρησιμοποιήθηκε το περιέχει η βιβλιοθήκη keras της python, και δίνει τη δυνατότητα δημιουργίας νέων νευρωνικών δικτύων. Στα μοντέλα αυτά παρατηρήθηκε η μεγαλύτερη αποτελεσματικότητα στην απόδοση των διαφορετικών μοντέλων όταν ήταν διαθέσιμα τα περισσότερα χαρακτηριστικά, ενώ η απόδοση του μοντέλου μειωνόταν εμφανώς. Επίσης, το νευρωνικό δίκτυο ήταν ιδιαίτερα χρονοβόρο στην εκπαίδευση του.

Έπειτα, ελέγχθηκε η απόδοση του αλγόριθμου με χρήση διαφορετικού αριθμού χαρακτηριστικών. Το μοντέλο, με τη χρήση δεκατριών χαρακτηριστικών είχε παρόμοια αποτελέσματα με τα μοντέλα Random Forest. Ωστόσο, με τη χρήση πέντε χαρακτηριστικών η απόδοση μειώθηκε εμφανώς, παραμένοντας βέβαια σε απόδοση πάνω από τη τιμή του 95%.

Συμπεραίνουμε λοιπόν, πως η χρήση αυτού του αλγόριθμου θα προτεινόταν για χρήση με τα περισσότερα δυνατά χαρακτηριστικά, αν και οι boosting αλγόριθμοι έχουν ακόμα καλύτερα αποτελέσματα.

Το μοντέλο έδειξε επίσης να χρησιμοποιεί σε μεγάλο βαθμό για τις προβλέψεις του το χαρακτηριστικό follow back. Αυτό μας δείχνει πως βασίζεται κυρίως στο αν ο followee ακολουθεί ήδη τον follower. Ακόμα, δίνει έμφαση στα χαρακτηριστικά shortest path, cosine similarity για τους followers και Preferential Attachment για τους followers, που έχουμε αναλύσει παραπάνω.

6.6 Μελλοντικές επεκτάσεις

Σε μελλοντικές επεκτάσεις της διπλωματικής θα μπορούσαν να υπάρχουν μερικές ακόμα προσθήκες που θα έκαναν τις συστάσεις του συστήματος πιο αποδοτικές.

Αρχικά, θα ήταν πολύ χρήσιμη η διαθεσιμότητα και διαφορετικού είδους πληροφορίας. Αν και το δίκτυο μεταξύ των ακολούθων παρέχει πολύ αποδοτικά αποτελέσματα, η διαδικασία της σύστασης θα μπορούσε να ενισχυθεί και με άλλες πληροφορίες. Αυτές θα μπορούσαν να είναι δεδομένα όπως η τοποθεσία που βρίσκεται ο χρήστης σε διάφορες χρονικές στιγμές για καλύτερη προσέγγιση των ανθρώπων που μπορεί να έχει συναναστραφεί. Ακόμα, η δραστηριότητα των λογαριασμών στη πλατφόρμα θα μας έδειχνε με ποια είδη λογαριασμών έχει περισσότερες αλληλεπιδράσεις.

Επιπρόσθετα, θα ήταν πολύ ενδιαφέρον να υπήρχε πληροφορία για στιγμιότυπα του δικτύου σε διαφορετικές χρονικές στιγμές και να γίνει η διαχείριση του ως ένα χρονικό δίκτυο

(temporal network). Έτσι θα γινόταν πιο εμφανές ποιους χρήστες ακολουθεί και συναναστρέφεται ο κάθε λογαριασμός ανά χρονικές περιόδους, με σκοπό τις πιο καίριες προβλέψεις.

Βιβλιογραφία

[1] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[2] Facebook recruiting competition dataset. `https://www.kaggle.com/competitions/FacebookRecruiting/data`.

[3] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[4] William Cukierski, Benjamin Hamner, and Bo Yang. Graph-based features for supervised link prediction. In *The 2011 International joint conference on neural networks*, pages 1237–1244. IEEE, 2011.

[5] Jiawei Zhang and S Yu Philip. Link prediction across heterogeneous social networks: A survey. *Social networks*, 2014.

[6] Antonio Pecli, Bruno Giovanini, Carla C Pacheco, Carlos Moreira, Fernando Ferreira, Frederico Tosta, Júlio Tesolin, Marcio Vinicius Dias, Silas P Lima Filho, Maria Cláudia Cavalcanti, et al. Dimensionality reduction for supervised learning in link prediction problems. In *ICEIS (1)*, pages 295–302, 2015.

[7] Matthew Partridge and Rafael A Calvo. Fast dimensionality reduction and simple pca. *Intelligent data analysis*, 2(3):203–214, 1998.

[8] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael Jordan. Nonparametric link prediction in dynamic networks. *arXiv preprint arXiv:1206.6394*, 2012.

[9] Zhiqiang Teng, Shuai Teng, Jiqiao Zhang, Gongfa Chen, and Fangsen Cui. Structural damage detection based on real-time vibration signal and convolutional neural network. *Applied Sciences*, 10(14):4720, 2020.

[10] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020.

[11] Weakly connected components wolphram alpha. `https://mathworld.wolfram.com/WeaklyConnectedComponent.html`.

[12] Weakly connected components geeks 4 geeks. `https://www.geeksforgeeks.org/find-weakly-connected-components-in-a-directed-graph/`.

[13] Punit Patel and Kanu Patel. A review of pagerank and hits algorithms. *Int J Adv Res Eng Sci Technol*, pages 2394–2444, 2015.

[14] Pagerank wikipedia. `https://en.wikipedia.org/wiki/PageRank`.

[15] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45(3):751–782, 2015.

[16] Allan H Murphy. The finley affair: A signal event in the history of forecast verification. *Weather and forecasting*, 11(1):3–20, 1996.

[17] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[18] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

[19] Xgboost documentation. `https://xgboost.readthedocs.io/en/stable/`.

[20] Welcome to lightgbm's documentation! `https://lightgbm.readthedocs.io/en/v3.3.2/`.

[21] Carter Chiu and Justin Zhan. Deep learning for link prediction in dynamic networks using weak estimators. *IEEE Access*, 6:35937–35945, 2018.

[22] Linear vs nonlinear classification image. `http://ieeeucsd.org/mlbootcamp/images/nonlinear.png`.