ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές μηχανικής μάθησης για την εξόρυξη δεδομένων και την πρόβλεψη επιτυχίας νεοφυών επιχειρήσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παπακωνσταντίνου Ε. Παναγιώτης

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Τεχνικές μηχανικής μάθησης για την εξόρυξη δεδομένων και την πρόβλεψη επιτυχίας νεοφυών επιχειρήσεων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παπακωνσταντίνου Ε. Παναγιώτης

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 1η Ιουλίου 2021.

.....	.....	.....
Σταφυλοπάτης Ανδρέας-Γεώργιος	Κόλλιας Στέφανος	Στάμου Γεώργιος
Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Αναπληρωτής Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2021

.....
Παναγιώτης Ε Παπακωνσταντίνου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ταχύτατη εξέλιξη είναι η φράση που χαρακτηρίζει τη σύγχρονη κοινωνία και ιδιαίτερα τη σημερινή επιχειρηματική αγορά. Η εξάπλωση της τεχνολογίας, το συνεχώς υψηλότερο επίπεδο γνώσεων και ο ανταγωνισμός μεταβάλλουν ραγδαία τον επιχειρηματικό κόσμο. Συνεχώς προκύπτουν ευκαιρίες με υπέρογκες ανταμοιβές, αλλά ταυτόχρονα απαιτούν υψηλό ρίσκο. Οι νεοφυείς επιχειρήσεις (start-ups), τα τελευταία χρόνια, αποτελούν πόλο έλξης για μεγάλο αριθμό επιχειρηματιών, αφού σε ελάχιστο χρονικό διάστημα αποφέρουν τεράστια κέρδη.

Στόχος της παρούσας εργασίας αποτελεί η μελέτη των παραγόντων που οδηγούν μια εταιρία στην επιτυχία, η αναγνώριση των κοινών μοτίβων, ώστε αυτή η γνώση να αξιοποιηθεί μελλοντικά από επενδυτές και επιχειρηματίες. Η πρόβλεψη επιτυχίας μιας εταιρίας θεωρείται ένα απαιτητικό εγχείρημα, το οποίο μπορεί να αποφέρει ένα σημαντικό πλεονέκτημα στους εμπλεκόμενους.

Ως σήμερα, το πρόβλημα μελετάται σε θεωρητικό επίπεδο και έμφαση δίνεται σε επιχειρηματικά μοντέλα, δομές και οικονομικά στοιχεία. Ωστόσο, η επιτυχία ή αποτυχία μιας εταιρίας έχει πλήθος παραγόντων που συχνά είναι δύσκολο να συνδυαστούν.

Στην παρούσα εργασία, προσεγγίζεται η παραπάνω πρόκληση εφαρμόζοντας τεχνικές μηχανικής μάθησης σε αυστηρά μετρήσιμους παράγοντες. Συγκεκριμένα, αναλύονται και επεξεργάζονται τα δεδομένα 55,585 εταιριών παγκοσμίως από διαφορετικούς τομείς και στη συνέχεια δοκιμάζονται σε 6 αλγορίθμους. Μεγάλη έμφαση δίνεται στον τομέα δραστηριοποίησης και τη χώρα ίδρυσης, στην τακτική χρηματοδότησης και την εμπειρία των επενδυτών. Ως επιτυχημένη εταιρία θεωρήθηκε αυτή που προσφέρει μεγάλα χρηματικά ποσά στους ιδρυτές, τους επενδυτές και τους πρώτους υπαλλήλους (i) με την εισαγωγή της στο χρηματιστήριο ή (ii) με την εξαγορά της από άλλη εταιρία. Οπότε, οι εταιρίες χωρίστηκαν σε δύο κατηγορίες σύμφωνα με το προηγούμενο κριτήριο, ενώ επιλέχθηκαν προσεκτικά χαρακτηριστικά που θα προσφέρουν πληροφορία προκειμένου να διευκολύνουν τους αλγορίθμους να αξιολογούν σωστά και να εντοπίσουν τα απαιτούμενα μοτίβα.

Αρχικά, περιγράφεται το θεωρητικό υπόβαθρο των start-ups και παρουσιάζονται τα αποτελέσματα παλαιότερων σχετικών μελετών. Ακολουθεί η ανάλυση και προσεκτική επεξεργασία των δεδομένων, ενώ επιλέγονται τα κατάλληλα χαρακτηριστικά ώστε να εμπλουτιστούν τα σύνολα. Στη συνέχεια, τα σύνολα χαρακτηριστικών αξιοποιούνται από τους αλγορίθμους, σχολιάζονται τα αποτελέσματα για την εξαγωγή χρήσιμων συμπερασμάτων και την ενίσχυση της επιστημονικής κοινότητας με μελλοντικές προτάσεις. Ταυτόχρονα, σημειώνονται τα προβλήματα και οι μέθοδοι με τις οποίες αντιμετωπίστηκαν.

Λέξεις κλειδιά:

startups, προβλέψεις, επιτυχημένη εταιρία, επιστήμη δεδομένων, μηχανική μάθηση, ανάλυση δεδομένων, εισαγωγή στο χρηματιστήριο, εξαγορά ή συγχώνευση, XGBoost

Abstract

Rapid evolution is the phrase that characterizes modern society and especially today's business market. The spread of technology, the ever-higher level of knowledge and competition are rapidly changing the business world. Opportunities are constantly emerging with tremendous rewards, but at the same time they require high risk taking. Start-ups, in recent years, are a pole of attraction for a large number of entrepreneurs, since it is possible, they make huge profits in a short period of time.

The aim of the present work is to study the factors that lead a company to success and recognize the common patterns, so that this knowledge can be used in the future by investors and entrepreneurs. Predicting a company's success is considered a demanding undertaking, which can bring a significant advantage to those involved.

To date, the problem is studied at a theoretical level and emphasis is placed on business models, structures and financial data. However, the success or failure of a company has a number of factors that is difficult to be combined.

In the present work, the above challenge is approached by applying machine learning techniques to strictly measurable factors. Specifically, the data of 55,585 companies worldwide from different sectors are analyzed and processed and then tested in 6 algorithms. Great emphasis is given to the field of activity and the country of establishment, to the financing tactics and the investors' experience. A successful company was considered to be the one that offers large sums of money to the founders, investors and first employees (i) by initial public offering (IPO) or (ii) by acquiring it from another company (M&A). So, the companies were divided into two categories according to the previous criterion, while features were carefully selected to provide information in order to facilitate the algorithms to evaluate correctly and to identify the required patterns.

First, the theoretical background of start-ups is described and the results of previous relevant studies are presented. This is followed by the analysis and careful processing of the data, while the appropriate features are selected to enrich the sets. Then, the feature sets are utilized by the algorithms, the results are commented on to draw useful conclusions and to strengthen the scientific community with future proposals. At the same time, the problems and the methods by which they were dealt with are noted.

Key words:

startups, predictions, successful company, data science, machine learning, data analysis, initial public offering (IPO), mergers or acquisitions (M&A), XGBoost

Ευχαριστίες

Η παρούσα εργασία πραγματοποιήθηκε στο πλαίσιο του προπτυχιακού κύκλου σπουδών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών το έτος 2021. Αποτελεί το τελευταίο βήμα για την απόκτηση του διπλώματός μου και αισθάνομαι την ανάγκη να ευχαριστήσω όλους αυτούς που συνέβαλαν στην πορεία μου μέχρι εδώ.

Πρώτα από όλα, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της εργασίας μου, κύριο Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή ΕΜΠ, που μου επέτρεψε να ασχοληθώ με το συγκεκριμένο θέμα και να διευρύνω τους ορίζοντές μου στο αντικείμενο. Ακόμη, θα ήθελα να ευχαριστήσω τον κ. Αθανάσιο Τασάκο, υποψήφιο διδάκτορα του ΕΜΠ, που με καθοδήγησε ορθά κατά τη διάρκεια εκπόνησης της εργασίας, ώστε να προσεγγίσω το αντικείμενο από όλες τις οπτικές.

Ένα μεγάλο ευχαριστώ οφείλω σε όλους τους φίλους και συμφοιτητές μου που ήταν δίπλα μου όλα αυτά τα χρόνια και συνεργαστήκαμε για εργασίες και διαβάσαμε μαζί για εξεταστικές.

Το μεγαλύτερο «ευχαριστώ» στους γονείς μου, που με στήριξαν όλα τα χρόνια και συμπαραστάθηκαν στις δυσκολίες μου συμβάλλοντας καθοριστικά στην επιτυχία μου.

Παναγιώτης Ε Παπακωνσταντίνου,

Αθήνα, Ιούνιος 2021

Πίνακας περιεχομένων

Κεφάλαιο 1: Εισαγωγή

1.1 Εισαγωγή

Τα τελευταία χρόνια, οι νεοφυείς επιχειρήσεις start-ups ανθίζουν σε κάθε επιχειρηματικό τομέα αφού περισσότερα Πανεπιστήμια, κυβερνήσεις και ιδιωτικές εταιρίες επενδύουν και ενθαρρύνουν τους ανθρώπους να τολμήσουν να ασχοληθούν με νέες ιδέες. Οι εταιρίες κερδίζουν εκατομμύρια μόλις αναγνωριστούν ως unicorn (αξιολόγηση της start-up με αξία άνω του ενός εκατομμυρίου δολαρίων), γεγονός που επιτυγχάνεται σε σύντομο χρονικό διάστημα, αφού κατά μέσο όρο 6 χρόνια χρειάζεται μια νεοφυής επιχείρηση για να το καταφέρει. Μια εφαρμογή μηνυμάτων, η Slack, το κατάφερε σε λιγότερο από ενάμιση χρόνια λειτουργίας. Ταυτόχρονα, η κοινωνία επηρεάζεται σημαντικά από τεχνολογικούς κολοσσούς και καινοτόμες ιδέες. Παραδείγματα όπως η Facebook και η Airbnb αλλάζουν τις συνήθειες και τον τρόπο διασκέδασης των ανθρώπων. Οι νεοφυείς επιχειρήσεις έχουν τέτοια απήχηση που αποτελεί διακαής πόθος κάθε επενδυτή να είναι μέλος μεγάλων εξαγορών, όπως για παράδειγμα η εξαγορά της WhatsApp από την Facebook όπου επέφερε στην Sequoia (εταιρία επενδυτικού κεφαλαίου) την επένδυσή της επί 50. Ωστόσο, οι start-ups υπολογίζεται πως έχουν πολύ υψηλή πιθανότητα αποτυχίας, το οποίο σημαίνει πως πολλές επενδύσεις αποτυγχάνουν. Σύμφωνα με το Fortune.com, 9 στις 10 αποτυγχάνουν, ενώ το Statisticbrain.com ισχυρίζεται πως μόνο το 50% των παραδοσιακών επιχειρήσεων επιβιώνει στη βιομηχανία μετά από 4 χρόνια.

Οι start-ups έχουν αναγνωριστεί παγκοσμίως ως καθοριστικός παράγοντας για τις εθνικές οικονομίες. Συμβάλλουν στην ανάπτυξη της τεχνολογίας, αφού προσφέρουν σημαντικά ποσά για έρευνα και ανάπτυξη, βελτιώνουν την ποιότητα ζωής των πολιτών, ενώ παράλληλα δημιουργούν και θέσεις εργασίας. Κάθε χρόνο δημιουργούνται περίπου 305 εκατομμύρια εταιρίες, με τα 1.35 εκατομμύρια να σχετίζονται με την τεχνολογία. Το 2019 έγιναν επενδύσεις 295 δισεκατομμυρίων δολαρίων παγκοσμίως, ενώ το 2020 μόνο στη βόρεια Αμερική επενδύθηκαν 133 δισεκατομμύρια δολάρια (CrunchBase.com). Γίνεται σαφές πως αρκετοί ασχολούνται με τη δημιουργία νεοφυών επιχειρήσεων.

Την εξήγηση αυτού του φαινομένου προσπαθεί να εξηγήσει ο Steve Blank, μετά από 25 χρόνια ως επιχειρηματίας σε τεχνολογικές επιχειρήσεις, θέτοντας τέσσερα επιχειρήματα στο βιβλίο του “The Four Steps to the Epiphany” (Blank, 2020):

-Οι start-ups πλέον απαιτούν για τη δημιουργία τους χιλιάδες αντί για εκατομμύρια δολάρια: Με τη μείωση του κόστους παραγωγής των προϊόντων είναι φθηνότερο από ποτέ να δημιουργηθεί τεχνολογία. Η πρόσβαση σε εργαλεία, ανοικτό κώδικα, φθηνότερους servers, ενώ ταυτόχρονα με την επέκταση της κοινότητας των προγραμματιστών που συνεισφέρουν στη διάδοση της τεχνολογίας παγκοσμίως επιτρέπεται στον καθένα να σχεδιάσει, να δοκιμάσει και να μοιραστεί προϊόντα.

-Η αποφασιστική βιομηχανία του επιχειρηματικού κεφαλαίου: Το επιχειρηματικό κεφάλαιο, παλαιότερα, απαιτούνταν να ξοδευτεί σε εκατομμύρια δολάρια, στοιχηματίζοντας λίγα, αλλά μεγάλα στοιχήματα. Ωστόσο, με το κόστος της τεχνολογίας να μειώνεται κάθε χρόνο δημιουργείται η ευκαιρία για άλλου τύπου επενδύσεις: πρώιμου σταδίου, επιτάχυνσης και μικρο-κεφάλαιο επιχειρηματικών συμμετοχών. Οι οργανισμοί, με μικρότερους λογαριασμούς μπορούν να

πραγματοποιήσουν πολλά περισσότερα μικρά στοιχήματα και να ενισχύσουν περισσότερες start-ups. Τα χρήματα αυτά αποτελούν σανίδα σωτηρίας για τις μικρές start-ups που μπορούν να μην αναζητούν επιπλέον χρηματοδότηση μέχρι τα αργότερα στάδια ανάπτυξής τους.

-Η επιχειρηματικότητα αναπτύσσει τη δική της επιστήμη διοίκησης: Στα τέλη του ’70, εφαρμόζονταν οι ιδέες που είχε προτείνει ο Henry Ford. Ωστόσο, μετά την φούσκα των dotcom στα τέλη του 20ου αιώνα, πολλοί επιχειρηματίες συνειδητοποίησαν τη διαφορετικότητα των start-ups. Το 2011, ο Eric Ries με το “The Lean Start-up” έθεσε τα θεμέλια για την ανάπτυξη νέας γνώσης στην διοικητική επιστήμη των start-ups, η οποία έγινε γνωστή με τον όρο the Lean Start-up Movement.

-Αυξήθηκε η ταχύτητα ενστερνισμού νέων τεχνολογιών: Καθώς το ιντερνέτ γίνεται ευρέως προσβάσιμο, οι start-ups μπορούν να είναι – από την πρώτη μέρα - αυτό που είπε ο Steve Blank, μια «μικρή πολυεθνική» και όλοι οι άνθρωποι να έχουν εύκολη πρόσβαση σε προϊόντα από την κάθε γωνιά του πλανήτη (Blank, 2020). Η Google and η Facebook απέδειξαν πως η τοποθεσία είναι ασήμαντη. Ακόμα και η ιδεολογία των μεγάλων επιχειρήσεων άλλαξε, αφού είναι πλέον πρόθυμες να δοκιμάσουν φθηνότερες, ταχύτερες και πιο εκλεπτυσμένες τεχνολογίες που παρέχονται από τις αναπτυσσόμενες start-ups. Για παράδειγμα, η Slack, η εταιρία που κατάφερε στο συντομότερο χρονικό διάστημα να αξίζει 1 δις δολάρια (μέσα σε 1.25 χρόνια έγινε unicorn) κατάφερε στα πρώτα τρία χρόνια λειτουργίας της να έχει πελάτες κολοσσούς όπως η Airbnb, η Buzzfeed, η eBay, η Expedia, η NASA και η Salesforce μέσα από το φθηνό λογισμικό και το εκλεπτυσμένο προϊόν.

Η ευκολία πρόσβασης σε παγκόσμιο επίπεδο από χρήστες και καταναλωτές, η αυξανόμενη ταχύτητα ενστερνισμού τεχνολογικών ιδεών από καταναλωτές και επιχειρήσεις, αλλά και η πρόσβαση σε ενημερωμένα δεδομένα και τεχνικές μάθησης έδωσε στους επιχειρηματίες πρόσβαση σε περισσότερη γνώση, ώστε να αποφεύγονται λάθη του παρελθόντος και να αξιολογούνται ορθά θεμελιώδη χαρακτηριστικά των επιχειρήσεων, με αποτέλεσμα να ενεργοποιηθούν οι start-ups και να αναπτύσσονται με σημαντικά γρηγορότερο ρυθμό. Η διαδικασία λήψης αποφάσεων μέσα από την αξιολόγηση δεδομένων, που υποστηρίζεται από τεχνικές μηχανικής μάθησης, μειώνει το ρίσκο για έναν επενδυτή, το οποίο τελικά σημαίνει περισσότερο κέρδος.

Εύλογα προκύπτει το ερώτημα γιατί μερικές start-ups αποτυγχάνουν, ενώ άλλες τα καταφέρνουν, του οποίου η απάντηση δεν είναι απλή. Πρέπει πρώτα να εξεταστεί η ειδοποιός διαφορά μεταξύ μιας start-up και μιας παραδοσιακής νέας επιχείρησης, ενώ ύστερα θα οριστεί η επιτυχία μιας επιχείρησης.

1.2 Ορισμός

Με τον όρο start-up γίνεται αναφορά σε μια εταιρία στα πρώτα στάδια λειτουργίας της. Έχει ιδρυθεί από έναν ή περισσότερους επιχειρηματίες που επιθυμούν να αναπτύξουν ένα προϊόν ή υπηρεσία που θεωρούν ότι θα έχει ζήτηση. Οι συγκεκριμένες επιχειρήσεις ξεκινούν με υψηλά κόστη και περιορισμένα κέρδη και αυτός είναι ο λόγος που αναζητούν κεφάλαιο από πολλές πηγές, όπως οι επενδυτές κεφαλαίου.

Οι εταιρίες που χαρακτηρίζονται start-ups επικεντρώνονται αποκλειστικά σε ένα προϊόν ή υπηρεσία το οποίο οι ιδρυτές του θέλουν να βγει στην αγορά. Συνήθως, δεν

έχουν ολοκληρωμένα ανεπτυγμένο ένα επιχειρηματικό μοντέλο και κυρίως δυσκολεύονται λόγω έλλειψης κεφαλαίου, με τις περισσότερες να έχουν αρχικά χρηματοδοτηθεί από τους ίδιους τους ιδρυτές (Investopedia).

Οι startups είναι αποτέλεσμα επιχειρηματικής δραστηριότητας. Σύμφωνα με τον Kirzner, οι επιχειρηματίες εκμεταλλεύονται ευκαιρίες που οι υπόλοιποι δεν βλέπουν (Kirzner, 2015). Για τον καθηγητή του Harvard, Stevenson η επιχειρηματικότητα είναι η επιδίωξη ευκαιριών ενώ δεν έχουν εξασφαλιστεί οι απαιτούμενοι πόροι (Stevenson et al., 2007). Και οι δύο ορισμοί βασίζονται στην εκμετάλλευση ευκαιριών εξισορροπητικής ισορροπίας, αλλά οι σύγχρονες startups απαιτούν κάτι περισσότερο από την αξιοποίηση ευκαιριών, να δημιουργούν ταυτόχρονα και ευκαιρίες. Συνεπώς, χρειάζεται ένας νέος διευρυμένος ορισμός. Ο Faltinand Ripsas πρότεινε το σχεδιασμό του επιχειρηματικού μοντέλου ως τον κορμό της επιχειρηματικής διαδικασίας. Οπότε, επιχειρηματικότητα θεωρείται η διαδικασία ανάπτυξης καινοτόμων επιχειρηματικών μοντέλων, που θα εξυπηρετούν πελάτες με νέα προϊόντα ή/και υπηρεσίες, με στόχο να αλλάξουν τον τρόπο που οι άνθρωποι ζουν και εργάζονται.

Οι start-ups είναι εταιρίες οι οποίες παράγουν προϊόντα και τολμούν να επιχειρούν σε τομείς και αγορές με πρωτοπόρο τρόπο. Είναι ριψοκίνδυνες και απρόβλεπτες, καθώς το νέο προϊόν ή υπηρεσία ενδέχεται να μην ικανοποιήσει το κοινό και να απαιτηθούν συνεχείς τροποποιήσεις πριν ενσωματωθεί στην αγορά. Ουσιαστικά, μια start-up είναι μια εταιρία υψηλού ρίσκου στα πρώτα στάδια λειτουργίας της και συχνά σχετίζεται με την τεχνολογία (Ries, 2011).

Ο συν-ιδρυτής της PayPal, Peter Thiel, ορίζει μια start-up ως το δημιουργό μιας κάθετης και όχι οριζόντιας καινοτομίας. Κάθετη καινοτομία θεωρείται η δημιουργία μιας νέας τεχνολογίας, ενώ οριζόντια καινοτομία είναι η διαδικασία της διεθνοποίησης, η χρησιμοποίηση ήδη υπάρχουσα τεχνολογίας σε τομείς που δεν έχει εφαρμοστεί ακόμη. Ο Thiel είναι ένθερμος υποστηρικτής της άποψης πως μια start-up πρέπει να στοχεύει να δημιουργήσει ένα μονοπώλιο σε μια αγορά και ύστερα να επεκταθεί σε άλλους τομείς (Thiel & Masters, 2014). Ο Thiel έδωσε έντονη έμφαση στα χαρακτηριστικά μιας τεχνολογικής παράτολμης επιχείρησης. Αυτό επιδεικνύεται από την οπτική του για εκθετικά αυξανόμενη ανάπτυξη και τοποθέτηση της επιχείρησης στην αγορά ως μονοπωλιακός ρυθμιστής σε πρώτο στάδιο. Παραδείγματα τέτοιων επιχειρήσεων θεωρούνται η Alibaba, το Facebook και το YouTube.

Ο Paul Graham, ιδρυτής της Y Combinator, το έθεσε πιο απλά στην έκθεση του: «μια start-up είναι μια εταιρία σχεδιασμένη να αναπτύσσεται γρήγορα». Σε αντίθεση με τον μεγιστάνα της τεχνολογίας, Peter Thiel, ο Graham δεν πιστεύει πως η τεχνολογία είναι καθοριστική για τις start-ups. Για αυτόν, σημασία έχει μόνο να αναπτύσσονται γρήγορα οι εταιρίες. Όπως είπε «το να είναι μια εταιρία νεοϊδρυθείς δεν την καθιστά από μόνο του start-up. Ούτε είναι αναγκαίο για μια start-up να σχετίζεται με την τεχνολογία, ή να λαμβάνει χρηματοδότηση κεφαλαίων. Το μόνο καθοριστικό είναι η ανάπτυξη» (Graham, 2012). Συμπεριλαμβάνοντας κατηγορηματικά μη τεχνολογικές επιχειρήσεις ως start-ups, ο Graham δίνει έναν πιο ρεαλιστικό ορισμό για τις σημερινές επιχειρήσεις, λέγοντας πως εάν μικρές αλλά παράτολμες επιχειρήσεις επιτύχουν θετική ρευστότητα σε σύντομο χρονικό διάστημα, χωρίς να αυξάνονται τα χρέη επιτρέποντάς τους να συγκεντρωθούν στην εξάπλωσή τους όσο πιο γρήγορα μπορούν, θεωρούνται start-ups.

Από την άλλη, ο Steve Blank, συγγραφέας του Four Steps to the Epiphany, παρόλο που δίνει παρόμοιο ορισμό με τον Graham, την ορίζει διαφορετικά προσθέτοντας τη σημαντική έννοια της επεκτασιμότητας: μια start-up είναι ένας προσωρινός οργανισμός που χρησιμοποιείται για την εύρεση ενός επαναλαμβανόμενου και επεκτάσιμου επιχειρηματικού μοντέλου. Μόλις μια start-up βρει το μοντέλο, παύει να θεωρείται start-up (Blank, 2020).

Οι Ripsas και Tröger ορίζουν πιο λεπτομερώς μια startup ως μια εταιρία που:

- Λειτουργεί κάτω από 10 χρόνια.
- Διαθέτει καινοτόμο επιχειρηματικό μοντέλο ή χρησιμοποιεί καινοτόμες τεχνολογίες.
- Παρουσιάζει σημαντική ανάπτυξη στον αριθμό των υπαλλήλων ή τις πωλήσεις.

Πρέπει να ικανοποιείται οπωσδήποτε το πρώτο κριτήριο και τουλάχιστον ένα από τα επόμενα δύο για να θεωρηθεί startup μια επιχείρηση (Ripsas and Tröger, 2014). Δίνεται έμφαση όχι μόνο στην ηλικία της εταιρίας, αλλά κυρίως στην καινοτομία και την ανάπτυξη. Στόχος για τους συγκεκριμένους επιχειρηματίες δεν αποτελεί η επίτευξη μιας σταθερής, μικρής και σίγουρης ροής χρημάτων, αλλά η ραγδαία ανάπτυξη (Ripsas et al., 2018).

Αυτές οι παράτολμες επιχειρήσεις συχνά λαμβάνουν μια αρχική ώθηση από τους επιχειρηματίες ιδρυτές τους, καθώς αυτοί επιχειρούν να χρηματοδοτήσουν την δημιουργία του προϊόντος ή την ικανοποίηση της υπηρεσίας. Εξαιτίας του περιορισμένου εισοδήματος και του ραγδαία αυξανόμενου κόστους, οι περισσότερες αυτές μικρής κλίμακας επιχειρήσεις δεν είναι μακροπρόθεσμα βιώσιμες χωρίς επιπλέον χρηματοδότηση από επενδυτές κεφαλαίων (venture capitalists). Στα τέλη της δεκαετίας του 1990, ο πιο ευρέως διαδεδομένος τύπος start-up ήταν γνωστός ως “dotcom”. Καθώς το ιντερνέτ διαδόθηκε και οι υπολογιστές διαδραμάτισαν σημαντικό ρόλο στην ανθρώπινη καθημερινότητα, το επιχειρηματικό κεφάλαιο έγινε εξαιρετικά εύκολο να αποκτηθεί, γεγονός που οφείλεται στον ενθουσιασμό των επενδυτών να βγάλουν χρήματα μέσα από τις καινοτόμες λειτουργίας που αυτές οι νεοσύστατες επιχειρήσεις ικανοποιούν. Δυστυχώς, μεταξύ 1997 και 2001, στην κρίση γνωστή ως “dotcom bubble”, οι περισσότερες αυτές start-up χρεοκόπησαν εξαιτίας του ελλιπούς επιχειρηματικού σχεδιασμού, όπως η έλλειψη διαθέσιμων πόρων.

Επίσης, σημαντική καθίσταται η κατανόηση της διαφοράς μεταξύ start-up και παραδοσιακής μικρής επιχείρησης. Η πρώτη έχει πιθανότητα αποτυχίας 90% (embroker.com), ενώ η δεύτερη 20% (Bureau of Labor Statistics). Ωστόσο, ένα συνεργείο αυτοκινήτων ή ένα καθαριστήριο δύσκολα θα επιτύχουν να μπουν στον κατάλογο Fortune 500 (οι 500 επιχειρήσεις με το μεγαλύτερο εισόδημα στις ΗΠΑ), ενώ εκατοντάδες start-ups ανήκουν εκεί.

Είναι ένα παιχνίδι με μεγαλύτερο ρίσκο, αλλά και μεγαλύτερες ανταμοιβές. Γνωρίζοντας το μεγάλο ρίσκο και το υψηλό ποσοστό αποτυχίας των start-ups, αλλά και την εκθετική ανάπτυξη τους στις ΗΠΑ και την Ευρώπη, καθώς επίσης και την αυξανόμενη αξία τους για τις εθνικές οικονομίες, μοιάζει να αποτελεί μεγάλης αξίας πρόκληση η μελέτη του φαινομένου, το οποίο απασχολεί τόσους ανθρώπους παγκοσμίως.

1.3 Καθοριστικοί παράγοντες

Τοποθεσία

Οι νεοφυείς επιχειρήσεις μπορούν να λειτουργούν μέσω διαδικτύου σε κεντρικό γραφείο ή από το σπίτι των εργαζομένων. Ανάλογα με το προϊόν ή την υπηρεσία ενδεχομένως να απαιτείται και φυσικό μαγαζί όπου θα εκθέτονται τα προϊόντα.

Νομική Μορφή

Η κάθε επιχείρηση πρέπει να αποφασίσει με ποια νομική μορφή θα συσταθεί. Συνήθως προτιμάται η εταιρία περιορισμένης ευθύνης (Ε.Π.Ε.) ή η ιδιωτική κεφαλαιουχική εταιρία (Ι.Κ.Ε.) με τη δεύτερη να είναι πιο δημοφιλής εξαιτίας της υψηλής ευελιξίας της.

Χρηματοδότηση

Συνήθως, στα πρώτα στάδια η πηγή χρηματοδότησης είναι φίλοι και οικογένεια, ενώ αν καταφέρουν να εξασφαλίσουν πόρους από επενδυτές κεφαλαίου έχουν ένα σημαντικό βοήθημα. Οι τελευταίοι είναι επαγγελματίες επενδυτές που ειδικεύονται σε νεοφυείς επιχειρήσεις.

Ο Thiel θέτει 7 ερωτήματα που καθορίζουν την επιτυχία μιας εταιρίας:

- *Μηχανική*: χρησιμοποιείται καινοτόμα τεχνολογία αντί για συνεχείς βελτιώσεις;
- *Συγχρονισμός*: μπαίνει η εταιρία στην αγορά την κατάλληλη στιγμή;
- *Μονοπώλιο*: είναι έτοιμη η επιχείρηση να καταλάβει μεγάλο μερίδιο μιας μικρής αγοράς και μετά να επεκταθεί σε μεγαλύτερες αγορές;
- *Άνθρωποι*: αποτελείται η εταιρία από την κατάλληλη ομάδα;
- *Διανομή*: ακολουθείται η βέλτιστη στρατηγική διανομής των προϊόντων στους πελάτες;
- *Διάρκεια*: θα έχει ζήτηση στην αγορά το προϊόν ή η υπηρεσία σε 10 με 20 έτη;
- *Μυστικό*: διαθέτει η εταιρία γνώση που είναι άγνωστη σε άλλους ή την αγνοούν;

1.4 Επιτυχημένες start-ups

Έχουν τεθεί πολλά διαφορετικά κριτήρια προκειμένου μια εταιρία να θεωρηθεί επιτυχημένη. Αυτά ταξινομούνται σε υποκειμενικά και αντικειμενικά, με την πρώτη κατηγορία να αποτελείται από την προσφορά στην ανθρωπότητα, τον διαθέσιμο ελεύθερο χρόνο και το ομαδικό πνεύμα μεταξύ των συνεργατών. Στην παρούσα εργασία, ωστόσο, έμφαση δίνεται σε αντικειμενικά κριτήρια και συγκεκριμένα στα οικονομικά κέρδη.

Προκειμένου ένας επιχειρηματίας να βγάλει άμεσα κέρδος πρέπει να ακολουθήσει μια στρατηγική εξόδου (exit strategy), ρευστοποιώντας το μερίδιο του από την επιχείρηση. Υπάρχουν 5 κύριες στρατηγικές:

- *Συγχώνευση και εξαγορά* (Merger & Acquisition, M&A). Αυτό σημαίνει συγχώνευση με παρόμοια εταιρία ή εξαγορά από μια μεγαλύτερη. Αποτελεί μια κερδοφόρα κατάσταση και για τις δύο πλευρές όταν έχουν συμπληρωματικές ικανότητες και μπορούν να εξοικονομήσουν πόρους με την ένωση. Για τις μεγαλύτερες επιχειρήσεις αποτελεί αποτελεσματικός και άμεσος τρόπος αύξησης των κερδών σε σύγκριση με τη δημιουργία προϊόντος από την αρχή.
- *Εισαγωγή στο χρηματιστήριο* (Initial Public Offering, IPO). Αποτελεί την πιο δημοφιλή επιλογή και το γρηγορότερο τρόπο για πλούτη. Μετά τη φούσκα του 2000, οι εισαγωγές στο χρηματιστήριο έχουν μειωθεί και συνήθως δεν προτιμάται ως επιλογή, αφού οι μέτοχοι είναι απαιτητικοί και δύσπιστοι.
- *Πώληση σε ιδιώτη*. Κάποιος με περισσότερη όρεξη, που πιστεύει στην ιδέα και έχει νέες προτάσεις αγοράζει την επιχείρηση.
- *Πηγή εισοδήματος*. Διατήρηση της επιχείρησης που βγάζει σταθερό εισόδημα, ως μια μόνιμη και σταθερή πηγή κέρδους, δίνοντας τη διαχείριση σε κάποιον άλλον. Ωστόσο, η διατήρηση της επιχείρησης είναι συνήθως απαιτητική, καθώς συνεχώς προκύπτουν ζητήματα και χρειάζονται πόροι.
- *Ρευστοποίηση και κλείσιμο*. Οι έμπειροι επιχειρηματίες καταλαβαίνουν πότε είναι η κατάλληλη στιγμή για να τερματίσουν μια επιχείρηση.

Ο όρος στρατηγική εξόδου, για πολλούς, ακούγεται ως κάτι αρνητικό. Ο αποτελεσματικότερος τρόπος είναι να λειτουργεί κανείς προνοητικά, βελτιστοποιώντας μια ευνοϊκή κατάσταση, παρά αποφεύγοντας μια δύσκολη. Με αυτό τον τρόπο θα μπορεί να εργαστεί αφοσιωμένος και να γίνει η εταιρία του ακόμη πιο ελκυστική σε μελλοντικούς αγοραστές.

Στην παρούσα μελέτη, μια start-up θεωρείται επιτυχημένη όταν προσφέρει ένα σημαντικό χρηματικό ποσό στους ιδρυτές, επενδυτές και πρώτους υπαλλήλους της. Αυτό επιτυγχάνεται με δύο τρόπους, είτε μοιράζοντας μετοχές (IPO, Initial Public Offering) αφού εισαχθεί στο χρηματιστήριο (πχ η Facebook μπήκε στο χρηματιστήριο επιτρέποντας στον καθένα να επενδύσει στην εταίρα αγοράζοντας μετοχές), είτε αν εξαγοραστεί (ή συγχωνευτεί) από μια άλλη εταιρία (πχ. Η Microsoft εξαγόρασε την LinkedIn για 26 δισεκατομμύρια δολάρια), όπου αυτοί που είχαν προηγουμένως επενδύσει λαμβάνουν άμεσα μετρητά για τις μετοχές τους. Η συγκεκριμένη προσέγγιση συχνά επικρατεί ως στρατηγική εξόδου. Και οι δύο στρατηγικές, εισαγωγή στο χρηματιστήριο, εξαγορά ή συγχώνευση, θεωρούνται ως κρίσιμα γεγονότα για την ταξινόμηση μιας start-up ως επιτυχημένης.

Οι συγχωνεύσεις και οι εξαγορές συνήθως ορίζονται ως M&As και παίζουν σημαντικό ρόλο στην εταιρική αναδιάρθρωση. Η συγχώνευση είναι η στρατηγική ένωσης δύο εταιριών για το σχηματισμό μίας ενιαίας εταιρίας, συχνά με νέο όνομα, για να αυξηθούν τα κέρδη, ενώ σε μη τεχνολογικές εταιρίες, πιο συχνά συμβαίνουν μεταξύ παρόμοιων εταιριών σε μέγεθος και στάτους. Οι συγχωνεύσεις είναι ιδιαίτερα καθοριστικές για τις υψηλά τεχνολογικές βιομηχανίες, καθώς συχνά χρησιμοποιούνται για την απόκτηση υπερσύγχρονων τεχνολογιών ή την ραγδαία επέκταση δυνατοτήτων έρευνας και ανάπτυξης. Αντίστοιχα, μια εταιρία εξαγοράζει μια μικρότερη για να αποκτήσει την ταλαντούχα ομάδα της. Η εταιρίας γονέας αγοράζει μαζί, τεχνολογία και υπαλλήλους. Αυτού του είδους η εξαγορά παρέχει μια γρήγορη στρατηγική ανάπτυξης σε ανταγωνιστικές αγορές. Η λογική πίσω από την εξαγορά ή συγχώνευση είναι πως οι

δύο εταιρίες έχουν μεγαλύτερη αξία μαζί παρά ως ξεχωριστές εταιρίες. Αυτή η εδραίωση των δύο εταιριών είναι καθοριστική για την εταιρική στρατηγική ώστε να διατηρηθούν τα ανταγωνιστικά πλεονεκτήματά τους. Η αξία των συγχωνεύσεων και εξαγορών παγκοσμίως σχεδόν άγγιξε τα 5 τρις δολάρια τις χρονιές 2007 και 2015 (statista 2021), με τον αριθμό των συναλλαγών να είναι περίπου 47,400 (imaa-institute.org).

Η εισαγωγή στο χρηματιστήριο είναι η πρώτη πώληση μιας μετοχής από μια ιδιωτική επιχείρηση στο κοινό και αποτελεί καθοριστικό γεγονός για τον κύκλο ζωής της επιχείρησης. Με την εισαγωγή στο χρηματιστήριο, η επιχείρηση μοιράζει μετοχές με αποτέλεσμα να μπορεί να λάβει επιπλέον χρηματοδότηση, ενώ επιτρέπει στα μέλη της να πουλήσουν τις μετοχές τους ρευστοποιώντας την ιδιοκτησία τους.

1.5 Στόχοι

Με στόχο να διευκρινιστεί με ποιον τρόπο μια start-up ή ένας επενδυτής θα μπορούσε να έχει γνώση για να λάβει καλύτερες αποφάσεις για την επενδυτική του στρατηγική και να βγάλει μεγαλύτερο χρηματικό κέρδος, η παρούσα έρευνα στοχεύει εφαρμόζοντας μεθόδους εξόρυξης δεδομένων και τεχνικές μηχανικής μάθησης να δημιουργήσει ένα μοντέλο προβλέψεων που να ταξινομεί, αν μια εταιρία είναι (ήδη) επιτυχημένη ή όχι (δυαδική ταξινόμηση).

Για τη δημιουργία των προβλεπτικών μοντέλων δοκιμάστηκαν 6 τεχνικές: Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines, SVM), Λογιστική Παλινδρόμηση (Logistic Regression, LR), κ Κοντινότεροι Γείτονες (KNN), Τυχαία Δέντρα (Random Forests, RF), Γκαουσιανός Αφελής Μπέυζ (Gaussian Naïve Bayes, GNB), XGBoost. Όλοι οι παραπάνω αλγόριθμοι αξιοποίησαν τα ίδια χαρακτηριστικά του σετ δεδομένων προσφέροντας γρήγορη και απλή τεχνική εφαρμογή. Η δημιουργία προβλεπτικών μοντέλων που να εξηγούν το φαινόμενο είναι ένας πολύ καλός δείκτης των δυνατοτήτων των τεχνικών εξόρυξης δεδομένων για το πόση πληροφορία μπορούν οι μελετητές να αποκτήσουν από τα διαθέσιμα δεδομένα. Εάν είναι εφικτό να ταξινομηθεί με ακρίβεια μια start-up ως επιτυχημένη με βάση τα γεγονότα που συνέβησαν από την ίδρυσή της, προσφέρεται ανεκτίμητη αξία τόσο στους εμπλεκόμενους με start-ups όσο και στην ακαδημαϊκή κοινότητα και τη βιομηχανία, με την εφαρμογή διαφορετικών τεχνικών και χαρακτηριστικών για τη δημιουργία μοντέλων με μεγάλη προβλεπτική ακρίβεια.

Προηγούμενες μελέτες επικεντρώνονται κυρίως σε διοικητικά χαρακτηριστικά ή τη σύνοψη οικονομικών χαρακτηριστικών σχετικών με την χρηματοδότηση (ιδιαίτερα το επενδυτικό κεφάλαιο). Στόχος αποτελεί να μειωθεί αυτό το χάσμα δημιουργώντας χαρακτηριστικά σχετικά με την χρηματοδότηση με μεγάλη προβλεπτική ισχύ για την ταξινόμηση των εταιριών. Επιπρόσθετα, υπάρχει χώρος για βελτίωση της ποιότητας του δείγματος με την πιο προσεκτική επιλογή των εταιριών και με την αποδοτικότερη διαχείριση των ελλιπών δεδομένων του σετ.

1.5.1 Τεχνικοί Στόχοι

Κατά τη διαδικασία, αναμένεται να επιτευχθούν ορισμένοι τεχνικοί στόχοι.

Κατά την πρώτη φάση της ανάλυσης δεδομένων, αναμένεται η πλήρης κατανόηση των δεδομένων που παρέχονται από το CrunchBase, ακολουθουμένη από τη διαδικασία

του καθαρισμού των δεδομένων (ελλιπείς τιμές, διπλότυπα, περιττά δεδομένα). Έχοντας ένα τόσο ευρύ σετ δεδομένων είναι καθοριστικής σημασίας η επιλογή και ο συνδυασμός των καταλληλότερων χαρακτηριστικών που θα δώσουν την περισσότερη χρήσιμη και αξιοποιήσιμη πληροφορία. Με την επεξεργασία, τον συνδυασμό και την μετατροπή των δεδομένων θα δημιουργηθούν νέα χαρακτηριστικά τα οποία θα αποτελέσουν και την τελική γνώση για την εκπαίδευση των μοντέλων.

Κατά τη δεύτερη φάση που αποτελείται από τις δοκιμές και τα αποτελέσματα, τα πειράματα θα διεξαχθούν δοκιμάζοντας διαφορετικούς αλγόριθμους μηχανικής μάθησης για τη δημιουργία των πιο αποτελεσματικών μοντέλων που θα εκπαιδεύονται με επιβλεπόμενη μάθηση και θα προσπαθήσουν να υπερβούν τα προηγούμενα αποτελέσματα ερευνών.

1.6 Δομή εργασίας

Ακολουθούν τέσσερα κεφάλαια: αρχικά γίνεται μια βιβλιογραφική παρουσίαση προηγούμενων μελετών στο αντικείμενο. Στη συνέχεια, πραγματοποιείται μια θεωρητική προσέγγιση του αντικειμένου της Μηχανικής Μάθησης και της Ανάλυσης Δεδομένων. Ύστερα, παρουσιάζεται η διαδικασία δημιουργίας του τελικού σετ δεδομένων από το CrunchBase, όπου περιλαμβάνεται η προ-επεξεργασία των δεδομένων η δημιουργία νέων μεταβλητών, τα προβλήματα που προέκυψαν με τις λύσεις τους και τα αποτελέσματα των αλγορίθμων. Τέλος, προκύπτουν τα τελικά συμπεράσματα και γίνονται προτάσεις για μελλοντική μελέτη.

Κεφάλαιο 2: Παρουσίαση υπάρχουσας βιβλιογραφίας

2.1 Πρόβλεψη επιτυχίας επιχειρήσεων βάση οικονομικών χαρακτηριστικών

Η περισσότερη έρευνα για την βελτιστοποίηση των προβλέψεων απόδοσης των επιχειρήσεων επικεντρώνεται στην ανάλυση ποσοτικών οικονομικών μεταβλητών για τις εταιρίες ως προς το μέγεθος, την αγοραστική προς τη λογιστική αξία, τις ταμειακές ροές, τον δείκτη χρεών προς ίδιο κεφάλαιο και τον δείκτη μερίσματος τιμής (Ali-Yrkkö et al., 2005), (Gugler and Konrad, 2002), (Meador et al., 1996). Χρησιμοποιούνται μερικά ακόμα διοικητικά χαρακτηριστικά ως προς τις διακυμάνσεις της βιομηχανίας, τη διοικητική αναποτελεσματικότητα (Ali-Yrkkö et al., 2005) και την πληθώρα πόρων (Meador et al., 1996). Οι περισσότεροι μέθοδοι ανάλυσης χρησιμοποιούν Λογιστική Παλινδρόμηση ή Πολυωνυμική Λογιστική Παλινδρόμηση (Logistic Regressions or Multinomial Logistic Regressions) για τη δημιουργία μοντέλων προβλέψεων συγχωνεύσεων και εξαγορών (Ali-Yrkkö et al., 2005), (Gugler and Konrad, 2002), (Meador et al., 1996), (Ragothaman, 2003).

Οι Hyytinen και Ali-Yrkko (2005) παρουσίασαν «πώς οι πολυωνυμικές λογιστικές εκτιμήσεις δείχνουν ότι εάν σε μια φιλανδική εταιρία ανήκουν πατέντες καταγεγραμμένες στο Ευρωπαϊκό Γραφείο Πατεντών (European Patent Office, EPO), οι πατέντες αυξάνουν την πιθανότητα η εταιρία να εξαγοραστεί από ξένη εταιρία». Οι ερευνητές έλαβαν υπόψη τους και άλλες μεταβλητές για τα μοντέλα τους, όπως το μέγεθος της επιχείρησης, την ταμειακή ροή σε σχέση με τα περιουσιακά στοιχεία και την απόδοση των επενδύσεων (ROI, return on investment) για την προσομοίωση της διοικητικής απόδοσης (Ali-Yrkkö et al., 2005). Ένα σχετικό εύρημα στην εργασία τους είναι πως το μέγεθος, ως λογαριθμική συνάρτηση των συνολικών περιουσιακών στοιχείων που ανήκουν σε μια επιχείρηση, παίζει σημαντικό ρόλο. Όσο μεγαλύτερη η εταιρία, τόσο πιο πιθανό να εξαγοραστεί. Ωστόσο, το δείγμα τους με 815 φιλανδικές επιχειρήσεις είναι πολύ μικρό για να δοκιμαστεί με πιο ισχυρές τεχνικές.

Οι Wei et al. (2008), μελέτησαν, επίσης, την αξία των πατεντών που μια επιχείρηση έχει ενισχύοντας τις προβλέψεις εξαγοράς και συγχώνευσης. Μέσα από Μπεϋζιανά μοντέλα (Bayesian models) για την ταξινόμηση μιας εταιρίας ως υποψήφιας εταιρίας στόχου για συγχώνευση ή εξαγορά από πλειοδότριες εταιρίες ή όχι, όρισαν ένα σετ χαρακτηριστικών όπως ο αριθμός των πατεντών που πραγματοποίησε μια επιχείρηση, τον αριθμό και την επίδραση των τελευταίων πατεντών και την τεχνολογική ποσότητα που ανέπτυξε η επιχείρηση. Τα αποτελέσματά τους, σε ένα σετ με 2394 εξαγορές, κινείται μεταξύ ενός βαθμού ακρίβειας από 42.9% έως 46.4% για τις εταιρίες που συγχωνεύτηκαν ή εξαγοράστηκαν (Wei et al., 2008). Παρόλο που έγιναν σχετικά βήματα για την πρόβλεψη εξαγορών και συγχωνεύσεων που περιλαμβάνουν τεχνολογικές μεταβλητές, τα αποτελέσματά τους περιορίστηκαν με τον αποκλεισμό όλων των υπόλοιπων χαρακτηριστικών όπως διοικητικών και οικονομικών.

2.2 Αναγνώριση χρεοκοπίας

Επίσης, μεγάλες μελέτες έχουν επικεντρωθεί στις αποτυχίες και χρεοκοπίες εταιριών. Ο καθηγητής Edward Altman, γνωστός για τη δημιουργία του (Altman) Z-score, πρότεινε πολλές οικονομικές αναλογίες, ως χαρακτηριστικά για τις στατιστικές

αναλύσεις πολλαπλών χαρακτηριστικών, στην εργασία του για την πρόβλεψη χρεοκοπίας. Ο Altman επέκτεινε την πρώτη έρευνά του στην πρόβλεψη πτώχευσης σιδηροδρομικών εταιριών στην Αμερική χρησιμοποιώντας ένα σύνολο 21 σιδηροδρομικών εταιριών που χρεοκόπησαν μεταξύ 1939 και 1970. Συγκριμένα, ο Altman με ένα μοντέλο 5 παραμέτρων χρησιμοποίησε την πολλαπλή μεροληπτική ανάλυση, αναλύοντας αναλογίες, όπως συνήθη μέτρα ρευστότητας, φερεγγυότητας και επιρροής, και μέτρα αποδοτικότητας συν δείκτες απόδοσης με πολύ ακριβή ταξινόμηση σε ένα και δύο χρόνια πριν την χρεοκοπία (πετυχαίνοντας ακρίβεια 97.7%) (Altman, 1968).

2.3 Αναγνώριση απάτης

Πιο πρόσφατα, οι Ravisankar et al. (2011), χρησιμοποίησαν 6 αλγόριθμους μηχανικής μάθησης, Πολυστρωματικά Εμπρόσθιας Τροφοδότησης Νευρωνικά Δίκτυα (Multilayer Feed Forward Neural Network, MLFF), Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines, SVM), Γενετικό Προγραμματισμό (Genetic Programming, GP), Ομαδική Μέθοδο Χειρισμού Δεδομένων (Group Method of Data Handling, GMDH), Λογιστική Παλινδρόμηση (Logistic Regression, LR) και Πιθανοτικά Νευρωνικά Δίκτυα (Probabilistic Neural Network, PNN), για να κατανοήσουν τις διαφορές μεταξύ ενός συνόλου 202 επιχειρήσεων που συμμετέχουν σε πολλά κινέζικα χρηματιστήρια, χρησιμοποιώντας 35 οικονομικά χαρακτηριστικά. Το σύνολο δεδομένων αποτελούνταν από 101 έντιμες εταιρίες και 101 ανέντιμες. Το Πιθανοτικό Νευρωνικό Δίκτυο (Probabilistic Neural Network) ξεπέρασε όλους τους άλλους ταξινομητές με ποσοστό Θετικών Αληθών 98% προβλέποντας ποιες εταιρίες είναι ανέντιμες (Ravisankar et al., 2011). Τα νούμερά τους είναι εντυπωσιακά, αλλά η χρήση ενός τόσου μικρού δείγματος 202 εταιριών και η έλλειψη εξερευνητικής ανάλυσης των χαρακτηριστικών που χρησιμοποιήθηκαν επιτρέπουν την εξαγωγή του συμπεράσματος πως εξαιρετικές διαφορές υπάρχουν μεταξύ έντιμων και ανέντιμων επιχειρήσεων που διακρίνονται με ευκολία κατά τη διαδικασία μάθησης. Η προσέγγισή τους έχει τα υψηλότερα αποτελέσματα, ωστόσο δεν επικεντρώνεται σε εξαγορές επιχειρήσεων, αλλά την πρόληψη απάτης.

2.4 Μελέτη επενδυτικής συμπεριφοράς

Η επενδυτική συμπεριφορά εταιριών επενδυτικού κεφαλαίου και άλλων επενδυτών νεοφυών επιχειρήσεων είναι ένας άλλος τομέας έρευνας. Οι Liang and Daphne Yuan (2012) χρησιμοποίησαν τη βάση δεδομένων του CrunchBase για την πρόβλεψη επενδυτικών συμπεριφορών συνδυάζοντας τα μέσα κοινωνικής δικτύωσης και την επιβλεπόμενη μάθηση. Παραδοσιακά, η έρευνα στην επενδυτική συμπεριφορά επικεντρώνεται σε ψυχολογικούς παράγοντες, ταύτιση απόψεων και προηγούμενη εμπειρία των επενδυτών. Στην εργασία τους έδωσαν πρακτικούς κανόνες σε εταιρίες που αναζητούν επενδυτές βασισμένους σε κοινωνικές σχέσεις. Μοντελοποίησαν την επενδυτική συμπεριφορά μέσα από την κλασσική σύνδεση προβλήματος, καθώς σύγκριναν κάθε ζευγάρι Επενδυτή-Εταιρίας για την πρόβλεψη εάν ο επενδυτής θα επενδύσει σε μια εταιρία βασιζόμενοι στο πόσο όμοιοι ή διαφορετικοί είναι σε όρους κοινωνικών σχέσεων. Τον Μάιο του 2012, η βάση δεδομένων τους αποτελούνταν από 89,370 εταιρίες και 28,108 επενδυτικούς γύρους. Χρησιμοποιώντας Δένδρα Αποφάσεων ως αλγόριθμους μάθησης, κατάφεραν ρυθμό ορθών θετικών (TPR, True Positive Rate) 87.53% με περιοχή κάτω από την καμπύλη (AUC, Area Under Curve)

77%. Παρόλο που δεν προβλέπουν απευθείας εξαγορές, η μελέτη τους αναγνωρίζει επιτυχημένες εταιρίες (Liang and Daphne Yuan, 2012).

Συγγραφείς	Τίτλος	Έτος	Τεχνική	Αποτελέσματα
El Altman	Οικονομικές αναλογίες, μεροληπτική ανάλυση και πρόβλεψη εταιρικής χρεοκοπίας	1968	Πολλαπλή μεροληπτική στατιστική	94% ακρίβεια
AL Meador, PH Church, LG Rayburn	Ανάπτυξη προβλεπτικών μοντέλων για οριζόντιες και κάθετες συγχωνεύσεις	1996	Δυαδική λογιστική παλινδρόμηση	77.27% σωστή πρόβλεψη για όσες συγχωνεύτηκαν 50% σωστή πρόβλεψη για όσες δεν συγχωνεύτηκαν
CP Wei, YS Jiang, CS Yang	Ανάλυση πατεντών για την υποστήριξη προβλέψεων συγχώνευσης και εξαγοράς: Μια προσέγγιση εξόρυξης δεδομένων	2008	Μπευζιανά μοντέλα	88% ακρίβεια
P Ravisankar, V Ravi, GR Rao, I Bose	Αναγνώριση οικονομικής απάτης και επιλογή χαρακτηριστικών χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων	2011	Πιθανοτικά Νευρωνικά Δίκτυα	98% ποσοστό Ορθών Αληθών
Liang & Daphne Yuan	Οι επενδυτές είναι κοινωνικά ζώα: Προβλέποντας την συμπεριφορά του επενδυτή χρησιμοποιώντας χαρακτηριστικά κοινωνικών δικτύων μέσω επιβλεπόμενης μάθησης	2012	Δένδρα Αποφάσεων	87% ποσοστό ορθών θετικών

Πίνακας 2.1 – Προηγούμενες μελέτες σχετικές με επιχειρήσεις και επενδύσεις

2.5 Πρόβλεψη επιτυχίας επιχειρήσεων με μεθόδους μηχανικής μάθησης

Το ACQTARGE είναι ένα εργαλείο για την ταξινόμηση εταιριών σε εξαγοράσιμους (ή όχι) στόχους που χρησιμοποιεί μεροληπτική ανάλυση και επαγωγή κανόνων στα μοντέλα του. Στόχος του είναι να βοηθήσει αναλυτές και επενδυτές στις αποφάσεις τους σχετικά με αγορές ή πωλήσεις επιχειρήσεων. Το εργαλείο αναπτύχθηκε με μια βάση 97 εξαγορασμένων και 97 μη εξαγορασμένων εταιριών (65-65 για εκπαίδευση και 32-32 για αξιολόγηση), με ποσοστό ακρίβειας 82.9% (Ragothaman et al., 2003).

Παρόλο που το αποτέλεσμα είναι ευνοϊκό, η μικρή βάση δεδομένων και η χρήση μόνο 8 οικονομικών χαρακτηριστικών περιορίζουν τη χρήση του.

Χρησιμοποιώντας δομημένη βάση δεδομένων, αλλά δίνοντας έμφαση στις εξαγορές νεοφυών επιχειρήσεων και τις επενδύσεις μέσω επενδυτικού κεφαλαίου, οι Xiang et al. (2012), προβλέπουν την εξαγορά επιχειρήσεων σε παρόμοιους κλάδους συνδυάζοντας δομημένα δεδομένα από το CrunchBase και την εξόρυξη κειμένου ειδήσεων από την ιστοσελίδα TechCrunch. Το ποσοστό ορθών θετικών προβλέψεων του μοντέλου τους κυμαίνεται μεταξύ 60% και 79.8% για διαφορετικές κατηγορίες επιχειρήσεων με μικρό αριθμό ελλιπών χαρακτηριστικών χρησιμοποιώντας Μπεϋζιανά Δίκτυα (Bayesian Network, BN) ως αλγορίθμους μηχανικής μάθησης. Το ποσοστό λάθος θετικών FPR (False Positive Rate) κυμαίνεται μεταξύ 0 και 8.3% στις κατηγορίες με λίγα ελλιπή δεδομένα στο CrunchBase (Xiang et al., 2012).

Οι Chenchen Pan et al. (2018) παρατηρώντας την ραγδαία άνθιση των νεοφυών επιχειρήσεων σε ΗΠΑ και Κίνα θέλησαν να μελετήσουν ποια χαρακτηριστικά είναι καθοριστικά για την επιτυχία μιας εταιρίες και πώς να πραγματοποιούν πετυχημένες προβλέψεις. Χρησιμοποιήθηκε ο αλγόριθμος των κ Κοντινότερων Γειτόνων, τα αποτελέσματα του οποίου ήταν εμφανώς βελτιωμένα συγκριτικά με προηγούμενες μελέτες που χρησιμοποιήθηκαν οι αλγόριθμοι της Λογιστικής Παλινδρόμησης και των Τυχαίων Δένδρων (Chenchen Pan et al., 2018).

Αντίστοιχα, κινητοποιημένος από την άμεση επιτυχία κολοσσών όπως η Facebook, η Apple, η Airbnb και η Uber, ο Bento (2018) αφοσιώθηκε στη μελέτη νεοφυών επιχειρήσεων των ΗΠΑ, και συγκεκριμένα των 5 πιο δημοφιλών πολιτειών. Στην εργασία του έγινε φανερό πως οι εταιρίες στην Καλιφόρνια και ειδικότερα της Silicon Valley πετυχαίνουν πολύ πιο γρήγορα να εισαχθούν στο χρηματιστήριο ή να εξαγοραστούν, προσφέροντας αρκετά χρήματα στους ιδρυτές, τους πρώτους επενδυτές και εργαζομένους τους. Χρησιμοποιώντας τη βάση δεδομένων CrunchBase ανέδειξε την σημασία του επενδυτικού κεφαλαίου για την ταχεία ανάπτυξη μιας νέας επιχείρησης (Bento, 2018).

Οι Sharchilev et al. (2018) αξιοποιώντας πληροφορίες που βρίσκονται στο διαδίκτυο, αλλά και στο CrunchBase επιχείρησαν να προβλέψουν την επιτυχία νεοφυών επιχειρήσεων στα πρώτα στάδια ανάπτυξής τους. Συγκεκριμένα, εξετάζουν αν εταιρίες που έχουν λάβει μια πρώτη χρηματοδότηση, θα λάβουν και επιπλέον σε ένα σύντομο χρονικό διάστημα. Με μοντέλα ενίσχυσης κλίσης (gradient boosting) και αναζήτηση πληροφοριών σε ανοιχτές βάσεις δεδομένων (open source databases) διαφοροποιήθηκαν από τις μέχρι τότε μελέτες, αφού αυτές αφορούσαν αποκλειστικά δομημένες βάσεις δεδομένων. Οι πληροφορίες που βρίσκονται στο διαδίκτυο για την κάθε εταιρία βελτιώνουν σημαντικά τις προβλέψεις, καθώς λαμβάνεται υπόψη η αθροιστική γνώμη των πελατών για αυτές (Sharchilev et al., 2018).

Οι Antretter et al. (2019) μελέτησαν την γνησιότητα των μέσων κοινωνικής δικτύωσης, προκειμένου να αξιολογήσουν κατά πόσο μπορούν να μελετηθούν και να συνεισφέρουν στην βελτίωση των προβλέψεων στον τομέα των επιχειρήσεων. Αναλύοντας πάνω από 187 χιλιάδες αναρτήσεις και 253 νέες επιχειρήσεις κατάφεραν με καλό ποσοστό (74%) να ξεχωρίσουν τις εταιρίες που έκλεισαν, από αυτές που επιβίωσαν. Για τη μελέτη τους χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης εξειδικευμένες στην ανάλυση κειμένου (Antretter et al., 2019).

Προσφάτως, πολλές μελέτες επικεντρώνονται στην αποτελεσματικότερη πληροφόρηση των επενδυτών. Συγκεκριμένα, οι Arroyo et al. (2019) και οι Żbikowski et al. (2021) χρησιμοποιώντας δεδομένα από το CrunchBase επιχείρησαν να προβλέψουν ποιες εταιρίες θα εξαγοραστούν ή θα μπουν στο χρηματιστήριο, γεγονός που θεωρείται επιτυχία και στην παρούσα εργασία.

Επίσης, έρευνες γίνονται για τις τακτικές εξόδου των επιχειρήσεων. Οι G. C. Calafiore et al. (2020) και οι Bargagli-Stoffi et al. (2020) εφαρμόζοντας τεχνικές μηχανικής μάθησης επιχείρησαν να προβλέψουν αν μια εταιρία θα εξαγοραστεί, θα μπει στο χρηματιστήριο ή θα χρεοκοπήσει σε ένα εύλογο χρονικό διάστημα.

Οι περισσότερες μελέτες που δεν χρησιμοποιούν βάσεις CrunchBase έχουν λίγα και πολύ εξειδικευμένα δεδομένα, και παρά το γεγονός πως πετυχαίνουν ευνοϊκά αποτελέσματα, η επεκτασιμότητα τους περιορίζεται. Ακόμη, οι περισσότερες ερευνητικές εργασίες επικεντρώνονται σε διοικητικά γνωρίσματα τα οποία δεν αντικατοπτρίζουν τη συνολική εικόνα της επιχείρησης ή τις δυνατότητες εξαγοράς της. Οι μελέτες που χρησιμοποιούν βάσεις CrunchBase, επίσης, δεν χρησιμοποιούν όλα τα διαθέσιμα δεδομένα προκειμένου να μην δημιουργήσουν πολλά χαρακτηριστικά σχετιζόμενα με την επίδραση του επενδυτικού κεφαλαίου, όπως το πλήθος των επενδυτών, τους γύρους χρηματοδότησης και το συνολικό ποσό χρηματοδότησής τους μεταξύ άλλων. Προς υπεράσπισή τους, πρέπει να ειπωθεί ότι μερική από την πληροφορία που είναι διαθέσιμη σήμερα ενδέχεται να μην ήταν διαθέσιμη τη στιγμή που διεξαγόταν η έρευνα.

Συγγραφείς	**Τίτλος**	**Έτος**	**Τεχνική Μηχανικής Μάθησης**	**Αποτελέσματα**
Ragothaman, S., Naik, B., & Ramakrishnan, K.	Πρόβλεψη εξαγορών σε επιχειρήσεις: Μια εφαρμογή αβέβαιης λογικής που χρησιμοποιεί επαγωγή κανόνων	2003	Μεροληπτική ανάλυση και επαγωγή κανόνων	81.3% για εξαγορασμένες επιχειρήσεις, 65.6% για μη εξαγορασμένες επιχειρήσεις
Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C. and Liu, C	Μια επιβλεπόμενη προσέγγιση για την πρόβλεψη εξαγορών επιχειρήσεων με πραγματικά και τοπικά χαρακτηριστικά χρησιμοποιώντας προφίλ και άρθρα ειδήσεων στο TechCrunch	2012	Μπευζιανά Δίκτυα	69.4% (κατά μέσο όρο) ποσοστό ορθών θετικών και 4,15% ποσοστό λανθασμένων θετικών
Chenchen Pan, Yuan Gao, Yuzi Luo	Πρόβλεψη μηχανικής μάθησης της επιχειρηματικής επιτυχίας εταιριών	2018	K Κοντινότεροι Γείτονες	44.45% F1-σκορ, 73.70% ακρίβεια

Francisco Ramadas da Silva Ribeiro Bento	Πρόβλεψη επιτυχίας νεοφυών επιχειρήσεων με μηχανική μάθηση	2018	Τυχαία Δένδρα	94.1% ποσοστό ορθών θετικών 7,8% ποσοστό λανθασμένων θετικών
Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M.	Πρόβλεψη επιτυχίας νεοφυών επιχειρήσεων βασισμένη στο διαδίκτυο	2018	Ενίσχυση κλίσης	85.4% περιοχή κάτω από την καμπύλη (ROC-AUC)
Arroyo, J., Corea, F., Jimenez-Diaz, G.,Recio-Garcia, J. A.	Αξιολόγηση επίδοσης της μηχανικής μάθησης για υποστήριξη αποφάσεων σε επενδύσεις κεφαλαίου	2019	Ενίσχυση κλίσης δένδρου (Gradient Tree Boosting)	82% ακρίβεια
Antretter, T., Blohm, I., Grichnik, D., & Wincent, J.	Προβλέποντας την επιβίωση νέων επενδύσεων: Μια προσέγγιση μηχανικής μάθησης βασισμένη στο Twitter για την μέτρηση της σε σύνδεση νομιμότητας	2019	Τυχαία Δένδρα και ενίσχυση κλίσης	76% ακρίβεια
G. C. Calafiore, M. Hillary Morales, V. Tiozzo and S. Marquie	Μοντέλο ψηφοφορίας ταξινομητών για την για την πρόβλεψη εξόδου ιδιωτικών επιχειρήσεων	2020	Λογιστική Παλινδρόμηση, Τυχαία Δένδρα, Μηχανή Διανυσματικής Υποστήριξης	63% ακρίβεια
Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M.	Επιβλεπόμενη μάθηση για την πρόβλεψη της δυναμικής εμπορικών οίκων	2020	Βαθειά μάθηση	
Żbikowski, K., & Antosiuk, P.	Μια αντικειμενική προσέγγιση μηχανικής μάθησης για την πρόβλεψης επιτυχίας επιχειρήσεων χρησιμοποιώντας	2021	Λογιστική Παλινδρόμηση, Μηχανές Διανυσματικής Υποστήριξης, Ταξινομητής	57% ακρίβεια, 34% ανάκληση, 43% F1-σκορ

	δεδομένα του CrunchBase		Ενισχυμένης Κλίσης	

Πίνακας 2.2 – Προηγούμενες μελέτες για πρόβλεψη επιτυχημένων επιχειρήσεων

Κεφάλαιο 3: Θεωρητική Προσέγγιση

3.1 Ανάλυση Δεδομένων

«Πνιγόμαστε από πληροφορίες, αλλά πεινάμε για γνώση» Rutherford D. Roger

Η κοινωνία, αυτή τη στιγμή, κατακλύζεται από συναλλαγές μέσω υπολογιστών μεταξύ επιχειρήσεων, επιστημόνων και κυβερνήσεων, ενώ ταυτόχρονα οι ψηφιακές συσκευές και τα μέσα κοινωνικής δικτύωσης δημιουργούν νέα πληροφορία. Οι επιστήμονες των δεδομένων έχουν να αντιμετωπίσουν την πρόκληση αξιοποίησης των απότομα αυξανόμενων πληροφοριών δημιουργώντας νέες τεχνικές και αυτοματοποιημένα εργαλεία, με στόχο τη μετατροπή των τεράστιων βάσεων δεδομένων σε χρήσιμη πληροφορία και ιδιαίτερα σε γνώση.

Η Ανάλυση Δεδομένων είναι η διαδικασία επιλογής, καθαρισμού, μετασχηματισμού και μοντελοποίησης των δεδομένων με στόχο την ανακάλυψη χρήσιμης πληροφορίας και την εξαγωγή συμπερασμάτων. Χρησιμοποιείται από επιχειρήσεις, επιστήμονες, αλλά και σε κοινωνικούς τομείς. Για τις επιχειρήσεις στόχος είναι, μέσω αυτής, να υποστηρίξουν την λήψη αποφάσεων με μια επιστημονική μέθοδο. Για να προκύψουν συμπεράσματα από τα δεδομένα και να αντιμετωπίσουν το πρόβλημα του υπερβολικού όγκου δεδομένων, οι επιστήμονες που μελετούν τα δεδομένα όρισαν μια διαδικασία που σχετίζεται με την ανάπτυξη μεθόδων και τεχνικών ώστε να τυποποιήσουν την εξόρυξη δεδομένων, την οποία ονόμασαν Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases, KDD). Αποτελείται από τα εξής βήματα:

i. Προ-επεξεργασία δεδομένων
ii. Εξόρυξη δεδομένων
iii. Ερμηνεία και αξιολόγηση αποτελεσμάτων

3.1.1 Προ-επεξεργασία Δεδομένων

Η προ-επεξεργασία δεδομένων ορίζεται ως ο μετασχηματισμός των δεδομένων σε μορφή κατανοητή από τον υπολογιστή και τους αλγορίθμους.

Για την προ-επεξεργασία των δεδομένων δεν είναι πάντα συγκεκριμένος ο αριθμός βημάτων, αλλά εξαρτάται από το είδος του προβλήματος και τα διαθέσιμα δεδομένα. Οπότε κάποιο/κάποια βήματα μπορεί να παραλειφθούν. Γενικά αυτά είναι:

- Αξιολόγηση ποιότητας δεδομένων
- Συγκέντρωση χαρακτηριστικών
- Δειγματοληψία
- Μείωση διαστατικότητας
- Κωδικοποίηση χαρακτηριστικών

3.1.1.1 Αξιολόγηση ποιότητας δεδομένων

Καθώς, συχνά, τα δεδομένα συσσωρεύονται από πολλές και διαφορετικές πηγές, οι οποίες δεν είναι πάντα αξιόπιστες, ενώ έχουν και διαφορετική μορφολογία, ένα σημαντικό μέρος του χρόνου καταναλώνεται στην αντιμετώπιση προβλημάτων ποιότητας και συμβατότητας, πριν αυτά εισαχθούν σε αλγορίθμους μηχανικής

μάθησης. Είναι, προφανώς, μη ρεαλιστικό να αναμένονται τέλεια δεδομένα, ενώ οι περισσότερες ανακρίβειες προκύπτουν από ανθρώπινα λάθη, ανακρίβειες στις συσκευές μέτρησης και αλλαγή του ορισμού μεταβλητών ενώ έχει ξεκινήσει η συλλογή δεδομένων. Τα πιο σημαντικά από αυτά μαζί με τις λύσεις του είναι:

Ελλιπείς τιμές

Είναι συνηθισμένο να υπάρχουν ελλείψεις στις τιμές των δεδομένων και πρέπει να αντιμετωπιστούν.

Διαγραφή των γραμμών που έχουν ελλιπή δεδομένα:

- Η πιο απλή και συνήθως αποτελεσματική στρατηγική. Αποτυγχάνει όταν πολλά αντικείμενα έχουν ελλιπή τιμές. Εάν κάποιο χαρακτηριστικό λείπει συχνά, τότε πρέπει να βγει από την εξίσωση.

Υπολογισμός των χαμένων τιμών:

- Εάν το ποσοστό των ελλιπών τιμών είναι λογικό, τότε είναι πρακτικό να εφαρμοστούν μέθοδοι για να γεμίσουν τα κενά. Οι πιο συνηθισμένες τακτικές είναι να γεμίσουν τα κενά με το μέσο όρο, τη μέση ή την πιο συνηθισμένη τιμή.

Ασυνεπείς τιμές

Σχεδόν πάντα προκύπτουν από ανθρώπινο λάθος είτε λόγω λανθασμένης ανάγνωσης, είτε λάθους κατά την πληκτρολόγηση.

- Για αυτό προτείνεται η αξιολόγηση των δεδομένων εξετάζοντας τι τύπου δεδομένα πρέπει να υπάρχουν και εάν είναι τα κατάλληλα για όλα τα αντικείμενα.

Διπλότυπες τιμές

Συχνά σε ένα σετ δεδομένων κάποιο αντικείμενο υπάρχει πάνω από μία φορά και οφείλεται στην πολλαπλή υποβολή του ίδιου.

- Σχεδόν πάντα, διαγράφονται τα διπλότυπα προκειμένου να μην δημιουργήσουν προκατάληψη στους αλγόριθμους μηχανικής μάθησης.

3.1.1.2 Συγκέντρωση χαρακτηριστικών

Πραγματοποιείται ώστε να είναι τα δεδομένα σε μια πιο εύχρηστη μορφή.

- Μειώνεται ο χρόνος επεξεργασίας τους και η απαιτούμενη μνήμη.
- Δίνεται μια υψηλού επιπέδου οπτική, καθώς ομαδοποιούνται τα αντικείμενα.

3.1.1.3 Δειγματοληψία

Είναι μια συνηθισμένη μέθοδος να εργάζεται κανείς με ένα υποσύνολο των δεδομένων προκειμένου να κερδίσει σε χρόνο επεξεργασίας και χώρο αποθήκευσης. Δίνεται με αυτή τη μέθοδο η δυνατότητα να χρησιμοποιηθούν αποδοτικότεροι, αλλά ακριβότεροι αλγόριθμοι. Το καθοριστικό σημείο είναι η κατάλληλη επιλογή του δείγματος ώστε αυτό να αντιπροσωπεύει το αρχικό σετ δεδομένων. Υπάρχουν διάφορες στρατηγικές:

- *Απλή τυχαία δειγματοληψία* όπου υπάρχει ίση πιθανότητα επιλογής για κάθε αντικείμενο και πραγματοποιείται με ή χωρίς αντικατάσταση.

Οι δύο στρατηγικές που παρέχει η απλή τυχαία δειγματοληψία μπορούν να αποτύχουν όταν μια ομάδα αντικειμένων του σετ δεδομένων συναντάται σπάνια, για παράδειγμα σε μη ισορροπημένα σετ.

Μη ισορροπημένο σετ δεδομένων είναι αυτό που μια ή παραπάνω κατηγορίες του συναντώνται πολύ συχνά, με αποτέλεσμα άλλη ή άλλες να συναντώνται πολύ σπάνια.

Είναι μείζονος σημασίας οι σπάνιες κατηγορίες να αντιπροσωπεύονται επαρκώς στο δείγμα. Σε αυτή την περίπτωση χρησιμοποιείται η τεχνική που ονομάζεται *στρωματοποιημένη δειγματοληψία*, η οποία ξεκινά έχοντας ήδη ορίσει τις κατηγορίες. Υπάρχουν διάφορες μορφές αυτής της τεχνικής, με την πιο απλή να προτείνει την επιλογή ίσου αριθμού στιγμιότυπων από την κάθε κατηγορία, παρά το γεγονός πως κάθε μια έχει διαφορετικό πλήθος.

3.1.1.4 Μείωση διαστατικότητας

Οι περισσότερες βάσεις δεδομένων αποτελούνται από μεγάλο αριθμό χαρακτηριστικών. Η μείωση της διαστατικότητας έχει στόχο να μειώσει τα χαρακτηριστικά χωρίς όμως να επιλέξει ένα σύνολο από αυτά.

Η κατάρα της διαστατικότητας

Καθώς η διαστατικότητα των δεδομένων αυξάνεται, η ανάλυση τους γίνεται όλο και πιο σύνθετη, αφού είναι δύσκολο να μοντελοποιηθούν και να οπτικοποιηθούν.

Με τη μείωση της, προβάλλονται τα χαρακτηριστικά της βάσης σε μια χαμηλότερη διάσταση η οποία είναι οπτικοποιήσιμη (2 ή 3 διαστάσεις). Στόχος αποτελεί η δημιουργία νέων χαρακτηριστικών συνδυάζοντας τα παλιά. Τα κύρια πλεονεκτήματα της χαμηλότερης διαστατικότητας είναι:

- Οι αλγόριθμοι λειτουργούν αποδοτικότερα λόγω της εξάλειψης του θορύβου και άσχετων χαρακτηριστικών.
- Τα μοντέλα είναι πιο κατανοητά.
- Γίνεται εφικτή η οπτικοποίηση.

3.1.1.5 Κωδικοποίηση χαρακτηριστικών

Στόχος της προ-επεξεργασίας των δεδομένων, όπως ειπώθηκε νωρίτερα, είναι να κωδικοποιηθούν τα δεδομένα, ώστε να βρεθούν σε μορφή κατανοητή από τον υπολογιστή.

Η κωδικοποίηση χαρακτηριστικών είναι η μετατροπή των δεδομένων σε αποδεκτή μορφή εισόδου για τους αλγορίθμους μηχανικής μάθησης, ενώ διατηρούν τη σημασία τους.

Για *κατηγορικές* μεταβλητές (categorical variables) υπάρχουν δύο κατηγορίες:

- *Ονομαστική* (Nominal): ένα προς ένα χαρτογράφηση διατηρώντας τη σημασία της.
- *Διατάξιμη* (Ordinal): αλλαγή διατηρώντας την κατάταξη των τιμών.

Για *αριθμητικές* μεταβλητές (numerical variables):

- *Διαστηματική* (Interval): το αποτέλεσμα απλού μαθηματικού μετασχηματισμού.
- *Αναλογίας* (Ratio): αλλαγή μονάδας μέτρησης.

3.1.2 Εξόρυξη Δεδομένων

Η Εξόρυξη Δεδομένων ορίζεται ως μια διαδικασία εξαγωγής χρήσιμης γνώσης και μοτίβων από τεράστιο όγκο δεδομένων. Δημιουργώντας προγράμματα, τα οποία αναζητούν σε βάσεις δεδομένων, υπάρχει η δυνατότητα να ανακαλυφθούν ισχυρά μοτίβα, με τη βοήθεια των οποίων θα μπορούν να γενικευθούν πολλά δύσκολα προβλήματα και να πραγματοποιηθούν ακριβέστερες προβλέψεις σε μελλοντικές καταστάσεις. Ένα γνωστό παράδειγμα αποτελεί το πρόβλημα πρόβλεψης των καιρικών συνθηκών, για το οποίο οι Witten et al. (2005) στο βιβλίο τους έδειξαν πως χρησιμοποιώντας μόνο τέσσερα χαρακτηριστικά -ηλιοφάνεια, θερμοκρασία, υγρασία, άνεμο- μπορεί να βρεθεί ένα μοτίβο που να προβλέπει καιρικές συνθήκες. Μέσα από ένα απλό σετ κανόνων, μπορούν με ακρίβεια να ταξινομήσουν μια παρατήρηση σε ευνοϊκό ή όχι καιρό (Witten et al., 2005). Η μηχανική μάθηση αποτελεί τεχνική βάση για την εξόρυξη δεδομένων. Χρησιμοποιείται, συνήθως, για την εξαγωγή πληροφορίας από ανεπεξέργαστες βάσεις δεδομένων. Η διαδικασία ανακάλυψης αυτών των μοτίβων δεδομένων πρέπει να είναι αυτόματη ή ημι-αυτόματη (το οποίο συμβαίνει πιο συχνά), και τα ευρήματα πρέπει να είναι ουσιαστικά για να οδηγούν σε αξιοποιήσιμη γνώση. Μιας και οι δύο όροι συχνά σχετίζονται, είναι επίσης σημαντικό να γίνει κατανοητό πως η μηχανική μάθηση χρησιμοποιείται ως μαθηματικός αλγόριθμος για να δημιουργεί μοντέλα, ενώ η εξόρυξη δεδομένων ως η συνολική διαδικασία εξαγωγής γνώσης, η οποία μπορεί και να μη χρησιμοποιεί τεχνικές μηχανικής μάθησης.

Η εξόρυξη δεδομένων χρησιμοποιείται ευρέως από επιχειρήσεις, επιστημονική έρευνα, ακόμη και για την κυβερνητική ασφάλεια, αφού συνδυάζει μεθόδους μηχανικής μάθησης, στατιστική με διαχείριση βάσεων και ανάλυση δεδομένων. Παραδοσιακά, η εξόρυξη δεδομένων και η εξαγωγή γνώσης γίνονταν χειροκίνητα, ωστόσο, η διάδοση και η αυξανόμενη δύναμη της επιστήμης των υπολογιστών αύξησε δραματικά την ικανότητα συλλογής και αποθήκευσης. Οι βάσεις δεδομένων έχουν αυξηθεί σε μέγεθος και πολυπλοκότητα με αποτέλεσμα η άμεση χειροκίνητη ανάλυση δεδομένων να μην είναι αρκετή για τα σύγχρονα προβλήματα. Η αυτοματοποιημένη επεξεργασία δεδομένων βοηθούμενη από νέες ανακαλύψεις στον τομέα της επιστήμης των υπολογιστών, όπως τους γενετικούς αλγόριθμους (‘50), τη συσταδοποίηση (‘60), τα δένδρα αποφάσεων και τους κανόνες απόφασης(‘60), τα νευρωνικά δίκτυα (‘80), τις μηχανές διανυσμάτων υποστήριξης (‘90), έχουν εξοπλίσει τους επιστήμονες με υπερσύγχρονες τεχνικές προκειμένου να αντιμετωπίσουν τις σύνθετες, σύγχρονες προκλήσεις.

Εφαρμογές της εξόρυξης δεδομένων συναντώνται στην ιατροφαρμακευτική περίθαλψη, καθώς κρίνεται απαραίτητη η χρήση τους σε αυτό τον τομέα. Η αποτελεσματική αξιολόγηση θεραπειών συγκρίνοντας αιτίες, συμπτώματα και την πρόοδο ομάδων ασθενών που ακολουθούν διαφορετική φαρμακευτική θεραπεία για όμοια πάθηση καθορίζει ποια θεραπεία ταιριάζει καλύτερα σε κάθε ομάδα ασθενών (Kudyba and Stephan P, 2018). Επίσης, για την ενίσχυση της ιατροφαρμακευτικής διοίκησης και διαχείρισης, εφαρμογές εξόρυξης δεδομένων μπορούν να αναπτυχθούν προκειμένου να αναγνωρίζουν αποδοτικότερα χρόνιες παθήσεις και ασθενείς υψηλού κινδύνου, σχεδιάζοντας κατάλληλες παρεμβάσεις για τη μείωση των νοσοκομειακών εισαγωγών (Koh et al., 2011). Άλλες εφαρμογές εξόρυξης δεδομένων στην ιατροφαρμακευτική περίθαλψη περιλαμβάνουν αναγνώριση απατών, διαχείριση των πελατειακών σχέσεων, ακόμη και προγνωστική ιατρική.

Το μάρκετινγκ, επίσης, προσελκύει μεγάλη ανάπτυξη στον τομέα. Η πιο κοινή εφαρμογή εξόρυξης δεδομένων στο μάρκετινγκ πραγματοποιείται μέσω της κατάτμησης, η οποία δια μέσω της ανάλυσης των βάσεων πελατών επιτρέπει τον καθορισμό διαφορετικών ομάδων πελατών, ακόμη και την πρόβλεψη της συμπεριφοράς τους. Το πλήθος των δεδομένων που συλλέγονται έχουν τόση προοπτική που κάποια στιγμή, το Target (ένα αμερικάνικο κατάστημα λιανικής), αναγνώρισε μια νεαρή γυναίκα ως έγκυο πριν ακόμα μάθει ο πατέρας για την εγκυμοσύνη (Hill and Kashmir, 2012). Μια άλλη εφαρμογή του μάρκετινγκ στην εξόρυξη δεδομένων πραγματοποιείται δια μέσω της ανάλυσης προοπτικών της αγοράς, κατά την οποία αναγνωρίζονται μοτίβα στις καταναλωτικές συνήθειες των πελατών (Fayyad et al., 1996). Αυτό επιτρέπει την καλύτερη διαχείριση του αποθέματος και την κατανομή του χώρου αποθήκευσης στα σούπερ μάρκετ.

3.1.3 Ερμηνεία και Αξιολόγηση αποτελεσμάτων

Το τελευταίο βήμα στην ανακάλυψη γνώσης είναι η επαλήθευση πως τα μοτίβα που προέκυψαν από τους αλγόριθμους μηχανικής μάθησης αντιπροσωπεύουν το συνολικό σετ δεδομένων. Συχνά οι αλγόριθμοι ανακαλύπτουν μοτίβα τα οποία δεν είναι αντιπροσωπευτικά για το σύνολο των δεδομένων και το πρόβλημα αυτό ονομάζεται *υπερπροσαρμογή* (overfitting), δηλαδή ο αλγόριθμος δεν έχει γενικεύσει τα στιγμιότυπα, αντίθετα τα έχει μάθει σε βάθος χάνοντας τη γενική εικόνα. Για την τελική αξιολόγηση του αλγόριθμου, του δίνετε ένα σετ δεδομένων που συναντά πρώτη φορά και ανάλογα με τον αριθμό των σωστών ταξινομήσεων προκύπτει η ακρίβειά του.

3.2 Μηχανική Μάθηση

Τα τελευταία 60 χρόνια, η μηχανική μάθηση αναπτύχθηκε χάρη στις προσπάθειες του Arthur L.Samuel να εξερευνήσει, εάν οι μηχανές μπορούν να μάθουν να παίζουν παιχνίδια όπως οι άνθρωποι (Samuel, 1962), ώστε να διευρυνθεί η πειθαρχημένη διδασκαλία σε όλους τους τεχνολογικούς τομείς. Με την υπολογιστική ισχύ να αυξάνεται ραγδαία τις τελευταίες δεκαετίες, έγινε πιθανό όλες αυτές οι τεχνικές να εφαρμοστούν στην πράξη. Χρησιμοποιώντας τεχνικές όπως η παλινδρόμηση και οι μηχανές διανυσματικής υποστήριξης, η Google δημιούργησε την PageRank, την Google News και την ταξινόμηση των ανεπιθύμητων email της Gmail με αποτέλεσμα να εδραιωθεί ως μια από τις πιο ισχυρές εταιρίες παγκοσμίως. Αυτοί οι αλγόριθμοι διαδόθηκαν και οι νέες εφαρμογές που στηρίζονται σε αυτούς θεωρούνται πλέον συνηθισμένες.

Ο Tom M. Mitchell, υπεύθυνος του τμήματος μηχανικής μάθησης στο Carnegie Mellon University ορίζει τη μηχανική μάθηση ως «τη μελέτη υπολογιστικών αλγορίθμων που επιτρέπουν στον υπολογιστή να βελτιώνεται αυτόματα μέσω την εμπειρίας» (T.M. Mitchell, 1997).

Η μηχανική μάθηση έχει ένα πολύ ευρύ πεδίο εφαρμογών, καθώς καλύπτει διεργασίες μάθησης από αυτόνομα ρομπότ, μέχρι την εξόρυξη δεδομένων από αρχεία καταναλωτών για την πρόβλεψη της συμπεριφοράς τους και την αναζήτηση μηχανών που αυτόματα μαθαίνουν τις προτιμήσεις των χρηστών τους. Αποτελεί αποτέλεσμα της τομής της επιστήμης των υπολογιστών και τις στατιστικής και είναι η ικανότητα να εκπαιδευτεί μια μηχανή μέσα από εμπειρίες (δεδομένα) και πρωτότυπες ρυθμίσεις (αλγόριθμοι και οι παράμετροι τους).

«Η μηχανική μάθηση μελετά πώς να μαθαίνει αυτοματοποιημένα να πραγματοποιεί ακριβείς προβλέψεις βασισμένη σε προηγούμενες παρατηρήσεις» (Rob Schapire, 2015).

Ο σημερινός κόσμος είναι γεμάτος από εφαρμογές της μηχανικής μάθησης (σχεδόν) σε όλους τους τομείς της καθημερινότητας:

- Τράπεζες και άλλες επιχειρήσεις την αξιοποιούν ώστε να αναγνωρίσουν επενδυτικές ευκαιρίες και να ενημερώσουν τους επενδυτές πότε είναι η κατάλληλη στιγμή να δραστηριοποιηθούν. Χρησιμοποιώντας την εξόρυξη δεδομένων μπορούν επίσης να αναγνωριστούν πελάτες με προφίλ υψηλού κινδύνου ή να εντοπιστεί μια ενδεχόμενη απάτη.
- Ιστοσελίδες προωθούν προϊόντα που ενδεχομένως να ενδιαφέρουν των καταναλωτή βασιζόμενες σε προηγούμενες αγορές ή αναζητήσεις αξιοποιώντας παλαιότερες συμπεριφορές. Πιο πρόσφατα, η εξόρυξη κειμένου χρησιμοποιείται για τη σύγκριση βαθμολογίας και κειμένου κριτικής του καταναλωτή.
- Εταιρίες που προσφέρουν υπηρεσίες μεταφοράς πελατών χρησιμοποιούν τέτοιους αλγορίθμους προκειμένου να πραγματοποιούν αποδοτικότερα διαδρομές και να προβλέπουν ενδεχόμενα προβλήματα στοχεύοντας στην αύξηση των κερδών τους. Ακόμη, τα αυτοκίνητα χωρίς οδηγό χρειάζονται την μηχανική μάθηση για να προβλέπουν ατυχήματα και να βελτιστοποιούν δρομολόγια .
- Η βιομηχανία της υγείας τη χρησιμοποιεί ως εργαλείο για να βοηθήσει ιατρικές ομάδες να αναγνωρίσουν μοτίβα για τη σωστή διάγνωση ασθενειών. Πιο πρόσφατα, συσκευές που φοριούνται χρησιμοποιούν αισθητήρες για την παρακολούθηση της υγείας κάποιου σε πραγματικό χρόνο.

Η μηχανική μάθηση μπορεί να χωριστεί σε τέσσερις κατηγορίες μάθησης:

- *Επιβλεπόμενη* (supervised) περιλαμβάνει μαθηματικούς αλγορίθμους που δέχονται ως είσοδο δεδομένα τα οποία περιλαμβάνουν τόσο την είσοδο όσο και την επιθυμητή έξοδο.
- *Μη επιβλεπόμενη* (unsupervised) λαμβάνει δεδομένα ως είσοδο και αναζητά τη δομή τους μέσα από ομαδοποίηση και κατηγοριοποίηση.
- *Ημι-επιβλεπόμενη* (semi-supervised) είναι κάτι ανάμεσα στις προηγούμενες δύο κατηγορίες. Μερικά δεδομένα εισόδου έχουν την επιθυμητή ταμπέλα εξόδου και έχει αποδειχθεί πως αυξάνεται σημαντικά η ακρίβεια των αλγορίθμων.
- *Ενισχυμένη* (reinforcement) αφορά το πώς πρέπει οι πράκτορες να δρουν σε ένα περιβάλλον ώστε να έχουν τις μέγιστες ανταμοιβές.

Οι πρώτες δύο κατηγορίες είναι οι πιο δημοφιλείς, ενώ στην παρούσα εργασία εφαρμόζεται μόνο η πρώτη.

3.2.1 Επιβλεπόμενη μάθηση

Οι αλγόριθμοι επιβλεπόμενης μάθησης προβλέπουν βασιζόμενοι σε παραδείγματα. Έχουν x μεταβλητές εισόδου και μια μεταβλητή εξόδου y. Ο αλγόριθμος μαθαίνει να αντιστοιχίζει μια συνάρτηση (y=f(x)) και μπορεί να προβλέπει/ταξινομεί κάθε νέα έξοδο

αφού λάβει τις νέες εισόδους x. Οι πιθανές απαντήσεις της εξόδου είναι γνωστές. Όλα τα δεδομένα εισόδου έχουν ταμπέλα εξόδου και οι αλγόριθμοι μαθαίνουν να προβλέπουν την έξοδο από την είσοδο. Όλοι οι αλγόριθμοι επιβλεπόμενης μάθησης μπορούν να ταξινομηθούν σε προβλήματα ταξινόμησης ή παλινδρόμησης: Μια *συνάρτηση παλινδρόμησης* είναι αυτή που έχει ως μεταβλητή εξόδου μια πραγματική τιμή πχ. 88, 249, 14%. Μια *συνάρτηση ταξινόμησης* δημιουργεί μοντέλα όπου η έξοδος είναι μια κατηγορία πχ. κόκκινο/μπλε, αποκτήθηκε/δεν αποκτήθηκε.

Συχνά υπάρχει σύγχυση μεταξύ των όρων μηχανική μάθηση και ανάλυση δεδομένων. Ο πρώτος όρος αφορά μαθηματικούς αλγόριθμους που τρέχουν σε υπολογιστές για ένα πολύ ευρύ φάσμα εφαρμογών, μία από τις οποίες είναι η κατανόηση της δομής δεδομένων μιας βάσης. Αντίθετα, η ανάλυση δεδομένων είναι η επιστήμη που χρησιμοποιεί στατιστικά συμπεράσματα και αλγορίθμους μηχανικής μάθησης για την αναγνώριση μοτίβων σε μεγάλα όγκο δεδομένων.

Αφού η μηχανική μάθηση παρέχει τη δυνατότητα οι υπολογιστές να μαθαίνουν από παλαιότερες εμπειρίες ώστε να παρέχουν γνώση για το παρόν και το μέλλον, είναι φυσιολογικό να χρησιμοποιείται στις μελέτες εξαγοράς επιχειρήσεων. Παρακάτω θα δοθεί το θεωρητικό υπόβαθρο για τους αλγορίθμους που χρησιμοποιήθηκαν στην παρούσα εργασία.

3.3 Αλγόριθμοι

3.3.1 Λογιστική Παλινδρόμηση – Logistic Regression

Η Λογιστική Παλινδρόμηση χρησιμοποιείται για την απόκτηση ποσοστού αναλογίας πιθανότητας μεταξύ ενδεχόμενων αποτελεσμάτων βασιζόμενη σε δύο ή περισσότερες μεταβλητές. Ως αποτέλεσμα της προκύπτει η επίδραση κάθε μεταβλητής στον υπολογισμό του ποσοστού αναλογίας πιθανότητας των παρατηρούμενων γεγονότων ενδιαφέροντος. Το κύριο πλεονέκτημα της αποτελεί η αποφυγή σύγχυσης αναλύοντας τη συσχέτιση μεταξύ όλων των μεταβλητών μαζί. Επίσης, διαχειρίζεται συνεχείς μεταβλητές.

Η Λογιστική Παλινδρόμηση μοντελοποιεί την πιθανότητα ενός αποτελέσματος βασιζόμενη σε ξεχωριστά χαρακτηριστικά. Καθώς η πιθανότητα είναι μια αναλογία, αυτό που μοντελοποιείται είναι ο αλγόριθμος της πιθανότητας που δίνεται από:

$$\log\left(\frac{\Pi}{1-\Pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_\mu x_\mu \qquad (3.1)$$

όπου το Π είναι η πιθανότητα να συμβεί ένα γεγονός, το β_0 είναι το ύψος κλίσης της γραμμής παλινδρόμησης, τα υπόλοιπα β_i είναι οι συντελεστές παλινδρόμησης που δείχνουν τη συνεισφορά κάθε αντίστοιχης μεταβλητής x_i. Όταν ο συντελεστής είναι υψηλός σημαίνει πως η αντίστοιχη μεταβλητή επηρεάζει σημαντικά την έκβαση του αποτελέσματος, αντίθετα όταν είναι χαμηλή συνεισφέρει ελάχιστα.

Λύνοντας την 3.1 ως προς Π προκύπτει:

$$\Pi = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_\mu x_\mu)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_\mu x_\mu)} \qquad (3.2)$$

Στην 3.1 χρησιμοποιήθηκε ο μετασχηματισμός logit $\pi' = ln\left(\frac{\pi}{1-\pi}\right)$, καθώς η γραμμική παλινδρόμηση συναντά προβλήματα όπως τη μη κανονικότητα και τις άνισες

διασπορές των σφαλμάτων, καθώς και τον περιορισμό της συνάρτησης απόκρισης. Η σιγμοειδής μορφή της 3.2 διατηρεί τα πλεονεκτήματα της γραμμικότητας λύνοντας τα παραπάνω προβλήματα.

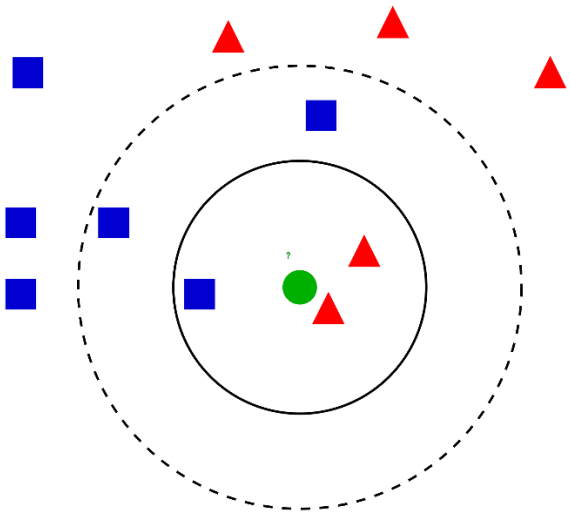
Κατά τη διαδικασία μάθησης ο αλγόριθμος αναθέτει ένα συντελεστή σε κάθε μεταβλητή ανάλογο με τη συνεισφορά της. Η ανάθεση πραγματοποιείται μέσω μιας επαναληπτικής διαδικασίας κατά την οποία, αρχικά ανατίθενται τυχαίες τιμές στους συντελεστές και στην πορεία επιδιώκεται η βελτιστοποίησή τους βάση μιας συνάρτησης σφάλματος. Ως συνάρτηση σφάλματος χρησιμοποιείται η εκτίμηση μέγιστης πιθανοφάνειας που παρουσιάστηκε παραπάνω. Μόλις ολοκληρωθεί η μάθηση, μέσω του τύπου 3.2 υπολογίζεται για κάθε εταιρία του σετ αξιολόγησης αν αυτή είναι επιτυχημένη (1) – αν η πιθανότητα Π είναι μεγαλύτερη του 0.5 – ή αποτυχημένη (0) διαφορετικά.

3.3.2 κ-Κοντινότεροι Γείτονες – k-Nearest Neighbors

Ο αλγόριθμος των κ-Κοντινότερων Γειτόνων θεωρείται ένας από τους πιο δημοφιλείς για προβλήματα ταξινόμησης προτύπων σε ομάδες. Είναι μη παραμετρικός, αφού δεν απαιτεί την εκμάθηση κάποιου συνόλου παραμέτρων, αντίθετα η μεταβλητή κ χρησιμοποιείται αποκλειστικά και συνήθως λαμβάνει την τιμή ενός μικρού περιττού αριθμού.

Τα δεδομένα αποθηκεύονται ως διανύσματα απαιτώντας ελάχιστο χρόνο για την «εκπαίδευση». Για την ταξινόμηση ενός νέου στοιχείου υπολογίζονται οι αποστάσεις του από τα κ κοντινότερα στοιχεία και λαμβάνει την τιμή της πλειοψηφίας. Οπότε, το κ είναι σημαντικός παράγοντας, αφού καθορίζει πόσα στοιχεία θα εξεταστούν.

Στο παρακάτω σχήμα υπάρχουν δύο κατηγορίες, τρίγωνα και τετράγωνα. Ο πράσινος κύκλος είναι το νέο στοιχείο που πρέπει να ταξινομηθεί. Αν επιλεγεί κ=3 τότε θα ταξινομηθεί ως τρίγωνο, ενώ για κ=5 ως τετράγωνο.



Σχήμα 3.1 – κ-Κοντινότεροι Γείτονες

3.3.3 Γκαουσιανός Αφελής Μπέυζ – Gaussian Naive Bayes

Ο Γκαουσιανός Αφελής Μπέυζ είναι ένας απλός αλγόριθμος που ανήκει στους πιθανοτικούς ταξινομητές και βασίζεται στην εφαρμογή του θεωρήματος του Μπέυζ θεωρώντας ισχυρή ανεξαρτησία μεταξύ των χαρακτηριστικών. Συγκεκριμένα,

σύμφωνα με το θεώρημα η πιθανότητα να συμβεί το ενδεχόμενο A δεδομένου ότι ισχύει το ενδεχόμενο B, προκύπτει από τον τύπο:

$$\mathrm{Prob}(A|B) = \frac{\mathrm{Prob}(B|A)\times\mathrm{Prob}(A)}{\mathrm{Prob}(B)} \qquad (3.3)$$

Όπου $\mathrm{Prob}(B|A)$ είναι η πιθανότητα να συμβεί το ενδεχόμενο B δεδομένου ότι ισχύει το A.

Αντίστοιχα, χρησιμοποιείται το θεώρημα Bayes για την ταξινόμηση ενός διανύσματος εισόδου. Έστω, ότι τα χαρακτηριστικά κάθε στοιχείου είναι N το πλήθος, δηλαδή:

$$X = (x_1, x_2, \dots, x_N)$$

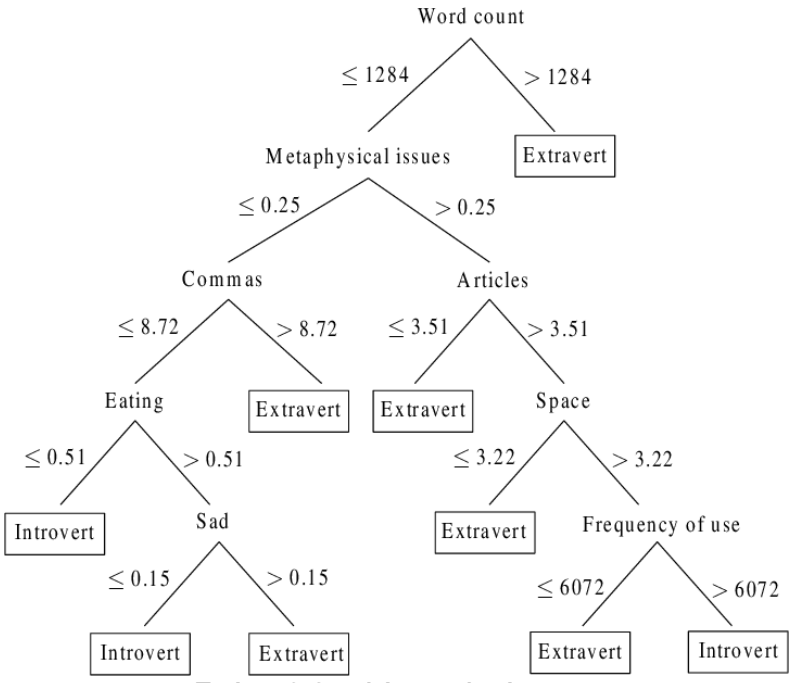
τότε η πιθανότητα που έχει το στοιχείο να ανήκει στην κλάση y προκύπτει από τη σχέση:

$$\mathrm{Prob}(y|x_1, x_2, \dots, x_N) = \frac{\mathrm{Prob}(x_1|y)\mathrm{Prob}(x_2|y)\dots\mathrm{Prob}(x_N|y))}{\mathrm{Prob}(x_1)\mathrm{Prob}(x_2)\dots\mathrm{Prob}(x_N)} \qquad (3.4)$$

όπου οι πιθανότητες $\mathrm{Prob}(x_1|y), \mathrm{Prob}(x_1)$ υπολογίζονται εύκολα από τα δεδομένα εκπαίδευσης. Η κάθε κατηγορία θα έχει διαφορετική πιθανότητα και προφανώς η είσοδος θα ταξινομηθεί σύμφωνα με τη μεγαλύτερη. Βασική προϋπόθεση του θεωρήματος είναι η στατική ανεξαρτησία μεταξύ των χαρακτηριστικών.

3.3.4 Τυχαία Δάση – Random Forest

Ο αλγόριθμος Τυχαία Δάση αποτελείται από ένα σύνολο πολλαπλών δέντρων αποφάσεων, τα οποία ενώνονται προκειμένου να αυξηθεί η ακρίβεια των προβλέψεων. Ένα δέντρο αποφάσεων διαχωρίζει συνεχώς τα δεδομένα εκπαίδευσης σε υποσύνολα μέχρι να φτάσει στο χαμηλότερο επίπεδο, τα φύλλα, όπου θα πραγματοποιηθεί η λήψη της απόφασης.



Σχήμα 3.2 – Δέντρο Απόφασης

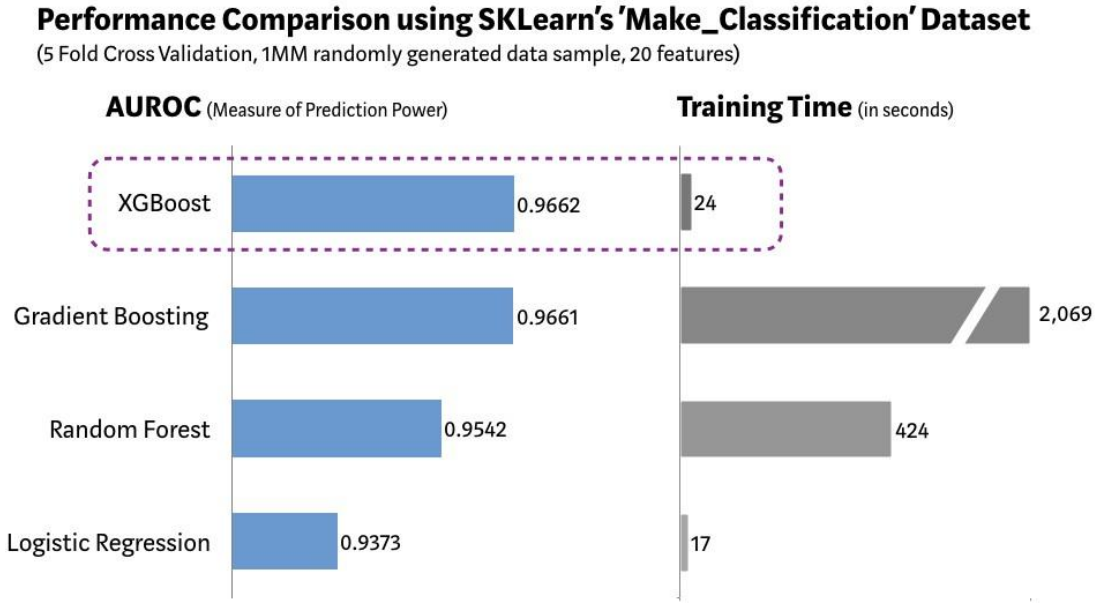
Συγκεκριμένα, κατά τον αλγόριθμο των Τυχαίων Δασών επιλέγονται τυχαία δείγματα με επανατοποθέτηση για τη δημιουργία του καθενός από τα Ν το πλήθος Δέντρα Αποφάσεων. Δημιουργώντας πολλά ανεξάρτητα Δέντρα Αποφάσεων εξαλείφεται η προκατάληψη (bias), αφού σημασία έχει σε ποια κλάση ταξινομήθηκε το αντικείμενο από την πλειοψηφία. Συνεπώς, υπάρχει η δυνατότητα λάθους χωρίς να κοστίσει, δίνοντας την ευκαιρία για τη δημιουργία ενός μοντέλου με υψηλή ικανότητα γενίκευσης.

3.3.5 XGBoost

Ο αλγόριθμος XGBoost (eXtreme Gradient Boosting) αποτελεί την ενισχυμένη κλίση δέντρων αποφάσεων που προσφέρει μεγαλύτερη ταχύτητα εκπαίδευσης και υψηλότερη απόδοση. Αποτελεί εξέλιξη της μεθόδου Τυχαίων Δέντρων (παρουσιάστηκε παραπάνω, ενότητα 3.3.4) χρησιμοποιώντας την ίδια φιλοσοφία, αλλά με μερικές προσθήκες που αυξάνουν σημαντικά την αποτελεσματικότητα.

Όσων αφορά την *ταχύτητα εκπαίδευσης* το σύστημα βελτιώθηκε μέσω της παραλληλοποίησης και του κλαδέματος των δέντρων (tree pruning) μειώθηκε η πολυπλοκότητα, ενώ ταυτόχρονα οι υπολογιστικοί πόροι αξιοποιήθηκαν αποτελεσματικότερα.

Για την *αυξημένη απόδοση* ευθύνονται οι αλγοριθμικές βελτιώσεις. Η ομαλοποίηση τιμωρεί με πιο σύνθετο τρόπο αποφεύγοντας την υπερπροσαρμογή (overfitting), η αραιότητα των χαρακτηριστικών αντιμετωπίζεται αποτελεσματικότερα και με τη χρήση του αλγορίθμου Weighted Quantile Sketch εντοπίζονται τα βέλτιστα σημεία διαχωρισμού μεταξύ των υποσυνόλων.



Σχήμα 3.3 – Σύγκριση αλγορίθμων υλοποιημένων από το SkLearn

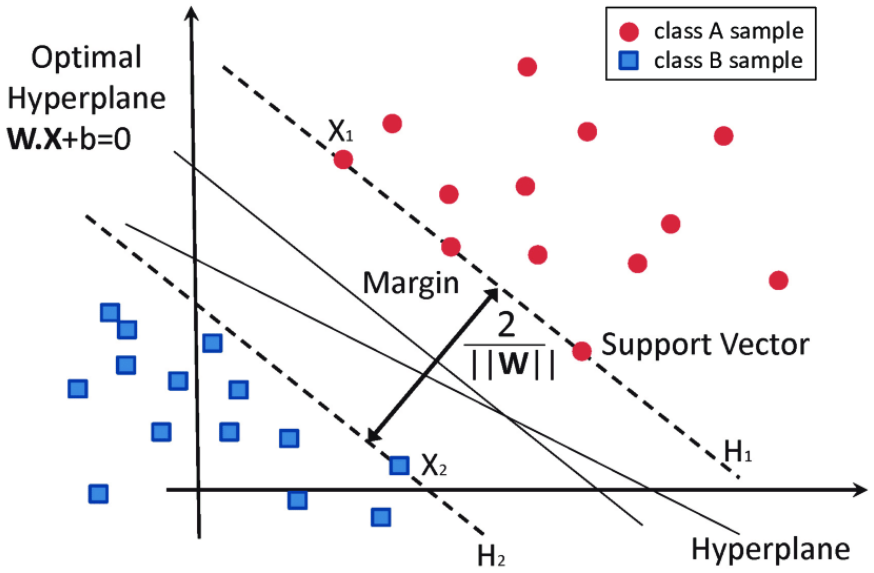
3.3.6 Μηχανή Διανυσματικής Υποστήριξης – Support Vector Machine

Ο αλγόριθμος βασίζεται στην ιδέα της εύρεσης υπερεπιπέδου με σκοπό να διαχωριστούν οι κλάσεις με βέλτιστο τρόπο, δηλαδή τη μεγιστοποίηση του περιθωρίου ανάμεσα στα αντικείμενα διαφορετικών κλάσεων. Για δεδομένα Ν χαρακτηριστικών, το υπερεπίπεδο είναι Ν-1 διαστάσεων και προκύπτει από την:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N = 0$$

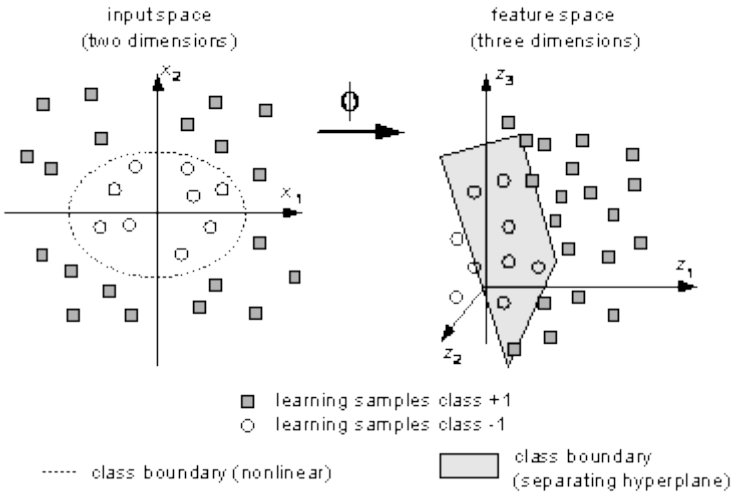
Εάν τα δεδομένα εκπαίδευσης είναι γραμμικώς διαχωρίσιμα, θα διαχωρίζονται μέσω ενός υπερεπιπέδου και το νέο αντικείμενο που πρέπει να ταξινομηθεί θα εξετάζεται σε ποια πλευρά του υπερεπιπέδου ανήκει.

Όταν οι κλάσεις είναι *γραμμικώς διαχωρίσιμες*, τότε υπάρχουν άπειρες ευθείες που τις διαχωρίζουν, ωστόσο μόνο μία το πραγματοποιεί βέλτιστα, μεγιστοποιώντας το περιθώριο (margin όπως φαίνεται στο σχήμα).



Σχήμα 3.4 – SVM

Όταν τα δεδομένα είναι *γραμμικώς μη διαχωρίσιμα*, χρησιμοποιείται μια συνάρτηση πυρήνα (kernel function) όπου απεικονίζει τα δεδομένα σε έναν χώρο μεγαλύτερων διαστάσεων στον οποίο οι κλάσεις είναι γραμμικώς διαχωρίσιμες.

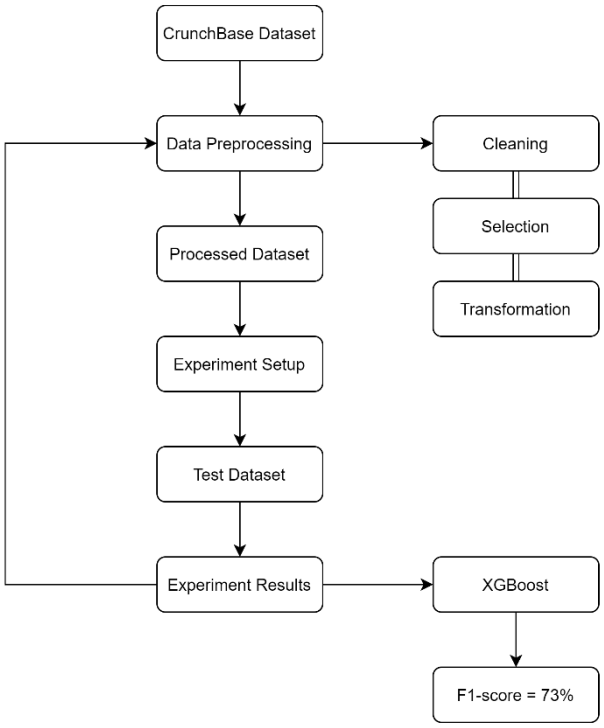


Σχήμα 3.5 – Συνάρτηση πυρήνα

Κεφάλαιο 4: Πειραματική διαδικασία και αποτελέσματα

4.1 Μεθοδολογία

Η μεθοδολογία (Σχήμα 4.1) που εφαρμόζεται σε αυτή την εργασία προσομοιάζει την προσέγγιση της Ανακάλυψης Γνώσης σε Βάσεις (Knowledge Discovery in Databases, KDD): **(1) Επιλογή** δεδομένων που θα επεξεργαστούν από ολόκληρη τη βάση του CrunchBase, **(2) Προ-επεξεργασία**, με καθάρισμα, επιλογή και μετασχηματισμό δεδομένων. Σε αυτό το στάδιο αντιμετωπίζονται ελλιπείς και ακραίες τιμές, γίνεται κωδικοποίηση των ετικετών καθώς λύνονται και άλλα προβλήματα. Μια εξερευνητική ανάλυση πραγματοποιείται πριν τους περεταίρω μετασχηματισμούς, **(3) Σύνθεση πειράματος**, όπου ορίζονται οι μετρικές αξιολόγησης και αντιμετωπίζονται δύο βασικά προβλήματα της βάσης δεδομένων, η ανεπάρκεια ορισμένων χαρακτηριστικών και η μη ισορροπία των (δύο) κατηγοριών στόχων, Πολλοί αλγόριθμοι μηχανικής μάθησης επιλέγονται για να δοκιμαστούν οι δυαδικοί ταξινομητές και να ταξινομήσουν τις παρατηρήσεις ως «επιτυχημένες» ή «μη επιτυχημένες», **(4) Αποτελέσματα του πειράματος**, όπου βγαίνουν συμπεράσματα και ερμηνεύονται τα αποτελέσματα.



Σχήμα 4.1 – Σύνοψη Μεθοδολογίας

4.1.1 Συλλογή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία ανήκουν στην βάση δεδομένων CrunchBase, λήφθηκαν από το github https://github.com/notpeter/crunchbase-data και έχουν εξαχθεί από τη βάση στις 5-12-2015.

Η εγκυρότητα των δεδομένων είναι αναμφίβολη καθώς πολλές συμβουλευτικές και εταιρίες επενδυτικού κεφαλαίου τα έχουν χρησιμοποιήσει, ενώ και προηγούμενες

μελέτες έχουν βασιστεί σε αυτά. Η ιστοσελίδα αποτελεί βάση δεδομένων για όλες τις νεοφυείς εταιρίες και τους επενδυτές.

4.1.2 Προ-επεξεργασία Δεδομένων

«Εάν υπάρχει πολύ άσχετη και περιττή πληροφορία, θόρυβος ή αναξιόπιστα δεδομένα, η ανακάλυψη γνώσης κατά τη διάρκεια της εκπαίδευσης γίνεται πιο δύσκολη»

(Kotsiantis, Kanellopoulos, & Pintelas, 2006)

Η προ-επεξεργασία συχνά έχει καθοριστική σημασία στην επίδοση της επιβλεπόμενης μηχανικής μάθησης. Θα ακολουθήσουν γενικές αλλαγές. Η φύση των δεδομένων είναι τέτοια που προτεραιότητα του προβλήματος αποτελεί η κατανόηση της ανεξαρτησίας τους και όχι η μείωση της συσχέτισής τους.

Η προ-επεξεργασία δεδομένων αποτελείται από διαδικασία τριών σταδίων:

- *Καθαρισμός δεδομένων*, όπου στόχος είναι η απομάκρυνση περιττής και άσχετης πληροφορίας από τη βάση όπως διπλότυπα, ελλιπείς και ακραίες τιμές.
- *Επιλογή δεδομένων*, όπου το πεδίο έρευνας καθορίζει ποιες χαρακτηριστικά θα επιλεχθούν για το σχηματισμό των τελικών δεδομένων.
- *Μετασχηματισμός δεδομένων*, αποτελείται από τη διαδικασία δημιουργίας νέων μεταβλητών ή τη συγκέντρωση δεδομένων από διαφορετικούς πίνακες.

4.1.2.1 Καθαρισμός Δεδομένων

«Ο καθαρισμός δεδομένων είναι μια χρονοβόρα και έντονα κοπιαστική διαδικασία, αλλά είναι απολύτως αναγκαία για την επιτυχημένη εξόρυξη δεδομένων»

(Yeates et al., 2000)

Χαρακτηριστικό	**Τύπος**
URL εταιρίας	Αλφαριθμητικό
Όνομα εταιρίας	Αλφαριθμητικό
Διεύθυνση ιστοσελίδας εταιρίας	Αλφαριθμητικό
Κατηγορίες δραστηριοποίησης επιχείρησης	Κατηγορία
Συνολικό ποσό χρηματοδότησης	Αριθμός
Κατάσταση επιχείρησης	Κατηγορία
Χώρα ίδρυσης επιχείρησης	Κατηγορία
Πολιτεία ίδρυσης επιχείρησης	Κατηγορία

Περιοχή ίδρυσης επιχείρησης	Κατηγορία
Πόλη ίδρυσης επιχείρησης	Κατηγορία
Γύροι χρηματοδότησης	Αριθμός
Ημερομηνία ίδρυσης	Ημερομηνία
Ημερομηνία 1ης χρηματοδότησης	Ημερομηνία
Ημερομηνία τελευταίας χρηματοδότησης	Ημερομηνία

Πίνακας 4.1 – Χαρακτηριστικά Επιχειρήσεων

Το *πρώτο βήμα* της προ-επεξεργασίας είναι η απαλλαγή από άσχετη και περιττή πληροφορία. Καθώς στη βάση του CrunchBase σημειώνονται πολλά χαρακτηριστικά και παρατηρήσεις που δεν σχετίζονται με την πρόβλεψη της μελλοντικής επιτυχίας νεοφυών επιχειρήσεων.

Από τα χαρακτηριστικά των επιχειρήσεων διαγράφηκαν:

- το URL της ιστοσελίδων, ο κωδικός της πολιτείας και η πόλη ίδρυσης, καθώς αποτελούν περιττές λεπτομέρειες.

Είναι επίσης σημαντική η διαγραφή των διπλότυπων.

Το *δεύτερο βήμα* είναι η εξάλειψη θορύβου και αναξιόπιστων δεδομένων που αποτελεί καθοριστικό παράγοντα για την εξαγωγή συμπερασμάτων. Οι δύο πιο συχνές περιπτώσεις είναι οι ελλιπείς και οι ακραίες τιμές.. «Οι περισσότερες μέθοδοι μηχανικής μάθησης πραγματοποιούν έμμεσα υποθέσεις και δεν δίνεται μεγάλη σημασία στο γεγονός ότι λείπει η τιμή σε κάποιο χαρακτηριστικό: η τιμή είναι απλώς άγνωστη» (Yeates et al., 2000).

Ακολουθώντας τη λογική «όσο μειώνεται το πλήθος των δεδομένων, τόσο αυξάνεται η ακρίβειά τους» (Kotsiantis et al., 2006), διαγράφηκαν τα στιγμιότυπα που τους έλλειπαν τιμές στα παρακάτω γνωρίσματα:

- κατάσταση εταιρίας, συνολικό ποσό χρηματοδότησης.

Διαγράφηκαν στιγμιότυπα με ακραίες τιμές στα χαρακτηριστικά:

- συνολικό ποσό χρηματοδότησης, πλήθος γύρων χρηματοδότησης.

4.1.2.2 Επιλογή Δεδομένων

Τα δεδομένα αποτελούνταν από 5 πίνακες που παρουσιάζονται παρακάτω:

Όνομα	**Παρατηρήσεις**	**Επιλέχθηκε**
Εξαγορές	18968	✓
Προσθήκες	2213	

Επιχειρήσεις	66368	✓
Επενδύσεις	168635	✓
Γύροι	114949	✓

Πίνακας 4.2 – Βάση CrunchBase

Τα χαρακτηριστικά των δεδομένων απεικονίζονται στον παρακάτω πίνακα:

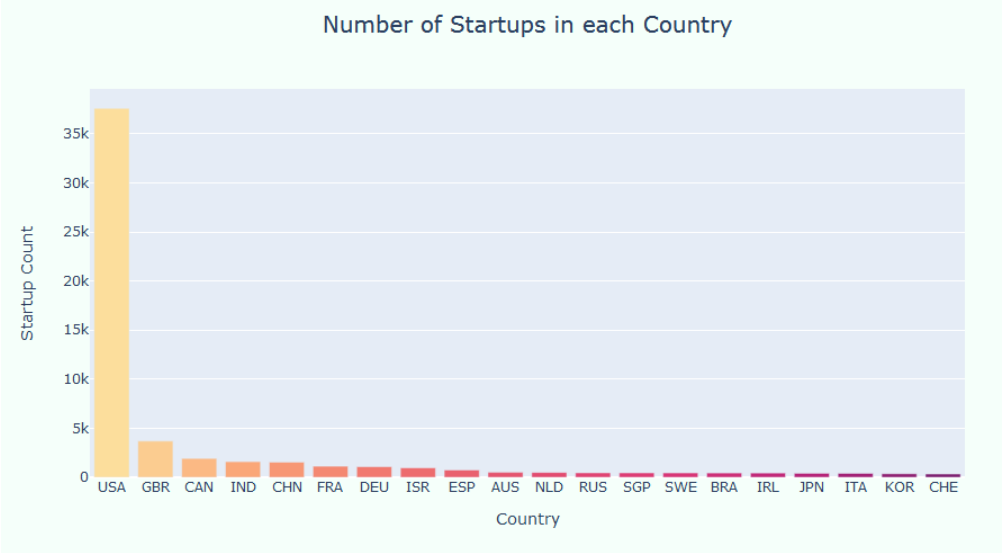
Χαρακτηριστικό	**Επιλέχθηκε**
URL εταιρίας	
Όνομα εταιρίας	
Κατηγορίες δραστηριοποίησης εταιρίας	✓
Χώρα που ιδρύθηκε η εταιρία	✓
Πολιτεία ίδρυσης εταιρίας	
Περιοχή ίδρυσης εταιρίας	
Πόλη ίδρυσης εταιρίας	
URL αγοραστή	
Όνομα αγοραστή	✓
Κατηγορίες δραστηριοποίησης αγοραστή	
Χώρα αγοραστή	
Πολιτεία αγοραστή	
Περιοχή αγοραστή	
Πόλη αγοραστή	
Ημερομηνία αγοράς	
Μήνας αγοράς	
Ποσό αγοράς	
Νόμισμα συναλλαγής	
Συνολικό ποσό χρηματοδότησης	✓
Κατάσταση εταιρίας	✓

Συνολικοί γύροι χρηματοδότησης	✓
Ημερομηνία ίδρυσης εταιρίας	✓
Ημερομηνία πρώτης χρηματοδότησης	✓
Ημερομηνία τελευταίας χρηματοδότησης	✓
URL επενδυτή	
Όνομα επενδυτή	✓
Χώρα επενδυτή	
Πολιτεία επενδυτή	
Περιοχή επενδυτή	
Πόλη επενδυτή	
Τύπος χρηματοδότησης	✓
Κωδικός χρηματοδότησης	✓
Ημερομηνία χρηματοδότησης	✓
Ποσό χρηματοδότησης	✓

Πίνακας 4.3 – Χαρακτηριστικά της βάσης CrunchBase

Σημείωση: Για την δημιουργία του συνόλου δεδομένων που εκπαιδεύονται τα μοντέλα χρησιμοποιήθηκαν μόνο τα «Επιλεγμένα» χαρακτηριστικά.

4.1.2.3 Ανάλυση Επιλεγμένων Δεδομένων



Σχήμα 4.2 – Οι 20 χώρες με τις περισσότερες νεοφυείς επιχειρήσεις

Στο πλαίσιο της παρούσας μελέτης, επιλέχθηκαν εταιρίες από ολόκληρο τον κόσμο για την εκμάθηση των μοντέλων. Για την παροχή πληροφορίας ως προς τη χώρα ίδρυσης και δραστηριοποίησης της κάθε εταιρίας επιλέχθηκαν όλες οι χώρες, οι οποίες στη συνέχεια κατηγοριοποιήθηκαν στις εννέα δημοφιλέστερες και στις έξι ηπείρους που ανήκουν.

Χώρα	**Πλήθος**	**Ποσοστό επί του πλήθους**	**Επιτυχημένες**	**Ποσοστό επιτυχημένων (στη χώρα)**
ΗΠΑ	31429	59%	4767	15%
Μ. Βρετανία	3136	6%	217	7%
Καναδάς	1579	3%	217	14%
Ινδία	1181	2%	47	4%
Κίνα	1219	2%	94	8%
Γαλλία	983	2%	80	8%
Γερμανία	669	1%	64	10%
Ισραήλ	819	2%	120	15%
Ισπανία	634	1%	32	5%
Άλλη/Άγνωστο	11934	22%	593	5%

Πίνακας 4.4 – Χώρες ίδρυσης των επιχειρήσεων

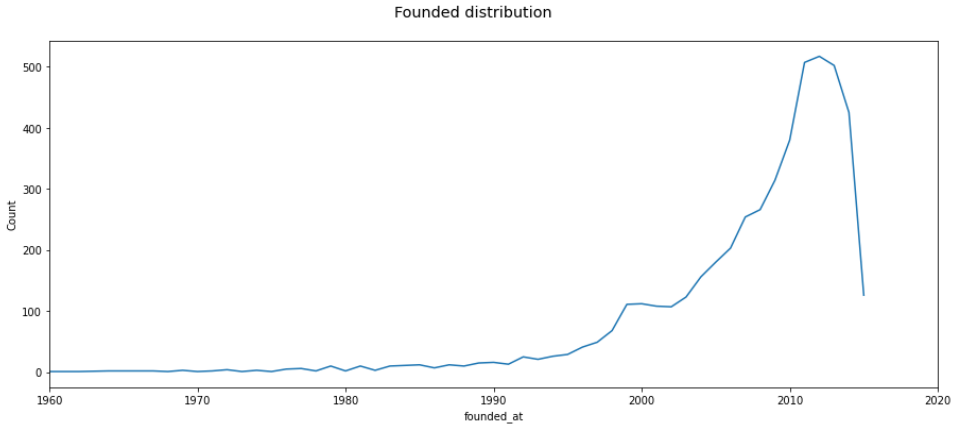
Ήπειρος	**Πλήθος**	**Ποσοστό επί του πλήθους**	**Επιτυχημένες**	**Ποσοστό επιτυχημένων**
Β. Αμερική	33189	62%	4989	15%
Ευρώπη	9144	17%	608	7%
Ασία	4980	9%	342	7%
Ν. Αμερική	945	2%	24	3%
Αυστραλία	501	1%	231	46%
Αφρική	207	0.4%	35	17%
Άγνωστη	4617	8.6%	0	0%

Πίνακας 4.5 – Ήπειροι ίδρυσης των επιχειρήσεων

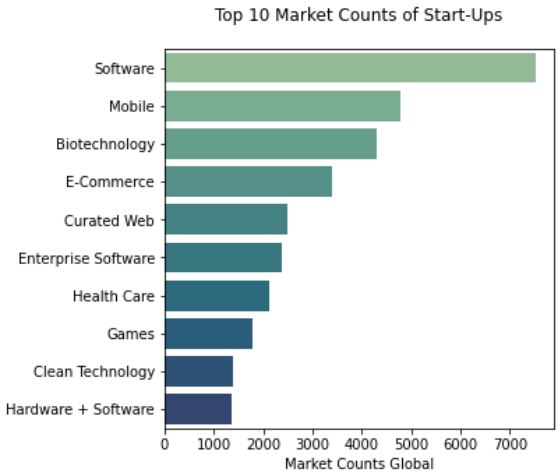
Εταιρίες που ιδρύθηκαν πριν το 2005: Παρόλο που αυτές οι εταιρίες δεν μπορούν να θεωρηθούν νεοφυείς λόγω της προχωρημένης ηλικίας τους χωρίς επιτυχία, κάποια στιγμή είχαν χαρακτηριστικά όπως γύροι χρηματοδότησης που τις έφεραν πιο κοντά στην επιτυχία, με αποτέλεσμα αυτές οι εταιρίες να παραμένουν στο σετ δεδομένων. Η λογική αυτής της επιλογής είναι η υπόθεση πως οι εταιρίες χρειάζονται χρόνο για να ωριμάσουν και να πετύχουν πρόοδο. Την ίδια στιγμή, καλύπτονται η «Dot-com bubble» του 1997 και η Παγκόσμια κρίση του 2008.

Ημερομηνία ίδρυσης	**Πλήθος**	**Ποσοστό επί του πλήθους**	**Επιτυχημένες**	**Ποσοστό επιτυχημένων (στην κατηγορία)**
Πριν το 2005	8571	16%	2494	29%
Μεταξύ 2005 και 2010	13765	26%	1791	13%
Μετά το 2010	19555	36%	608	3%
Άγνωστο	11692	22%	1338	11%

Πίνακας 4.6 – Ημερομηνίες ίδρυσης των επιχειρήσεων (κατηγορίες)



Σχήμα 4.3 – Ημερομηνίες ίδρυσης των επιχειρήσεων



Σχήμα 4.4 – Οι 10 δημοφιλέστερες κατηγορίες δραστηριοποίησης για νεοφυείς επιχειρήσεις

Εταιρίες με κατηγορία: Προκειμένου να συγκριθούν τα αποτελέσματα της παρούσας μελέτης με προηγούμενες δημοσιεύσεις, ο τομέας δραστηριοποίησης της κάθε εταιρίας αποτέλεσε κρίσιμο παράγοντα για την επιτυχίας της. Συνεπώς, επιλέχθηκαν εταιρίες με τουλάχιστον έναν τομέα δραστηριοποίησης για περεταίρω ανάλυση, ώστε να καθίσταται σαφές αν μια εταιρία ανήκει ή όχι στην τεχνολογική βιομηχανία.

Μετά τα φιλτραρίσματα που εφαρμόστηκαν, τα στιγμιότυπα των εταιριών είναι 53583.

4.1.2.4 Μετασχηματισμός Δεδομένων

Ο μετασχηματισμός των δεδομένων μπορεί να θεωρηθεί ως «η εφαρμογή μαθηματικών τροποποιήσεων στις τιμές μιας μεταβλητής» για να εξαχθεί περισσότερη αξία από την αρχική (Osborne, 2002). Στην παρούσα εργασία, ο μετασχηματισμός δεδομένων μπορεί να χωριστεί σε δύο διαδοχικά στάδια:

1. Αλλαγές στα πρωτότυπα δεδομένα
2. Δημιουργία νέων μεταβλητών

Αλλαγές στα πρωτότυπα Δεδομένα

Αυτές οι αλλαγές εφαρμόστηκαν με συνέπεια σε όλες τις μεταβλητές

- Ημερομηνία ίδρυσης εταιρίας: Οι εταιρίες κατηγοριοποιήθηκαν σε 3 χρονολογικές κατηγορίες (πριν_το_2005, μεταξύ_2005_και_2010, μετά_το_2010) και όσες δεν είχαν ημερομηνία ίδρυσης στο Άλλο.

Νέες Μεταβλητές

Χρησιμοποιώντας την παρούσα πληροφορία δημιουργήθηκαν οι παρακάτω μεταβλητές. Να σημειωθεί πως για τα περισσότερα χαρακτηριστικά έπρεπε να αντιμετωπιστεί το πρόβλημα των ελλιπών ή αραιών δεδομένων, το οποίο επηρεάζει τους αλγόριθμους μηχανικής μάθησης και απαιτεί ειδική μεταχείριση.

Χαρακτηριστικό	**Περιγραφή**	**Μέσος όρος**
Συνολικό ποσό χρηματοδότησης	Το συνολικό ποσό με το οποίο χρηματοδοτήθηκε η επιχείρηση	18,478,604
Γύροι χρηματοδότησης	Το πλήθος των γύρων χρηματοδότησης της επιχείρησης	1.88
ΗΠΑ	1 αν η επιχείρηση ιδρύθηκε εκεί, διαφορετικά 0	
Μ. Βρετανία	>>	
Καναδάς	>>	
Ινδία	>>	
Κίνα	>>	

Γαλλία	>>	
Γερμανία	>>	
Ισραήλ	>>	
Ισπανία	>>	
Β. Αμερική	>>	
Ν. Αμερική	>>	
Ευρώπη	>>	
Αφρική	>>	
Ασία	>>	
Αυστραλία	>>	
Άγνωστη χώρα	1 αν η χώρα ίδρυσης της επιχείρηση είναι άγνωστη, διαφορετικά 0	
Λογισμικό	1 αν η επιχείρηση δραστηριοποιείται στον συγκεκριμένο τομέα, διαφορετικά 0	
Βιοτεχνολογία	>>	
Ηλεκτρονικό εμπόριο	>>	
Κινητά	>>	
Τεχνολογία καθαρισμού	>>	
Ταξινόμηση διαδικτύου	>>	
Υλικό και Λογισμικό	>>	
Ιατρική περίθαλψη	>>	
Παιχνίδια	>>	
Λογισμικό για επιχειρήσεις	>>	
Άλλο τομέα	1 αν η επιχείρηση δραστηριοποιείται σε τομέα που δεν ανήκει στους 10 δημοφιλέστερους, διαφορετικά 0	

Α	Συνολικό ποσό χρηματοδότησης τέτοιου κωδικού	1,665,638
Β	>>	2,214,774
Γ	>>	1,684,205
Δ	>>	1,094,267
Ε	>>	672,113
ΣΤ	>>	261,985
Ζ	>>	101,991
Η	>>	31,485
Τολμηρή	Τύπος χρηματοδότησης	0.98
Σπόρου	>>	0.48
Χρεών	>>	0.13
Άγγελος	>>	0.1
Κρυφή	Που δεν έχει δημοσιευθεί	0.05
Συλλογικής ισότητας	Τύπος χρηματοδότησης	0.02
Ιδιωτικής ισότητας	>>	0.04
Επιχορήγησης	>>	0.04
Μετατρέψιμη	>>	0.03
Ισότητας μετά το ipo	>>	0.01
Συλλογική για το προϊόν	>>	0.01
Άνισης βοήθειας	>>	0.002
Χρέους μετά το ipo	>>	0.003
Δευτερεύουσας αγοράς	>>	0.002
Επενδυτές	Πλήθος επενδυτών που χρηματοδότησαν την επιχείρηση	2.24
Επιτυχίες επενδυτών	Πλήθος επιτυχιών των επενδυτών που	50.76

	χρηματοδότησαν την επιχείρηση	
Εμπειρία επενδυτών	Πλήθος επιχειρήσεων που έχουν επενδύσει οι επενδυτές που χρηματοδότησαν την επιχείρηση	160
Αγοραστές	Πλήθος αγοραστών που χρηματοδότησαν την επιχείρηση	0.17
Εμπειρία αγοραστών	Πλήθος επιχειρήσεων που έχουν αγοράσει οι αγοραστές που αγόρασαν την επιχείρηση	0.31
Πριν το 2005	1 αν έχει ιδρυθεί τη δεδομένη χρονική περίοδο, διαφορετικά 0	
2005 με 2010	>>	
Μετά το 2010	>>	
Άγνωστη ημερ. ίδρυσης	1 αν είναι άγνωστη η ημερομηνία ίδρυσης της επιχείρησης	
Χρόνος μέχρι την πρώτη χρηματοδότηση	Το χρονικό διάστημα (σε μήνες) από την ίδρυση μέχρι την πρώτη χρηματοδότηση της επιχείρησης	35
Άγνωστη χρηματοδότηση	1 αν λείπει η ημερομηνία ίδρυσης ή 1ης χρηματοδότησης	
Χρόνος μέχρι τελευταία χρηματοδότηση	Το χρονικό διάστημα (σε μήνες) από την 1η μέχρι την τελευταία χρηματοδότηση της επιχείρησης	12

Πίνακας 4.7 – Χαρακτηριστικά που χρησιμοποιήθηκαν

Πρόβλημα μη ισορροπημένων δεδομένων

Ένα μικρό ποσοστό των επιχειρήσεων πετυχαίνει, με αποτέλεσμα το σύνολο των δεδομένων να αποτελείται από ένα μικρό ποσοστό επιτυχημένων εταιριών (11.6%). Όμως, με τόσα λίγα παραδείγματα επιτυχημένων εταιριών, οι αλγόριθμοι δυσκολεύονται να αναγνωρίσουν το μοτίβο τους. Το συγκεκριμένο πρόβλημα έχει πολλές λύσεις, ωστόσο αυτή που δόθηκε στην παρούσα εργασία είναι η

υπερδειγματοληψία (oversampling). Υλοποιήθηκαν δύο τεχνικές της βιβλιοθήκης sklearn:

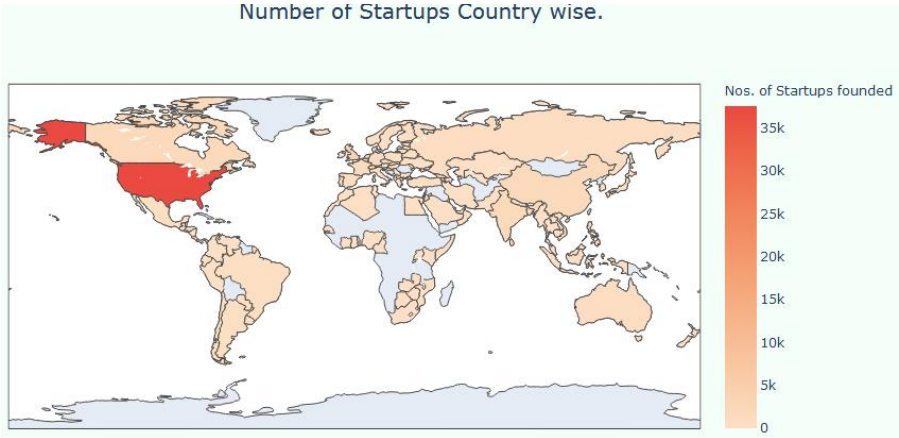
- SMOTE
- Oversampling

Ύστερα από δοκιμές, έγινε φανερό πως η πρώτη είχε σημαντικά καλύτερα αποτελέσματα με στρατηγική δειγματοληψίας ίση με 0.5.

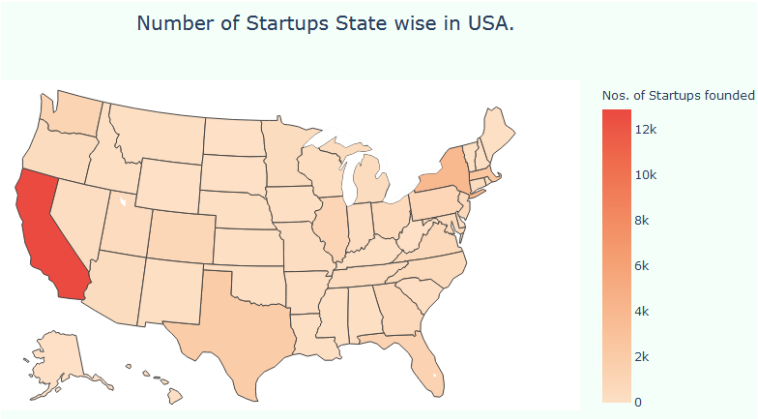
4.2 Ανάλυση Βάσης Δεδομένων

4.2.1 Χώρα ίδρυσης

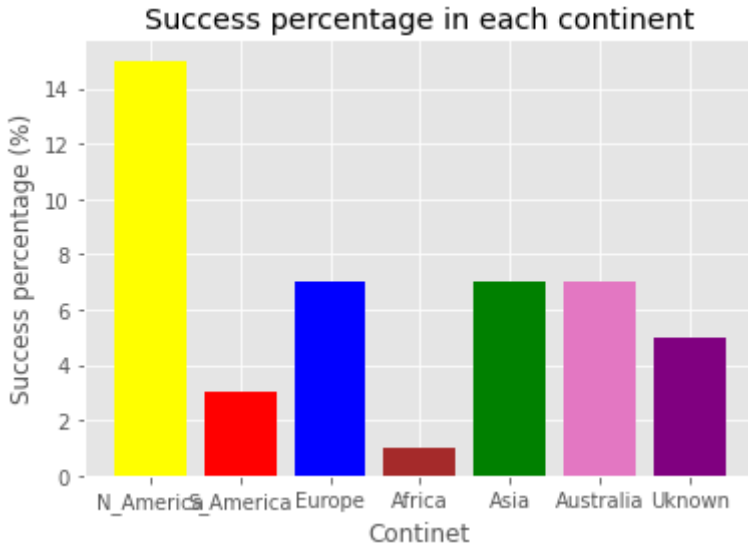
Επιλέχθηκαν εταιρίες από ολόκληρο τον κόσμο προκειμένου να αναδειχθούν τα διαφορετικά κριτήρια στην εκάστοτε περιοχή. Γίνεται όμως αντιληπτό πως η μεγάλη πλειοψηφία τους είναι επιχειρήσεις από τις ΗΠΑ. Συγκεκριμένα, στην πολιτεία της Καλιφόρνιας ανήκει η πιο διάσημη περιοχή για τεχνολογικές επιχειρήσεις στον κόσμο, η Silicon Valley, που αποτελεί σπίτι για πολλές τεράστιες τεχνολογικές επιχειρήσεις παγκοσμίως, συμπεριλαμβανομένων 39 επιχειρήσεων του Fortune 1000 που εδρεύουν εκεί και συνολικά 109 σε ολόκληρη την πολιτεία. Η Silicon Valley έγινε παγκόσμια συνώνυμο για την ηγετική της θέση στην έρευνα υψηλής τεχνολογίας, με τις εταιρίες της να αγγίζουν το ένα τρίτο του επενδυτικού κεφαλαίου σε ολόκληρη τη χώρα, προσελκύοντας πολλούς εργαζόμενους εξειδικευμένους στην υψηλή τεχνολογία.



Σχήμα 4.5 – Παγκόσμιος χάρτης με πυκνότητα των startups ανά χώρα

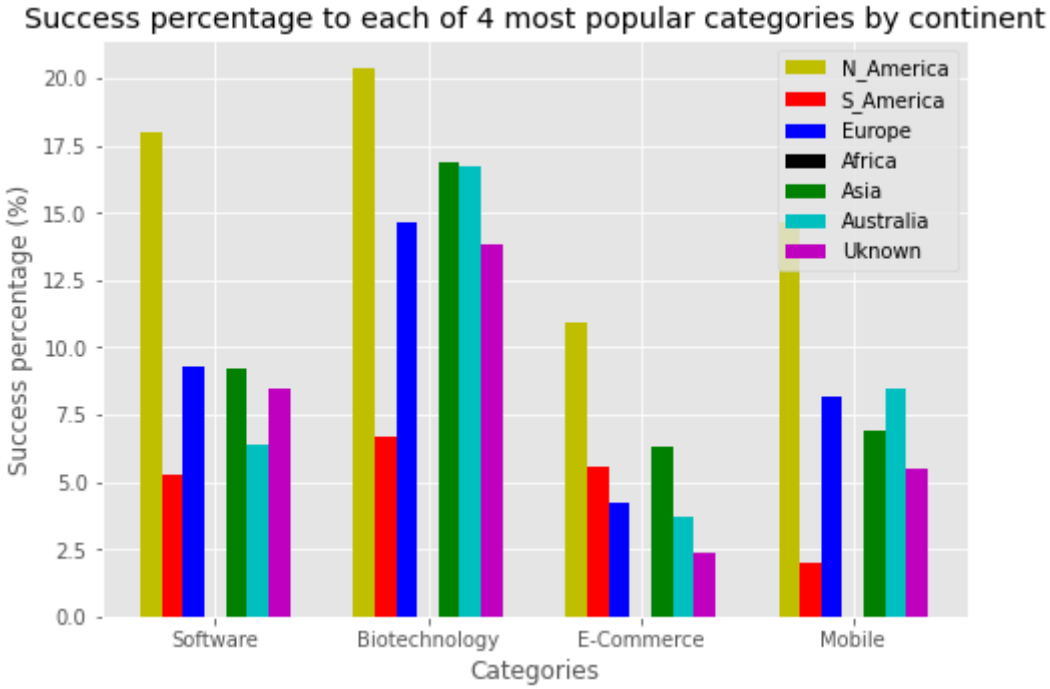


Σχήμα 4.6 – Χάρτης ΗΠΑ με πυκνότητα των startups ανά πολιτεία



Σχήμα 4.7 – Ποσοστό επιτυχημένων επιχειρήσεων ανά ήπειρο

Η Β. Αμερική έχει με διαφορά το μεγαλύτερο ποσοστό επιτυχίας στις επιχειρήσεις με 15%, ενώ με 7%, λιγότερο από το μισό, ακολουθούν Ευρώπη, Ασία και Ωκεανία. Γίνεται, λοιπόν, σαφές πως οι ΗΠΑ αποτελούν το πιο ευνοϊκό περιβάλλον για μια επιχείρηση.

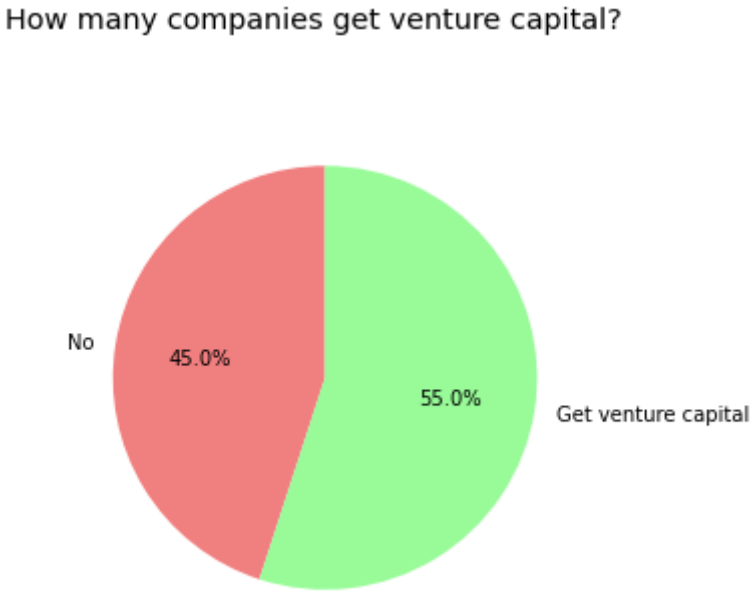


Σχήμα 4.8 – Ποσοστό επιτυχημένων επιχειρήσεων ανά ήπειρο για τις 4 δημοφιλέστερες κατηγορίες

4.2.2 Χρηματοδότηση

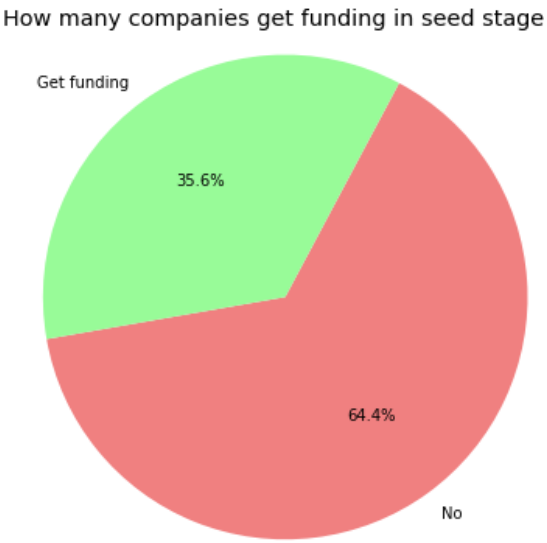
Η πλειοψηφία των επιχειρήσεων ασχολούνται με την τεχνολογία και κρίνεται καθοριστικής σημασίας η κατανόηση της χρηματοδότησής τους μέσω επενδυτικού κεφαλαίου και του τρόπου λήψης επενδυτικών πόρων, αντί να δανείζονται από τράπεζες, που επιτρέπει την ταχεία ανάπτυξή τους με στόχο την άμεση επιτυχία. Το

επενδυτικό κεφάλαιο (venture capital) είναι ένας τύπος χρηματοδότησης με τον οποίο οι επενδυτές παρέχουν στην (νεοφυή) επιχείρηση χρήματα με αντάλλαγμα ένα ποσοστό ιδιοκτησίας της επιχείρησης. Αυτός ο μηχανισμός έχει ως προϋπόθεση την προοπτική εκθετικής ανάπτυξης των νεοφυών επιχειρήσεων, αλλά συνδέεται και με υψηλό ρίσκο. Εξαιτίας της φύσης της τεχνολογίας, όπου συναντάται ταχύτερη δυνατότητα επεκτασιμότητας από τις παραδοσιακές μη τεχνολογικές επιχειρήσεις, το επενδυτικό κεφάλαιο συνδέεται στενά με τις τεχνολογικές επιχειρήσεις. Ο μέσος όρος χρηματοδότησης κεφαλαίου, χωρίς τα μηδενικά, είναι 11,755,736 δολάρια.



Σχήμα 4.9 – Ποσοστό επιχειρήσεων που έλαβαν χρηματοδότηση κεφαλαίου (venture capital)

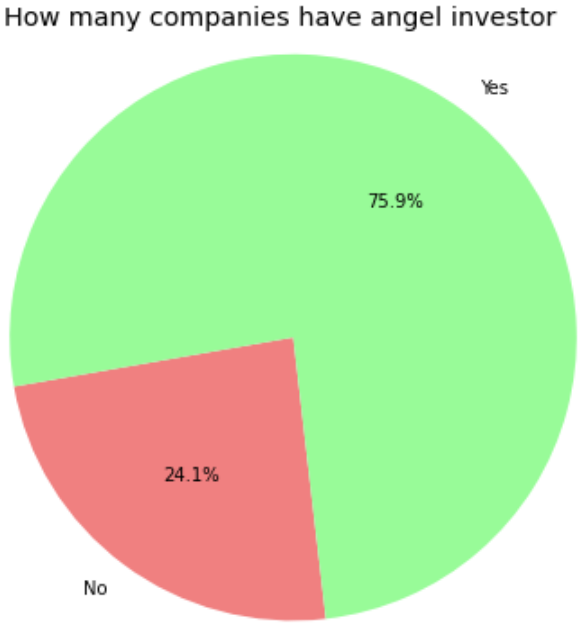
Ένας άλλος τύπος χρηματοδότησης είναι αυτός του σπόρου (seed funding). Τυπικά αποτελεί την πρώτη επίσημη λήψη κεφαλαίου για την επιχείρηση. Ωστόσο, για μερικές



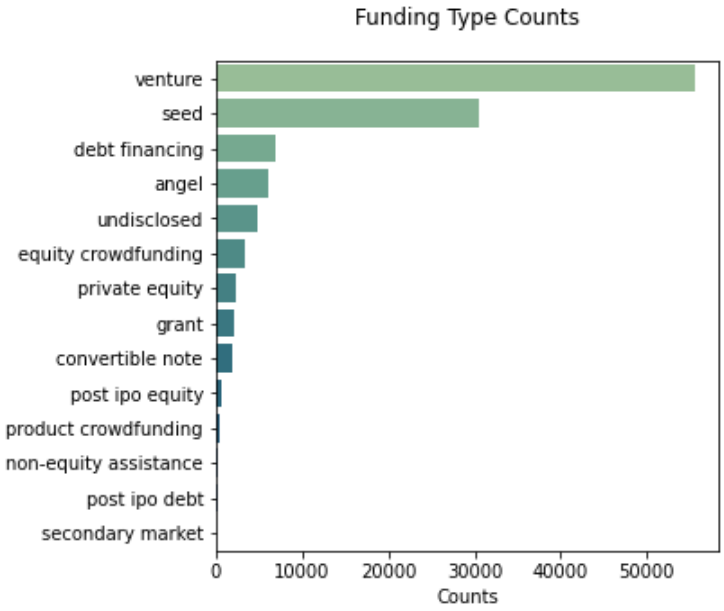
Σχήμα 4.10 – Ποσοστό επιχειρήσεων που λαμβάνουν χρηματοδότηση σπόρου (seed funding)

επιχειρήσεις αποτελεί και τη μοναδική φορά που λαμβάνουν χρηματοδότηση. Ο μέσος όρος χρηματοδότησης σπόρου, χωρίς τα μηδενικά, είναι 722.726 δολάρια.

Επιπρόσθετα, ο επιχειρηματικός άγγελος, γνωστός και ως ιδιωτικός επενδυτής, είναι κάποιος με σημαντικό κεφάλαιο που παρέχει χρηματική υποστήριξη σε μικρές επιχειρήσεις και επιχειρηματίες, συνήθως με αντάλλαγμα μερίδιο ιδιοκτησίας από την επιχείρηση. Συχνά, ως επιχειρηματικός άγγελος λειτουργεί κάποιο μέλος της οικογένειας ή κάποιος φίλος του επιχειρηματία. Οι πόροι που ο επιχειρηματικός άγγελος παρέχει είναι μια εφάπαξ επένδυση με στόχο να ενισχύσει την επιχείρηση να ξεκινήσει και να λειτουργήσει ως βοήθεια στα πρώτα δύσκολα στάδια της. Ο μέσος όρος χρηματοδότησης αγγέλου, χωρίς τα μηδενικά, είναι 964,849 δολάρια.



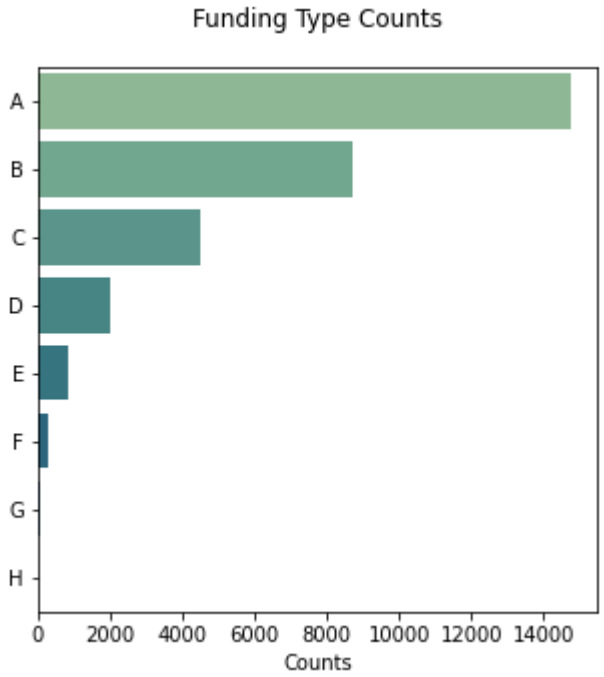
Σχήμα 4.11 – Ποσοστό επιχειρήσεων που λαμβάνουν χρηματοδότηση αγγέλου (Angel Funding)



Σχήμα 4.12 – Πλήθος τύπων χρηματοδότησης

Στον παραπάνω πίνακα παρουσιάζεται το πλήθος επενδύσεων κάθε τύπου χρηματοδότησης.

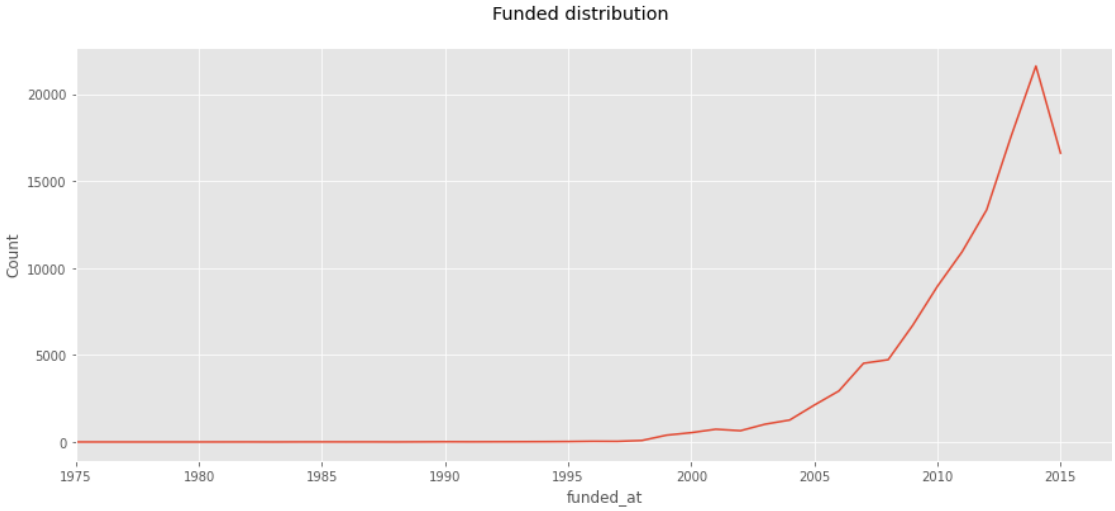
Ενώ στα επόμενα σχήματα φαίνονται οι γύροι χρηματοδότησης και ο μέσος όρος χρημάτων που δόθηκαν στον καθένα (χωρίς τα μηδενικά).



8,173,613
14,518,096
22,178,010
28,866,574
38,190,589
55,016,743
43,900,000
-

Σχήμα 4.13 – Γύροι χρηματοδότησης με μέσος όρος χρηματοδότησης ($)

Το έτος 2014 έγιναν οι περισσότερες χρηματοδοτήσεις στο σετ δεδομένων όπως παρουσιάζεται στο παρακάτω σχήμα.



Σχήμα 4.14 – Χρηματοδότηση ανά έτος

Επίσης, αφού όπως αναφέρεται ο κύριος λόγος αποτυχίας των νεοφυών επιχειρήσεων είναι η έλλειψη κεφαλαίου στα πρώτα τους στάδια (Griffith, 2014), κρίνεται καθοριστικής σημασίας ο χρόνος που μεσολαβεί μεταξύ ίδρυσης και πρώτης χρηματοδότησης. Σύμφωνα με το σετ δεδομένων το σύνολο των επιχειρήσεων,

επιτυχημένων και μη, έλαβαν την πρώτη τους χρηματοδότηση 45.4 μήνες (3.8 χρόνια) μετά την ίδρυσή τους, κατά μέσο όρο. Ο αντίστοιχος μέσος όρος για τις αποτυχημένες επιχειρήσεις μειώνεται στους 41.8 μήνες (3.5 χρόνια), ενώ για τις επιτυχημένες εκτοξεύεται στους 72.4 μήνες (6 χρόνια).

Κατηγορία	**Γενικός Μ.Ο.**	**Μ.Ο. Επιτυχημένων**	**Μ.Ο. Αποτυχημένων**
Λογισμικό	47.8	68.7	44.0
Βιοτεχνολογία	66.2	94.2	59.3
Ηλεκτρονική διαφήμιση	29.4	60.6	26.7
Κινητά τηλέφωνα	24.2	38.9	22.2
Τεχνολογία καθαρισμού	63.4	100.4	58.5
Ταξινόμηση διαδικτύου	23.9	30.7	22.8
Υλικό και Λογισμικό	69.2	142.4	57.7
Ιατρική περίθαλψη	56.0	87.5	49.2
Παιχνίδια	28.2	36.0	27.0
Λογισμικό για επιχειρήσεις	38.4	52.9	35.1
Άλλος τομέας	45.4	71.7	42.2
Συνολικά	45.4	72.4	41.8
Πριν το 2005	155.7	128.4	167.0
Μεταξύ 2005 και 2010	31.4	20.4	33.0
Μετά το 2010	10.7	7.6	10.8

Πίνακας 4.8 – Μήνες μεταξύ ίδρυσης και 1ης χρηματοδότησης

Γίνεται φανερό πως η ερμηνεία του χρόνου μεταξύ ίδρυσης της εταιρίας και πρώτης χρηματοδότησης είναι παραπλανητική χωρίς την εξέταση του χρονικού έτους ίδρυσης της. Όπως φαίνεται από τις τρεις τελευταίες γραμμές του παραπάνω πίνακα, πάντα οι επιτυχημένες επιχειρήσεις είναι αυτές που λαμβάνουν άμεση χρηματοδότηση, ενώ όσες καθυστερούν να εξασφαλίσουν κεφάλαιο αποτυγχάνουν. Όμως, οι τακτικές των επενδυτών έχουν αλλάξει δραματικά τα τελευταία χρόνια. Πριν το 2005, μια επιχείρηση λάμβανε χρηματοδότηση 155.7 μήνες, σχεδόν 13 χρόνια, μετά την ίδρυσή της, γεγονός που οφείλεται στην τακτική των επενδυτών να λαμβάνουν χαμηλά ρίσκα. Ο αριθμός αυτός μειώθηκε δραματικά την πενταετία μετά το 2005 στους 31.4 μήνες, λίγο παραπάνω από 2.5 χρόνια, ενώ μετά το 2010 ελαττώθηκε στους 10.7 μήνες, γεγονός που καταδεικνύει την τάση των επενδυτών να ρισκάρουν τα τελευταία χρόνια.

	Επενδυτές ανά επιχείρηση	Πλήθος εταιριών που χρηματοδοτούν οι επενδυτές	Ποσοστό επιτυχίας επενδυτών
Πριν το 2005	4.37	309	53.1%
Μεταξύ 2005 και 2010	4.86	437	42.3%
Μετά το 2010	4.85	438	30.1%
Σύνολο	4.4	348	45.7%

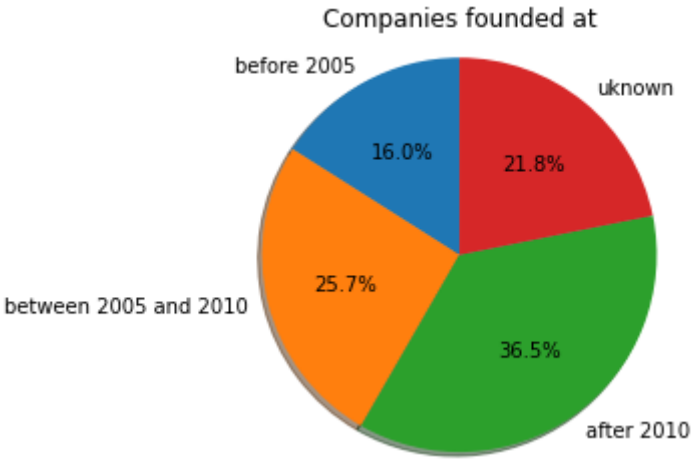
Πίνακας 4.9 – Συμπεριφορά επενδυτών επιτυχημένων επιχειρήσεων ανά χρονική περίοδο

	Επενδυτές ανά επιχείρηση	Πλήθος εταιριών που χρηματοδοτούν οι επενδυτές	Ποσοστό επιτυχίας επενδυτών
Πριν το 2005	3.54	224	56.3%
Μεταξύ 2005 και 2010	3.77	268	44.8%
Μετά το 2010	3.80	272	31.3%
Σύνολο	3.58	240	41.3%

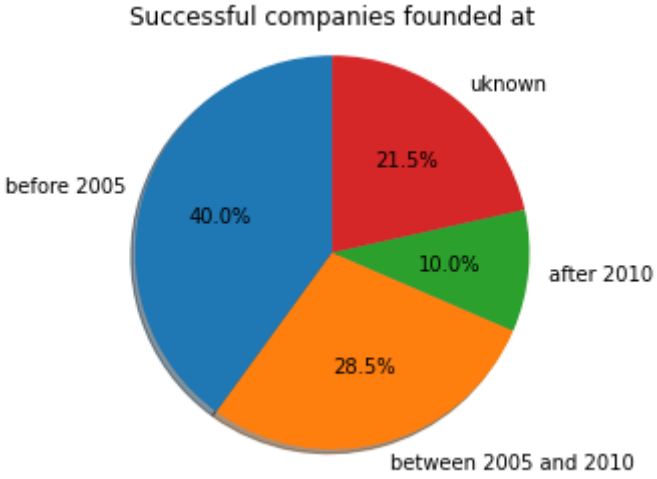
Πίνακας4.10 – Συμπεριφορά επενδυτών επιχειρήσεων ανά χρονική περίοδο

Όπως φαίνεται από τους προηγούμενους δύο πίνακες, οι επιτυχημένοι επενδυτές χρηματοδοτούν περισσότερες επιχειρήσεις. Η τακτική που ακολουθούν τα τελευταία χρόνια μειώνει το ποσοστό των επιτυχιών τους, αφού οι επενδυτές ρισκάρουν επενδύοντας σε περισσότερες, αλλά και ηλικιακά νεότερες επιχειρήσεις, όπως φαίνεται στους τρεις προηγούμενους πίνακες.

Η πλειοψηφία των επιχειρήσεων της δοσμένης βάσης είναι εταιρίες που ιδρύθηκαν μετά το 2010 και οι πληροφορίες για αυτές σταματούν το 2015. Οπότε, οι περισσότερες δεν είχαν την ευκαιρία σε τόσο σύντομο χρονικό διάστημα να εξαγοραστούν ή να εισαχθούν στο χρηματιστήριο. Ταυτόχρονα, οι περισσότερες επιτυχημένες επιχειρήσεις έχουν ιδρυθεί πριν το 2005 και είχαν την ευκαιρία να εξελιχθούν.



Σχήμα 4.15 – Έτος ίδρυσης εταιριών



Σχήμα 4.16 – Έτος ίδρυσης επιτυχημένων εταιριών

4.2.3 Αγοραστές

Η σύγχρονη τάση για ανάληψη μεγαλύτερου ρίσκου ακολουθήθηκε και από τους αγοραστές, όπου μετά το 2010 αύξησαν κατά 61.5% τις εξαγορές τους. Γίνεται σαφές πως τα τελευταία χρόνια οι εταιρίες εξαπλώνονται ραγδαία με συνέπεια να εξαγοράζονται και να συγχωνεύονται ταχύτατα.

	Πλήθος εταιριών που εξαγόρασαν οι αγοραστές
Πριν το 2005	17.9
Μεταξύ 2005 και 2010	20.8
Μετά το 2010	29.1
Σύνολο	18.9

Πίνακας 4.11 – Συμπεριφορά αγοραστών επιχειρήσεων ανά χρονική περίοδο

4.2.4 Κατηγορίες δραστηριοποίησης

Είναι γνωστό πως οι περισσότερες νεοφυείς επιχειρήσεις ασχολούνται καθαρά με την ανάπτυξη λογισμικού, γεγονός που αποτυπώνεται και στο δοσμένο σετ δεδομένων, αφού το 14% των επιχειρήσεων (7520 επιχειρήσεις) ανήκει στη συγκεκριμένη κατηγορία.

Ωστόσο, το ποσοστό επιτυχίας στην κάθε μια από τις 10 δημοφιλέστερες κατηγορίες του σετ δεδομένων ποικίλει από 8% έως 19%. Οι εταιρίες που ασχολούνται με την βιοτεχνολογία κατέχουν την πρώτη θέση στην ποσοστιαία επιτυχία, που ακολουθείται από τις εταιρίες που ασχολούνται με την ανάπτυξη λογισμικού για επιχειρήσεις (18%) και την ιατρική περίθαλψη (17%)

Κατηγορία	**Πλήθος επιχειρήσεων**	**Ποσοστό επί του πλήθους (%)**	**Ποσοστό επιτυχίας (%)**
Λογισμικό	7520	14	15
Κινητά τηλέφωνα	4792	9	11
Βιοτεχνολογία	4307	8	19
Ηλεκτρονική διαφήμιση	3402	6	8
Ταξινόμηση διαδικτύου	2493	5	14
Λογισμικό για επιχειρήσεις	2387	4	18
Ιατρική περίθαλψη	2120	4	17
Παιχνίδια	1792	3	13
Τεχνολογία καθαρισμού	1378	3	11
Υλικό και Λογισμικό	1371	3	13
Άλλος τομέας	39341	73	10

Πίνακας 4.12 – Κατηγορίες δραστηριοποίησης

4.3 Μετρικές Αξιολόγησης

Χρησιμοποιείται ο *πίνακας σύγχυσης* (confusion matrix) για την ακριβή κατανόηση των προβλέψεων του εκάστοτε μοντέλου. Αποτελεί τετραγωνικό πίνακα ίσο με το πλήθος των διαφορετικών κατηγοριών. Κάθε γραμμή αντιπροσωπεύει το σύνολο των στιγμιότυπων που ανήκουν σε μία κατηγορία, ενώ κάθε στήλη το σύνολο των στιγμιότυπων που ταξινομήθηκαν στην κατηγορία. Τα στιγμιότυπα της διαγώνιου είναι αυτά που έχουν ταξινομηθεί ορθά.

	Πρόβλεψη Αποτυχίας (0)	Πρόβλεψη Επιτυχίας (1)
Αποτυχημένες (0)	TN	FP
Επιτυχημένες (1)	FN	TP

Για την αξιολόγηση ενός μοντέλου μηχανικής μάθησης χρησιμοποιούνται οι μετρικές:

- *Ακρίβεια* (precision): είναι ο λόγος με αριθμητή τις σωστές προβλέψεις της θετικής κατηγορίας και παρονομαστή το άθροισμα των στιγμιότυπων που ταξινομήθηκαν στη θετική κλάση (σωστά ή λανθασμένα):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- *Ανάκληση* (recall): είναι ο λόγος με αριθμητή τις σωστές προβλέψεις της θετικής κατηγορίας και παρονομαστή το άθροισμα των λάθος ταξινομήσεων για την αρνητική κατηγορία:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- *F1-score*: λαμβάνει υπόψη τόσο την ακρίβεια όσο και την ανάκληση:

$$\text{f1\_score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Στην παρούσα εργασία, ως κύρια μετρική αξιολόγησης χρησιμοποιείται η macro F1-score δίνοντας στις δύο μετρικές, ακρίβεια και ανάκληση, την ίδια βαρύτητα.

4.4 Παρουσίαση συνόλων δεδομένων και Αποτελέσματα

Οι αλγόριθμοι που παρουσιάστηκαν στην ενότητα 3.3 εφαρμόστηκαν σε διάφορα σύνολα δεδομένων. Τα σύνολα δεδομένων δημιουργήθηκαν επιλέγοντας χαρακτηριστικά που θεωρήθηκαν χρήσιμα για την πρόβλεψη μελλοντικής επιτυχίας των επιχειρήσεων. Όπως είναι αναμενόμενο αρχικά επιλέχθηκαν λίγα, κύρια χαρακτηριστικά και στην πορεία προστέθηκαν όσα θεωρήθηκε πως θα προσθέσουν πληροφορία, αφού πρώτα δοκιμάστηκαν και έγινε κατανοητό ότι προσφέρουν μεγαλύτερη ακρίβεια στις προβλέψεις. Πριν την εκπαίδευση των αλγορίθμων τα σετ αυτά επεξεργάστηκαν για τη βελτιστοποίηση της μάθησης (i) κανονικοποιήθηκαν (Standard Scaler) και (ii) ισοζυγίστηκαν οι δύο κλάσεις (SMOTE). Τα ονόματά τους θα είναι ίσα με το πλήθος τον χαρακτηριστικών που αξιοποιούν και τα χαρακτηριστικά τους περιγράφονται παρακάτω μαζί με τα αποτελέσματα τους σε κάθε αλγόριθμο.

4.4.1 10 Χαρακτηριστικά

Αρχικά, τα χαρακτηριστικά που αξιοποιήθηκαν ήταν η συνολική χρηματοδότηση, το πλήθος των γύρων χρηματοδότησης και το ποσό που δόθηκε σε κάθε ένα από τους γύρους Α έως Η.

	Precision (%)	**Recall (%)**	**F1-score (%)**	**Training Runtime (sec)**	**Prediction Runtime (sec)**
XGBoost	61.0	66.0	62.5	2.7	<1
KNN	55.7	60.0	55.8	6.0	<2
SVM	56.1	61.9	55.7	389.3	41.6
Random Forest	55.7	60.4	55.5	<1	<1
Naïve Bayes	57.3	54.4	55.0	<1	<1
Logistic Regression	58.0	54.0	54.8	<1	<1

Πίνακας 4.13 – Αποτελέσματα αλγορίθμων για 10 χαρακτηριστικά

	Predicted 0	**Predicted 1**
Companies 0	8072	1390
Companies 1	668	587

Πίνακας 4.14 - XGBoost confusion matrix για 10 χαρακτηριστικά

4.4.2 998 Χαρακτηριστικά

Στη συνέχεια, προστέθηκαν στα προηγούμενα 10, ο κωδικός της χώρας ίδρυσης και η κατηγορία (κατηγορίες) δραστηριοποίησης κάθε εταιρίας, που συνολικά αθροίζονταν στα 998.

	Precision (%)	**Recall (%)**	**F1-score (%)**	**Training Runtime (sec)**	**Prediction Runtime (sec)**
XGBoost	61.4	64.9	62.7	112	<1
Logistic Regression	61.0	63.0	61.7	10	<1
Random Forest	59.3	58.7	59.0	7	<1
SVM	58.6	58.5	58.5	2252	413

KNN	56.3	60.6	56.8	<1	56
Naïve Bayes	53.9	53.1	53.4	<1	<1

Πίνακας 4.15 – Αποτελέσματα αλγορίθμων για 998 χαρακτηριστικά

	Predicted 0	**Predicted 1**
Companies 0	8332	1162
Companies 1	708	515

Πίνακας 4.16 - XGBoost confusion matrix για 998 χαρακτηριστικά

4.4.3 37 Χαρακτηριστικά

Καθώς τα 998 χαρακτηριστικά ήταν πάρα πολλά και καθυστερούσε σημαντικά η εκπαίδευση των μοντέλων αποφασίστηκε να διατηρηθούν οι 9 δημοφιλέστερες χώρες και να σημειώνεται σε ποια ήπειρο ανήκει η κάθε εταιρία. Επίσης, παρόμοια στρατηγική ακολουθήθηκε και για τις κατηγορίες δραστηριοποίησης όπου διατηρήθηκαν οι 10 δημοφιλέστερες και οι υπόλοιπες θεωρήθηκαν κατηγορία Άλλο. Όπως θα φανεί στα αποτελέσματα, τα επιπλέον 961 χαρακτηριστικά δεν προσέφεραν σημαντική πληροφορία.

	Precision (%)	**Recall (%)**	**F1-score (%)**	**Training Runtime (sec)**	**Prediction Runtime (sec)**
XGBoost	62.4	65.9	63.7	6	<1
KNN	59.2	62.4	60.2	<1	14
Random Forest	58.3	59.8	58.9	<1	<1
Logistic Regression	60.4	56.8	57.9	<1	<1
SVM	60.8	54.4	55.2	234	37
Naïve Bayes	56.0	54.3	55.1	<1	<1

Πίνακας 4.17 – Αποτελέσματα αλγορίθμων για 37 χαρακτηριστικά

	Predicted 0	**Predicted 1**
Companies 0	8262	1162
Companies 1	723	570

Πίνακας 4.18 - XGBoost confusion matrix για 37 χαρακτηριστικά

4.4.4 43 Χαρακτηριστικά

Σε αυτό το σημείο λήφθηκε υπόψη η χρηματοδότηση των εταιριών μέσω των διάφορων τύπων χρηματοδότησης. Στα προηγούμενα 37 χαρακτηριστικά προστέθηκαν οι 6 δημοφιλέστεροι τύποι χρηματοδότησης.

	Precision (%)	**Recall (%)**	**F1-score (%)**	**Training Runtime (sec)**	**Prediction Runtime (sec)**
XGBoost	63.0	65.5	64.0	6	<1
SVM	63.5	62.4	62.9	283	36
Logistic Regression	61.1	62.0	61.5	<1	<1
KNN	59.1	62.8	60.2	<1	14
Random Forest	58.3	58.5	58.4	<1	<1
Logistic Regression	60.4	56.8	57.9	<1	<1
Naïve Bayes	56.2	57.4	56.8	<1	<1

Πίνακας 4.19 – Αποτελέσματα αλγορίθμων για 43 χαρακτηριστικά

	Predicted 0	**Predicted 1**
Companies 0	8420	1029
Companies 1	737	531

Πίνακας 4.20 - XGBoost confusion matrix για 43 χαρακτηριστικά

4.4.5 51 Χαρακτηριστικά

Καθώς τα αποτελέσματα ήταν ενθαρρυντικά προστέθηκαν και οι υπόλοιποι τύποι χρηματοδότησης.

	Precision (%)	**Recall (%)**	**F1-score (%)**	**Training Runtime (sec)**	**Prediction Runtime (sec)**
Logistic Regression	70.7	70.7	70.7	<2	<1
SVM	70.4	70.9	70.7	271	36
XGBoost	70.4	69.8	70.1	9	<1

Random Forest	70.7	66.4	68.2	<1	<1
KNN	69.8	64.9	65.7	<1	15
Naïve Bayes	63.7	62.7	63.2	<1	<1

Πίνακας 4.21 – Αποτελέσματα αλγορίθμων για 51 χαρακτηριστικά

	Predicted 0	Predicted 1
Companies 0	8887	614
Companies 1	655	561

Πίνακας 4.22 - XGBoost confusion matrix για 51 χαρακτηριστικά

4.4.6 64 Χαρακτηριστικά

Αξιοποιήθηκαν πληροφορίες για τους επενδυτές και τους αγοραστές. Συγκεκριμένα, σημειώθηκε πόσοι επενδυτές έχουν χρηματοδοτήσει την κάθε εταιρία, το ποσοστό επιτυχίας και η εμπειρία τους. Τέλος, σημειώθηκε σε ποιο χρονικό διάστημα (πριν το 2005, μεταξύ 2005 και 2010, μετά το 2010) ιδρύθηκαν και μετά από πόσους μήνες λειτουργίας έλαβαν την πρώτη τους χρηματοδότηση.

	Precision (%)	**Recall (%)**	**F1-score (%)**	**Training Runtime (sec)**	**Prediction Runtime (sec)**
XGBoost	73.2	72.3	72.7	9.6	<1
SVM	71.7	73.4	72.5	282	28
Logistic Regression	70.1	73.5	71.6	<2	<1
Random Forest	73.9	68.1	70.3	<2	<1
Naïve Bayes	70.7	66.1	68.0	<1	<1
KNN	65.5	71.3	67.5	<1	16

Πίνακας 4.23 – Αποτελέσματα αλγορίθμων για 64 χαρακτηριστικά

	Predicted 0	Predicted 1
Companies 0	8824	585
Companies 1	643	665

Πίνακας 4.24 - XGBoost confusion matrix για 64 χαρακτηριστικά

Κεφάλαιο 5: Συμπεράσματα και Μελλοντικές Κατευθύνσεις

5.1 Συμπεράσματα

Τα αποτελέσματα θα σχολιαστούν, αρχικά αφαιρετικά και ύστερα ανά αλγόριθμο ταξινομημένο με βάση τη βαθμολογία του στο τελευταίο σετ δεδομένων όπου όλοι παρουσιάζουν τα βέλτιστα αποτελέσματα.

5.1.1 Γενικά

Ως 1ο σετ δεδομένων σημειώνεται '10*' και αποτελείται από τα δέκα χαρακτηριστικά του 1ου σετ με μόνη επεξεργασία την κανονικοποίησή τους. Ο λόγος ύπαρξης του είναι να αποτελέσει βάση σύγκρισης και να αξιολογηθεί πόση αξία προστέθηκε μετά την προσθήκη και επεξεργασία των χαρακτηριστικών.

Στα πρώτα 4 σετ δεδομένων δεν παρατηρούνται σημαντικές διαφοροποιήσεις στην απόδοση των αλγορίθμων, ωστόσο στο 5ο σετ (51 χαρακτηριστικά) η μετρική του f1_score αυξάνεται σημαντικά και στο 6ο (64 χαρακτηριστικά) βελτιστοποιείται. Γίνεται σαφές πως οι πληροφορίες σχετικά με τον τύπο χρηματοδότησης που έλαβε η κάθε εταιρία (5ο σετ) βελτιώνουν αισθητά τα αποτελέσματα. Τέλος, η χρονική στιγμή ίδρυσης της εταιρίας, η εμπειρία και το ποσοστό επιτυχίας επενδυτών και αγοραστών βελτιστοποιούν τα αποτελέσματα.

Η χρονική στιγμή ίδρυσης μια εταιρίας είναι καθοριστική για την τακτική που πρέπει να ακολουθήσει, καθώς τα τελευταία χρόνια η αγορά έχει αλλάξει δραματικά ρισκάροντας περισσότερο με γνώμονα τα εκθετικά υψηλότερα κέρδη, όπως παρουσιάστηκε στην ενότητα 4.2.2. Ταυτόχρονα, όπως αναμενόταν, η γνώση των επιτυχημένων επενδυτών με μεγάλη εμπειρία προσφέρει σημαντική και αξιοποιήσιμη πληροφορία στους ταξινομητές.

Οι αλγόριθμοι, εκτός του SVM, εκπαιδεύονται και προβλέπουν ταχύτατα σχεδόν σε κάθε περίπτωση σε λιγότερο από 2 δευτερόλεπτα, ενώ όλοι χρειάζονται περισσότερο χρόνο εκπαίδευσης στο 2ο σετ δεδομένων προκειμένου να επεξεργαστούν και τα 998 χαρακτηριστικά.

Το σύνολο με 998 χαρακτηριστικά έδωσε υψηλότερο σκορ 3% σε σύγκριση με το σύνολο των 37, όπου τα 998 ομαδοποιούνται, μόνο στους ταξινομητές LR και SVM, γεγονός που οφείλεται στην αποδοτική συμπύκνωση των πολυπληθών χαρακτηριστικών.

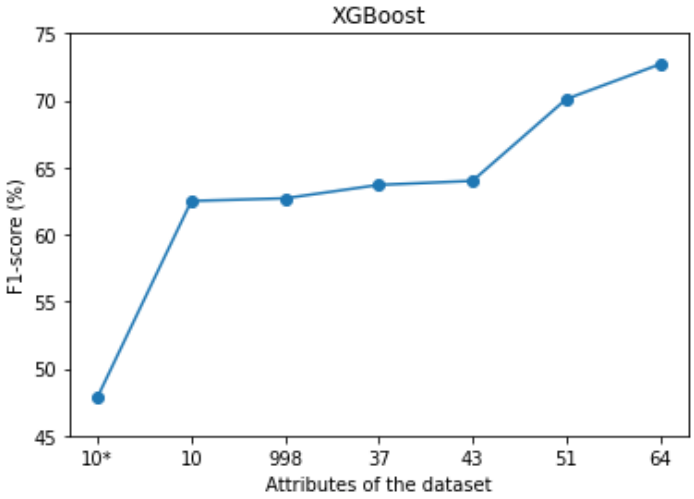
Οι αλγόριθμοι που τελικά πέτυχαν το υψηλότερο σκορ (πάνω από 71%) (XGBoost, SVM, LR) στο σετ '10*' είχαν τη μικρότερη βαθμολογία (περίπου 47.4%) που αποδεικνύει πως η επεξεργασία που πραγματοποιήθηκε στην παρούσα εργασία αύξησε καθοριστικά το σκορ στους αλγορίθμους με υψηλές δυνατότητες. Αντίθετα, οι αλγόριθμοι με χαμηλά σκορ ξεκίνησαν με f1-score άνω του 53%, γεγονός που καταδεικνύει τις περιορισμένες δυνατότητες των αλγορίθμων RF, GNB και kNN.

5.1.2 XGBoost

Η παρούσα εργασία είχε στόχο να βελτιστοποιήσει το σκορ του συγκεκριμένου αλγορίθμου, καθώς αποτελεί υπερσύγχρονο ταξινομητή που παρουσιάζει πολύ

ενθαρρυντικά αποτελέσματα σε διαγωνισμούς του Kaggle, αλλά και στην εκπαίδευση με δεδομένα από το CrunchBase. Για το λόγο αυτό επιλέχθηκε ο ισοζυγισμός των κλάσεων με τη μέθοδο SMOTE.

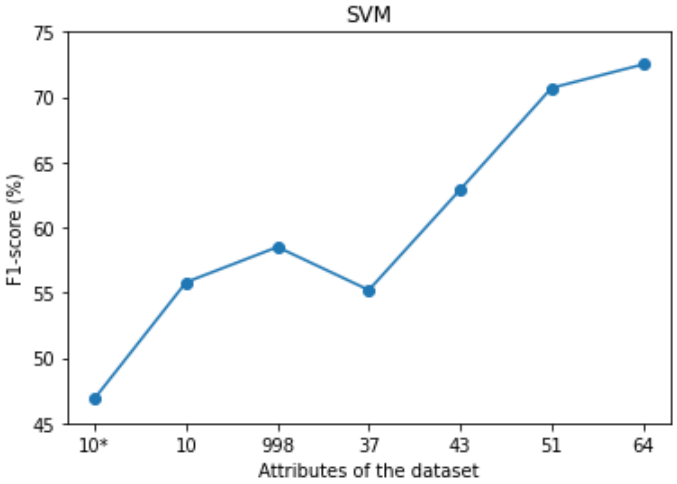
Το 72.7% αποτελεί το βέλτιστο σκορ για την παρούσα εργασία και είναι δραματικά ανώτερο από το 47.9% όπου ξεκίνησε ο συγκεκριμένος αλγόριθμος (αύξηση 52%).



Σχήμα 5.1 – XGBoost F1-score

5.1.3 SVM

Ο SVM σημείωσε το 2ο καλύτερο σκορ, ωστόσο στα πρώτα τρία σύνολα δεδομένων τα αποτελέσματά του ήταν αποθαρρυντικά. Τα χαρακτηριστικά σχετικά με τον τύπο χρηματοδότησης τον ανέδειξαν στην παρούσα εργασία.

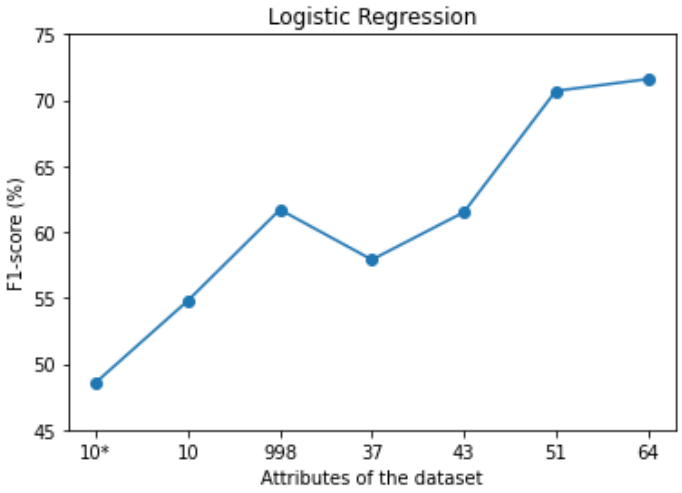


Σχήμα 5.2 – SVM F1-score

5.1.4 Logistic Regression

Ο ταξινομητής Logistic Regression δεν ήταν στα αρχικά πλάνα της εργασίας, ωστόσο προστέθηκε για την πληρότητα της, αφού σύγχρονες έρευνες στηρίζονται σε αυτόν (Calafiore, 2019), (Żbikowski, 2021).

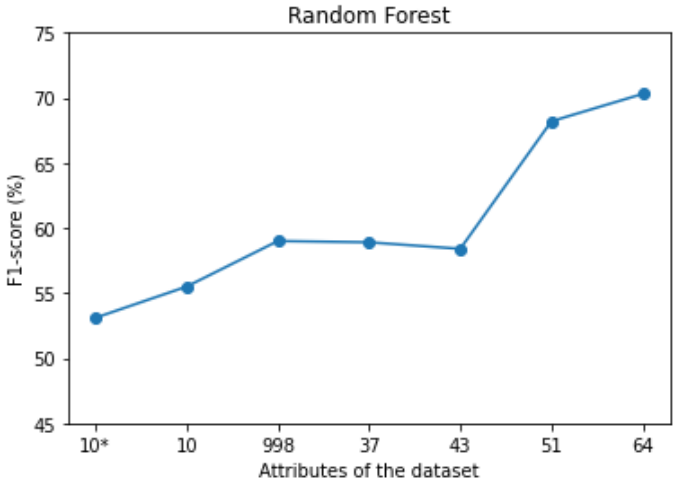
Αποτελεί το μοναδικό αλγόριθμο που σημείωσε τόσο μεγάλη βελτιώση στο σύνολο των 998 χαρακτηριστικών. Επίσης, η εμπειρία των επενδυτών και η χρονική στιγμή ίδρυσης της κάθε εταιρίας βελτίωσαν ελάχιστα την απόδοσή του.



Σχήμα 5.3 – Logistic Regression F1-score

5.1.5 Random Forest

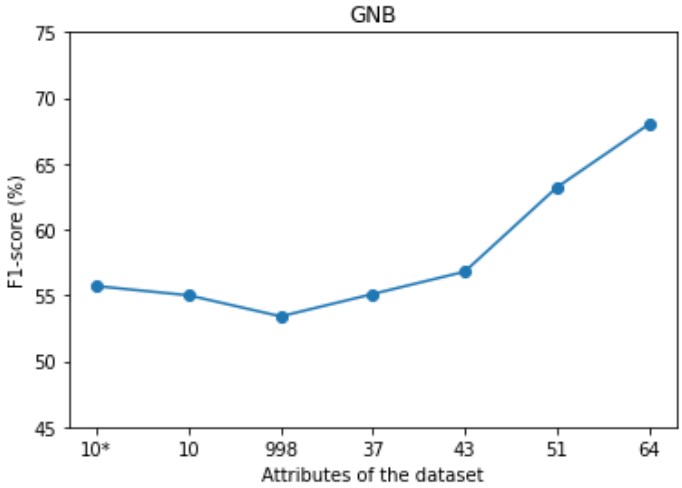
Ο ταξινομητής έγινε ανταγωνιστικός, κοντά στο 70%, μετά την εισαγωγή όλων των τύπων χρηματοδότησης. Με την προηγούμενη πληροφορία, μέχρι και τα 41 χαρακτηριστικά, η απόδοση του ήταν πολύ χαμηλή, λιγότερο από 60%.



Σχήμα 5.4 – Random Forest F1-score

5.1.6 Gaussian Naïve Bayes

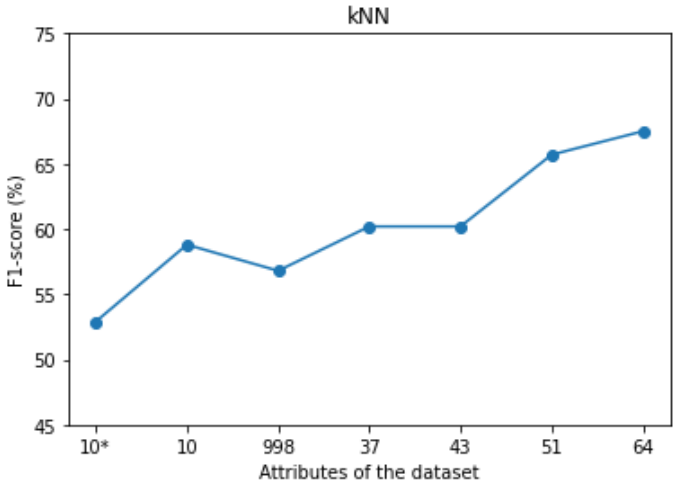
Οι περιορισμοί του συγκεκριμένου αλγορίθμου ήταν γνωστοί, οπότε χρησιμοποιήθηκε για την πληρότητα της εργασίας, ενώ ταυτόχρονα αποτελεί και βάση σύγκρισης με τους υπόλοιπους. Ήταν ο μόνος που τα πήγε σχετικά καλά με το σύνολο ‘10*’ και η απόδοση του μέχρι και το σύνολο των 43 χαρακτηριστικών δεν αυξήθηκε ιδιαίτερα.



Σχήμα 5.5 – GNB F1-score

5.1.7 k-Nearest Neighbors

Ενώ στα πρώτα σύνολα δεδομένων τα αποτελέσματά του ήταν αποδεκτά, στα τελευταία δύο σύνολα που όλοι οι ταξινομητές βελτιώθηκαν σημαντικά, αυτός κινήθηκε σε χαμηλά επίπεδα.



Σχήμα 5.6 – kNN F1-score

5.1.8 Συζήτηση

Το σύνολο που αξιοποιήθηκε για την αξιολόγηση των μοντέλων αποτελείται από 9,409 αποτυχημένες και 1,308 επιτυχημένες εταιρίες. Προφανώς, οι δύο κατηγορίες είναι μη ισοζυγισμένες, αλλά αυτό συμβαίνει και στην πράξη, καθώς περίπου το 10% των επιχειρήσεων επιτυγχάνει. Οπότε, καθοριστικής σημασίας είναι όχι μόνο πόσες επιτυχημένες επιχειρήσεις εντοπίζονται, αλλά και πόσες αποτυχημένες αποφεύγονται.

Στις προηγούμενες παραγράφους του κεφαλαίου έγινε κατανοητό πως ο ταξινομητής XGBoost παρουσιάζει τα βέλτιστα αποτελέσματα. Σύμφωνα με τον Πίνακα το ποσοστό των ορθών θετικών δειγμάτων (True Positive Rate, TPR) είναι 50.8%, που σημαίνει ότι εντοπίζονται λίγο πάνω από τις μισές επιτυχημένες εταιρίες της αγοράς. Αποτελεί ένα πολύ υψηλό ποσοστό, αν αναλογιστεί κανείς πως με τυχαία επιλογή έχει

πιθανότητα επιτυχίας 10%, ενώ οι πιθανότητες του μετά την αξιοποίηση του μοντέλου είναι 4 φορές περισσότερες.

Ωστόσο, το μεγαλύτερο επίτευγμα της παρούσας εργασίας είναι το μικρό ποσοστό λανθασμένων θετικών (False Positive Rate, FPR), που ισούται με 6.2%. Είναι καθοριστικής σημασίας, διότι οι πολλές ανεπιτυχείς εταιρίες (90%) δεν παραπλανούν το μοντέλο. Αντίθετα, με ανάκληση (Recall) 53.2%, είναι πολύ υψηλές οι πιθανότητες για την εταιρία που προβλέφθηκε ως επιτυχημένη να είναι και στην πραγματικότητα. Το γεγονός πως η πρόταση του μοντέλου έχει πιθανότητα να ισχύει πάνω από 50% δίνει ένα τεράστιο πλεονέκτημα σε αυτόν που θα την αξιοποιήσει, αφού τα κέρδη από μια σωστή επένδυση είναι εκθετικά περισσότερα.

5.2 Μελλοντικές Κατευθύνσεις

Η μηχανική μάθηση τα τελευταία δέκα χρόνια αξιοποιείται σε όλους τους επιστημονικούς και μη τομείς. Ωστόσο, η πρόβλεψη επιτυχίας επιχειρήσεων έχει μελετηθεί σε βάθος από ελάχιστους ερευνητές. Θεωρείται πως η έρευνα στον τομέα θα εντατικοποιηθεί στο άμεσο μέλλον και κρίνεται αναγκαία η παρουσίαση ορισμένων προτάσεων που θα συμβάλλουν στο έργο της επιστημονικής κοινότητας.

Αρχικά, πιο ακριβή και αξιόπιστα δεδομένα είναι απαραίτητα για την επίτευξη υψηλότερης απόδοσης από τα μοντέλα. Όταν οι ελλιπείς τιμές είναι περιορισμένες, τα μοντέλα μελετούν περισσότερη πληροφορία και μπορούν να καταλήξουν σε πιο εξειδικευμένα συμπεράσματα. Στην παρούσα μελέτη απουσίαζαν εντελώς πληροφορίες σχετικά με τους ιδρυτές, το ανθρώπινο δυναμικό και τους συνεργάτες της εκάστοτε εταιρίας, που αποτελούν καθοριστικούς παράγοντες για την ανάπτυξή της στα πρώτα στάδια. Η εμπειρία και το βιογραφικό τους σίγουρα μπορεί να συνεισφέρει στις μελλοντικές προβλέψεις. Επίσης, καθώς η αγορά μεταβάλλεται με ταχύτατους ρυθμούς, πρόσφατα δεδομένα πρέπει να έχουν μεγαλύτερη συμβολή. Ταυτόχρονα, μπορούν να αξιοποιηθούν γνώμες ειδικών του χώρου και οι δηλώσεις των ιδιοκτητών που δημοσιεύονται στο διαδίκτυο και στα μέσα κοινωνικής δικτύωσης.

Μια μελέτη με μεγαλύτερο όγκο δεδομένων θα μπορούσε να μελετήσει ξεχωριστά διαφορετικούς τομείς, ενώ αν διατίθεται ο απαιτούμενος χρόνος, οι προβλέψεις και η επικύρωσή τους μπορεί να πραγματοποιείται ζωντανά σε διάφορα χρονικά διαστήματα. Επιπλέον, ανταγωνιστικές εταιρίες στο ίδιο χρονικό διάστημα έχουν μοιρασμένες πιθανότητες επιτυχίας και πρέπει να λαμβάνεται υπόψη.

Τέλος, πολλές επιχειρήσεις που ιδρύθηκαν πρόσφατα δεν είχαν την ευκαιρία να επιτύχουν. Συνεπώς, μπορούν να διευρυνθούν τα κριτήρια επιτυχίας τους, καθώς αν είναι κερδοφόρες μετά από ένα εύλογο χρονικό διάστημα, μπορούν να θεωρηθούν πετυχημένες.

Βιβλιογραφία

Blank, S. (2020). The four steps to the epiphany: successful strategies for products that win. John Wiley & Sons.

Kirzner, I. M. (2015). Competition and entrepreneurship. University of Chicago press.

Stevenson, H. H., & Jarillo, J. C. (2007). A paradigm of entrepreneurship: Entrepreneurial management. In Entrepreneurship (pp. 155-170). Springer, Berlin, Heidelberg.

Reis, E. (2011). The lean startup. New York: Crown Business, 27.

Thiel, P. A., & Masters, B. (2014). Zero to one: Notes on startups, or how to build the future. Currency.

Graham, P. (2012). Startup= growth. Dostupno na: http://www. paulgraham. com/growth. html,[pristupljeno: 15.8. 2015.].

Startups eV, B. D., Ripsas, S., & Tröger, S. (2014). Deutscher Start-Up Monitor. KPMG in Deutschland.

Ripsas, S., Schaper, B., & Tröger, S. (2018). A startup cockpit for the proof-of-concept. In Handbuch entrepreneurship (pp. 263-279). Springer Gabler, Wiesbaden.

Jyrki Ali-Yrkkö , Ari Hyytinen & Mika Pajarinen (2005) Does patenting increasethe probability of being acquired? Evidence from cross-border and domestic acquisitions, AppliedFinancial Economics, 15:14, 1007-1017, DOI: 10.1080/09603100500186978

Gugler, Klaus, and Kai A. Konrad. "Merger target selection and financial structure." University of Vienna and Wissenschaftszentrum Berlin (WZB) (2002).

Meador, Anna Lee, Pamela H. Church, and L. Gayle Rayburn. "Development of prediction models for horizontal and vertical mergers." Journal of financial and strategic decisions 9.1 (1996): 11-23.

Ragothaman, Srinivasan, Bijayananda Naik, and Kumoli Ramakrishnan. "Predicting corporate acquisitions: An application of uncertain reasoning using rule induction." Information Systems Frontiers 5.4 (2003): 401-412.

Wei, Chih-Ping, Yu-Syun Jiang, and Chin-Sheng Yang. "Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach." Workshop on E-Business. Springer, Berlin, Heidelberg, 2008.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance, 23(4), 589-609.

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision support systems, 50(2), 491-500.

Eugene, L. Y., & Yuan, S. T. D. (2012, August). Where's the money? the social behavior of investors in facebook's small world. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 158-162). IEEE.

Xiang, G., Zheng, Z., Wen, M., Hong, J., Rose, C., & Liu, C. (2012, May). A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 6, No. 1).

Pan, C., Gao, Y., & Luo, Y. (2018). Machine Learning Prediction of Companies 'Business Success. CS229: Machine Learning, Stanford University.

Bento, F. R. D. S. R. (2018). Predicting start-up success with machine learning (Doctoral dissertation).

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018, October). Web-based startup success prediction. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 2283-2291).

Antretter, T., Blohm, I., Grichnik, D., & Wincent, J. (2019). Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy. Journal of Business Venturing Insights, 11, e00109.

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. Ieee Access, 7, 124233-124243.

Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. Information Processing & Management, 58(4), 102555.

Calafiore, G. C., Morales, M. H., Tiozzo, V., & Marquie, S. (2020, May). A Classifiers Voting Model for Exit Prediction of Privately Held Companies. In 2020 European Control Conference (ECC) (pp. 615-620). IEEE.

Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M. (2020). Supervised learning for the prediction of firm dynamics. arXiv preprint arXiv:2009.06413.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). Practical machine learning tools and techniques. Morgan Kaufmann, 578.

Kudyba, Stephan P. Healthcare informatics: improving efficiency through technology, analytics, and management. CRC Press, 2018.

Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." Journal of healthcare information management 19.2 (2011): 65.

Hill, Kashmir. "How target figured out a teen girl was pregnant before her father did." Forbes, Inc (2012).

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): 27-34.

Samuel, A. L. (1962). Artificial intelligence: a frontier of automation. The Annals of the American Academy of Political and Social Science, 340(1), 10-20.

Mitchell, T. M. (1997). Does machine learning really work?. AI magazine, 18(3), 11-11.

Schapire, R. (2015). Machine learning algorithms for classification. Princeton University, 10.

Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), pp.111-117.

Yeates, S., Bainbridge, D. and Witten, I.H., 2000. Using compression to identify acronyms in text. arXiv preprint cs/0007003.

Osborne, J. (2002). Notes on the use of data transformations. Practical assessment, research, and evaluation, 8(1), 6.

Griffith, E. (2014). Why startups fail, according to their founders. Fortune Magazine, September, 25.