Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής & Υπολογιστών

Τεχνικές Μηχανικής Μάθησης σε Προβλήματα Πρόβλεψης Επιτυχίας Εταιρειών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΠΑΕΥΘΥΜΙΟΥ ΔΑΦΝΗ ΠΕΛΑΓΙΑ

Επιβλέπων : Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής & Υπολογιστών

Τεχνικές Μηχανικής Μάθησης σε Προβλήματα Πρόβλεψης Επιτυχίας Εταιρειών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΑΠΑΕΥΘΥΜΙΟΥ ΔΑΦΝΗ ΠΕΛΑΓΙΑ

Επιβλέπων : Δημήτριος Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Μαρτίου 2020.

Δημήτριος Φωτάκης Αν. Καθηγητής Ε.Μ.Π.	Νικόλαος Παπασπύρου Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2020

Παπαευθυμίου Δάφνη Πελαγία

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

Περίληψη

Η αστάθεια και η αβεβαιότητα είναι οι λέξεις που χαρακτηρίζουν με τη μεγαλύτερη ακρίβεια το σύγχρονο επιχειρησιακό περιβάλλον. Ο ψηφιακός μετασχηματισμός, ο μεγάλος ανταγωνισμός και το όλο και πιο υψηλό επίπεδο γνώσεων των ατόμων που πλαισιώνουν τις διάφορες επιχειρήσεις μεταβάλουν ραγδαία τα δεδομένα στον επενδυτικό χώρο και περιορίζουν τους φραγμούς εισόδου στην αγορά. Την ίδια στιγμή, παρατηρείται μια πρωτόγνωρης κλίμακας στροφή των ατόμων σε ό,τι αφορά την ενασχόληση με τους συγκεκριμένους τομείς. Συνέπεια της κατάστασης αυτής αποτελεί η εκθετική αύξηση της ίδρυσης νεογνών εταιρειών - Startup - αλλά και ο επίσης αυξανόμενος ρυθμός αποτυχίας τους.

Στο πλαίσιο αυτής της διπλωματικής εργασίας μελετάται η πρόβλεψη της πιθανότητας επιτυχίας μιας σύγχρονης startup εταιρείας υπό οικονομικό, επιστημονικό και κοινωνικό πρίσμα. Η έγκυρη πρόβλεψη σχετικά με το μέλλον μιας τέτοιου είδους εταιρείας θεωρείται σήμερα ιδιαίτερα σημαντική κυρίως για άτομα και εταιρείες με μεγάλη επενδυτική δύναμη, καθώς είναι επιθυμητό η διαχείριση των κεφαλαίων τους να γίνει με τον πλέον αποδοτικό τρόπο. Έτσι τον τελευταίο καιρό έχει επιχειρηθεί να μοντελοποιηθεί το ”μοτίβο” της επιτυχίας μιας startup εταιρείας από αρκετούς ερευνητές.

Μέχρι σήμερα, το συγκεκριμένο πρόβλημα αποτελούσε κυρίως ένα πιο θεωρητικό πεδίο μελέτης της επιχειρησιακής έρευνας και για την προσέγγισή του εξετάζονταν και αξιολογούνταν ως επί των πλείστων επιχειρηματικά σχέδια και μοντέλα δόμησης των εταιρειών. Ωστόσο, τα ερεθίσματα που μπορεί να λάβει μια νέα εταιρεία και να την οδηγήσουν στην αποτυχία ή την επιτυχία είναι πολυδιάστατα και συχνά μη αλληλοσυνδεόμενα και δεν περιορίζονται μόνο στο επιχειρηματικό πλάνο που θα ακολουθήσει.

Στην παρούσα εργασία, επιχειρούμε να δώσουμε μία λύση στο πρόβλημα που παρουσιάστηκε, όπου θα λαμβάνονται υπόψη πιο πρακτικοί και εύκολα μετρήσιμοι παράγοντες και χαρακτηριστικά των εταιρειών, μέσω της εφαρμογής μεθόδων μηχανικής μάθησης. Καταγράφουμε έπειτα τα αποτελέσματα των διάφορων μεθόδων που εφαρμόσαμε και σχολιάζουμε την αποδοτικότητα και την καταλληλότητά τους για το συγκεκριμένο πρόβλημα. Αναλυτικότερα, περιορίζουμε το χώρο μελέτης μας στις εταιρείες που δραστηριοποιούνται στους τομείς της βιοτεχνολογίας και της υγείας. Επιδιώκουμε να αποφανθούμε για το αν, πέρα από τα οικονομικά κριτήρια, υπάρχει αντίστοιχη επιστημονική γνώση για να πλαισιώσει το έργο τους και να τους οδηγήσει στην εδραίωση της παρουσίας τους στον επιχειρηματικό χώρο και τελικά την επιτυχία.

Αρχικά λοιπόν, γίνεται περιγραφή του γενικού πλαισίου και της ιστορίας και παρουσίαση της περίπτωσης προβλήματος υπό εξέταση. Ακολουθεί η περιγραφή των δεδομένων, η ανάλυση των βασικών χαρακτηριστικών τους και η επεξεργασία που απαιτήθηκε για να είναι δυνατή η κατανόηση τους και η αναγνώριση συσχετίσεων.

Σχεδιάστηκαν στη συνέχεια τρία διαφορετικά μοντέλα μηχανικής μάθησης που εκπαιδεύονται με βάση τα δεδομένα αυτά και έπειτα μπορούν να γενικεύσουν αυτή τη γνώση που αποκτήσανε, για τον υπολογισμό της αναμενόμενης τιμής δεδομένων αγνώστων σε εκείνα. Τα μοντέλα αυτά παραμετροποιούνται και εκπαιδεύονται κατάλληλα ώστε να μπορούν να διαχωρίσουν τα δεδομένα σε δύο κλάσεις που περιγράφουν τις επιτυχημένες και τις αποτυχημένες εταιρείες αντίστοιχα . Στη συνέχεια παρουσιάζονται και αξιολογούνται τα αποτελέσματα της εφαρμογής τους τόσο στα δεδομένα εκπαίδευσης όσο και σε ένα άγνωστο σύνολο δεδομένων.

Παράλληλα, με την παρουσίαση της μεθοδολογίας που ακολουθούμε, των προβλημάτων που κληθήκαμε να αντιμετωπίσουμε καθώς και των λύσεων που επιστρατεύσαμε ευελπιστούμε να προ-

σφέρουμε, με την παρούσα διπλωματική, ένα μοτίβο και έναν ενδεικτικό τρόπο εργασίας για τους μελετητές που θα αποφασίσουν να ασχοληθούν με παρόμοια θεματολογία ώστε να πετύχουν τα βέλτιστα δυνατά αποτελέσματα στον καλύτερο δυνατό χρόνο.

Λέξεις κλειδιά

startups, επιτυχία εταιρείας, πρόβλεψη,τυχαία δάση , μηχανική μάθηση, λογιστική παλινδρόμηση, , σύνολο δεδομένων, ανάλυση δεδομένων

Abstract

Uncertainty and obscurity are the two most accurate words to describe the current business environment. The digital transformation, the high competition and the constantly increasing knowledge level of people who frame the various business are rapidly changing the investment industry and limiting the barriers to entry in the markets. At the same time there is a remarkable tendency of people, especially young ones, towards working in these industries. As a result of this situation, one can observe an exponential growth of newly founded companies - Startups - and at the same time an increasing number of companies which are ceasing operations due to business failure.

The present work is focusing on predicting the possibility of success of a startup company. We will attempt to study this problem from an economic, scientific and social point of view. The accurate prediction regarding the future of a newly born company is considered today highly important, especially for shareholders and investors as they opt for an efficient and profitable capital management. Thus, the last years, there have been numerous attempts to specify the profile of a successful company.

Until today, this specific problem was concerning a more theoretical field of operational research as in order to approach it researchers were mostly focusing on reviewing business plans and models. However, the factors that can contribute to the making of a successful company are multidimensional, often unrelated and certainly more complex.

With this work of ours, we attempt to provide solution to the previously described problem, taking into account more practical and easily measurable factors and features through the application of machine learning algorithms. Moreover, we present the results of the various algorithms that we have implemented and we comment on their suitability for the specific problem. We are going to restrict our field of interest in companies active in the health and biotechnology sectors. We seek to determine if, along with the economical features, the current scientific knowledge is adequate to frame theri work and to establish their presence in the business and finally to lead to their success.

In more detail, after describing the general problem and its history, we move on the presentation of the case that we take into consideration in this thesis. Then, we describe the complete process of forming our dataset and preprocessing it as well as present its features and investigate the relationships between them.

Consequently, three different machine models are being designed and trained based on the dataset in order to generalize the knowledge gained so as to be able to make predictions concerning the success of a startup company which is completely unknown to them . The models' parameters are being tuned so that we have a highly accurate classification of companies into successful and unsuccessful ones. Afterwards, an overall evaluation of our models' is being presented

Along with the presentation of the overall process, as well as the problems that we faced and our proposed solutions we are hoping to provide, with this present work, a baseline in order to guide the researchers that are going to take on a similar project, towards a more efficient and less time consuming solution.

Key words

startups, machine learning , logistic regression, dataset, data analysis, random forests, success prediction, gradient descent on decision trees.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Δημήτριο Φωτάκη για τη διακριτική του καθοδήγηση κατά τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας. Σε κάθε στάδιο της εργασίας μου, με βοήθησε να ξεπεράσω δυσκολίες και αδιέξοδα παρακινώντας με να θέτω τα σωστά ερωτήματα σε κάθε περίπτωση και να εμπιστεύομαι τις ικανότητές μου ώστε να οδηγούμαι στις σωστές απαντήσεις. Πιστεύω ότι αυτή η εργασία θα είναι η αρχή ενός μεγαλύτερου ταξιδιού στο χώρο της έρευνας, της μηχανικής μάθησης και των αλγορίθμων. Θα ήθελα ακόμα να ευχαριστήσω την οικογένειά μου για την αμέριστη υποστήριξή της στις εύκολες αλλά και κυρίως στις δύσκολες στιγμές που πέρασαν κατά τη διάρκεια όλων των φοιτητικών μου χρόνων. Τέλος, ένα μεγάλο ευχαριστώ στους φίλους μου που με συνοδεύουν γεμίζοντας με χαρά και αισιοδοξία στο ταξίδι της ζωής!

Παπαευθυμίου Δάφνη Πελαγία,

Αθήνα, 30η Μαρτίου 2020

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Νεοφυείς επιχειρήσεις - Startup

Οι ανάγκες του ανθρώπου για επιχειρηματική δραστηριότητα, ικανοποιούνται με διαφορετικούς τρόπους σε κάθε χρονική περίοδο της ιστορίας οι οποίοι συσχετίζονται με τα αντίστοιχα ποσοστά τεχνολογικής , κοινωνικής και πολιτικής κατάστασης της εκάστοτε εποχής. Τα τελευταία λοιπόν χρόνια, παρατηρείται η εμφάνιση και ο πολλαπλασιασμός ενός είδους εταιρείας που ακούει στο όνομα startup. Πρόκειται για μία έννοια της οποίας ο ορισμός δεν έχει καθοριστεί με σαφήνεια καθώς ακόμα και αυθεντίες στον τομέα των επιχειρήσεων δεν έχουν καταλήξει σε μία συγκεκριμένη περιγραφή. Για τον Eric Reis, συγγραφέα του βιβλίου The lean startup Ries (2011) η λέξη startup "Περιγράφει ένα ίδρυμα κατασκευασμένο από τον άνθρωπο με στόχο τη δημιουργία ενός νέου προϊόντος ή υπηρεσίας, κάτω από καταστάσεις ακραίας αβεβαιότητας." Πρόκειται για δομές που δημιουργούνται από μια μικρή ομάδα ατόμων οι οποίες έχουν στόχο να εισάγουν μια καινοτομία στην αγορά, να καλύψουν κάποιο πιθανό κενό, να ανατρέψουν τις κοινώς χρησιμοποιούμενες πρακτικές και να αποδειχτούν παράλληλα ανταγωνιστικές στο σύγχρονο, απαιτητικό επιχειρηματικό περιβάλλον. Αποτελούν εταιρείες στα πρώιμα στάδια λειτουργία τους που έχουν πρόθεση να επεκταθούν , να ισχυροποιηθούν και τελικά να γίνουν προσοδοφόρες.

Τα ποσοστά επιτυχίας όμως των δομών αυτών είναι εξαιρετικά μικρά. Ο Neil Patel, βραβευμένος συγγραφέας των New York Times και εταιρικός σύμβουλος αναγνωρισμένων εταιρειών ισχυρίζεται ότι μόλις μία στις δέκα startups καταφέρνει και επιβιώνει. Παράλληλα, ο καθηγητής του Harvard Buisiness school Shikhar Ghosh αποκαλύπτει πως ακόμα και στις startup που έχουν κάποιας μορφής αρχικό κεφάλαιο, τα ποσοστά αποτυχίας είναι 75%. Υποστηρίζει μάλιστα ότι «Λίγες startup πετυχαίνουν τους αρχικούς τους στόχους. Η αποτυχία είναι ο κανόνας ». Γεννούνται έτσι πολλά ερωτήματα σχετικά με τους παράγοντες που μπορούν να συμβάλλουν στην επιτυχία ή την αποτυχία των εταιρειών αυτών. Σαφώς δε λείπουν και σχετικές, ερευνητικές κυρίως μελέτες που προσπαθούν να απαντήσουν σε αυτά τα ερωτήματα ακολουθώντας μεθόδους που σχετίζονται με συλλογή απόψεων, εμπειριών, διαμόρφωση και ανάλυση ερωτηματολογίων και αφορούν τους άμεσα εμπλεκόμενους, οι οποίοι συνήθως είναι οι ιδρυτές, κάποιοι εργαζόμενοι και ίσως πελάτες.

Μια από τις πιο κοινές αιτιολογίες που παρουσιάζεται από τις μελέτες αυτές ως βασική αιτία της αποτυχίας, αφορά την κατάσταση και τις ανάγκες της αγοράς, υποστηρίζοντας ότι απλά το προϊόν ή η υπηρεσία που προσφέρει η startup δεν κάλυπταν κάποιο κενό της αγοράς τη δεδομένη χρονική στιγμή. Επιπλέον ποιοτικά χαρακτηριστικά που αναφέρονται ως αίτια της αποτυχίας αφορούν τα χαρακτηριστικά της ομάδας καθώς και τις επιχειρησιακές στρατηγικές που ακολουθούνται και πιθανόν δεν είναι οι κατάλληλες. Διαβάζοντας διάφορες μελέτες και εντοπίζοντας τα

κοινά τους σημεία, σχηματίζεται ένα μοντέλο της τέλειας υποψήφιας startup ως μια επιχείρηση που ιδρύεται από έμπειρα, εφευρετικά άτομα που έχουν εντοπίσει μία ανικανοποίητη ανάγκη ενός μεγάλου μέρους του πληθυσμού και λειτουργούν γνωρίζοντας τους 'νόμους' της αγοράς, εφαρμόζοντας τις κατάλληλες επιχειρησιακές πρακτικές όντας έτοιμα να προσαρμοστούν στις νέες συνθήκες ενός αστάθμητου περιβάλλοντος.

Και ενώ τα παραπάνω φαντάζουν πολύ θεωρητικά, παρόμοιες μελέτες συνεχίζουν να δημοσιεύονται με συχνότατο ρυθμό γεγονός που μπορεί να μας οδηγήσει παρά μόνο σε ένα συμπέρασμα, η τύχη αυτών των εταιρειών έχει πολύ μεγάλο αντίκτυπο στη διεθνή οικονομία. Ενδεικτικά, τα συνολικά ετήσια κεφάλαια επιχειρηματικών συμμετοχών που αφορούν startups μπορούν να φτάνουν εως και τα 155 δισεκατομμύρια. Στα πλαίσια αυτά, επενδυτές αλλά και ιδρυτές έχουν ανάγκη από ένα βοηθητικό μηχανισμό που να αποφαίνεται για την πιθανότητα επιτυχίας της νέας αυτής επιχείρησης.

1.2 Startup στον τομέα της υγείας

Από επενδυτικής άποψης μιλώντας, οι starup που ασχολούνται με τον τομέα της υγείας συγκεντρώνουν τα τελευταία χρόνια και τα μεγαλύτερα κεφάλαια. Από τη μία μπορούμε να φανταστούμε ότι μία εταιρεία βιοτεχνολογίας ή κατασκευής ιατρικού εξοπλισμού θα είχε περισσότερες δαπάνες από εταιρείες που καταπιάνονται με διαδικτυακά, για παράδειγμα, ζητήματα ή θέματα ασφάλειας, λογισμικού, ηλεκτρονικών παιχνιδιών κλπ. Από την άλλη, εντοπίζουμε σε αυτήν την επενδυτική τάση την ανάγκη που αισθάνεται ο άνθρωπος να καταφέρει να υπερνικήσει φυσικούς περιορισμούς που μπορούν να σταθούν εμπόδιο στην κανονικότητα της καθημερινότητάς του. Μελετώντας δεδομένα από οργανισμούς υγείας και στατιστικές μελέτες παρατηρήσαμε πως από τα τέλη του 20ου αιώνα η ίδρυση εταιρειών αυτού του αντικειμένου πολλαπλασιάζονταν με ταχύτατο ρυθμό. Βέβαια υψηλότατος είναι και ο ρυθμός αποτυχίας των συγκεκριμένων εταιρειών. Οι λόγοι πίσω από αυτή την παρατήρηση είναι λίγο πολύ εμφανείς. Αρχικά, όσον αφορά τις φαρμακευτικές εταιρείες ο μέσος όρος της εισαγωγής ενός φαρμάκου στην αγορά είναι 12 χρόνια και αυτά μη υπολογίζοντας το διάστημα της ανακάλυψης-δημιουργίας του νέου αυτού φαρμάκου κατά το οποίο είναι δυνατό η εταιρεία να οδηγηθεί σε αδιέξοδα. Στα ίδια πλαίσια, για την εισαγωγή ενός ιατρικού μηχανήματος στην αγορά (FDA Approval) πρέπει να ακολουθηθεί αυστηρό πρωτόκολλο αξιολόγησής του, το οποίο συχνά περιλαμβάνει πολλαπλές φάσεις στις οποίες η πιθανότητα επιτυχίας φθίνει σχεδόν εκθετικά. Στο χώρο αυτόν την τελευταία δεκαετία έχουν προστεθεί και οι εταιρείες που ασχολούνται με τη γενετική, που δρουν κυρίως προσπαθώντας να αποκωδικοποιήσουν το ανθρώπινο γονιδίωμα σε ένα ερευνητικό πλαίσιο, μη γνωρίζοντας κατά πόσο οι μελέτες τους θα αποδώσουν ουσιαστικούς καρπούς. Την τελευταία δεκαετία με την επανάσταση στη ψηφιακή τεχνολογία και την κινητή τηλεφωνία παρατηρούμε επίσης την ανάπτυξη ενός τομέα που ασχολείται με εφαρμογές ηλεκτρονικής υγείας που περιέχουν ένα ευρύ φάσμα προϊόντων, συστημάτων και εργαλείων, τα οποία βασίζουν τη λειτουργία τους στις εξελιγμένες τεχνολογίες πληροφοριών και επικοινωνιών. Πρόκειται για εγχειρήματα που απαιτούν σαφώς μικρότερη χρηματοδότηση από όσα προαναφέρθηκαν, αποδεικνύονται όμως εξίσου δύσκολα και απαιτητικά. Άλλοι τομείς που έχουν γνωρίσει ανάπτυξη τα τελευταία χρόνια παράλληλα με την τεχνολογική άνθιση αφορούν την αξιοποίηση της ρομποτικής στον τομέα της υγείας, της τεχνητής νοημοσύνης

στην ανακάλυψη νέων φαρμάκων και θεραπειών, τη διάγνωση ασθενειών καθώς και την απομακρυσμένη παρακολούθηση ασθενών μέσω συστημάτων αισθητήρων.

1.3 Παρουσίαση υπάρχουσας βιβλιογραφίας και βασικών παραγόντων διαφοροποίησης

Το συγκεκριμένο ερώτημα που προσπαθήσαμε να απαντήσουμε, το αν δηλαδή μια εταιρεία θα επιτύχει ή θα αποτύχει, έχουν προσπαθήσει να απαντήσουν λίγοι ακόμα μελετητές επιστρατεύοντας τεχνικές μηχανικής μάθησης. Η ίδια η ερώτηση όμως που θέτουμε δεν είναι σαφώς ορισμένη και επιτρέπει να τις δοθούν διαφορετικές ερμηνείες οδηγώντας και στην επίλυση διαφορετικών προβλημάτων. Αρχικά, πρέπει να αναρωτηθούμε και να ορίσουμε τι εννοούμε με τις λέξεις επιτυχία-αποτυχία. Το 2012 οι Xiang et al. (2012) ασχολήθηκαν με το ζήτημα της πρόβλεψης συγχώνευσης/εξαγοράς των εταιρειών, συλλέγοντας δεδομένα από την ίδια βάση που χρησιμοποιήσαμε και εμείς. Στην περίπτωση αυτή λοιπόν, επιτυχία θεωρήθηκε το συμβάν συγχώνευσης - εξαγοράς ενώ η απουσία του μεταφράστηκε ως αποτυχία. Έτσι αποτυχημένες θεωρήθηκαν όχι μόνο οι εταιρείες που πλέον δεν λειτουργούσαν αλλά και όσες ήταν εν λειτουργία καθώς και όσες εισήλθαν στο χρηματιστήριο μέσω αρχικής δημόσιας προσφοράς (IPO). Το σύνολο των εταιρειών που συγκεντρώθηκε περιείχε startup που άνηκαν σε διάφορες κατηγορίες ως προς το αντικείμενο ενασχόλησεις τους. Στην παραπάνω μελέτη η υλοποίηση του ταξινόμηση έγινε με μοντέλα λογιστικής παλινδρόμησης και Support Vector Machines .

Στη συνέχεια, το 2018 έγινε μια παρόμοια μελέτη από τον da Silva Ribeiro Bento (2018) που χρησιμοποίησε την ίδια βάση και προσπάθησε να απαντήσει στο ίδιο ερώτημα. Μετάφρασε ωστόσο το εν λόγω πρόβλημα στο εξής ακόλουθο «Η επιτυχία μιας startup ορίζεται συνήθως ως μια διπλής φύσεως περίπτωση καθώς μια εταιρεία μπορεί είτε να θέσει σε δημόσια κυκλοφορία μετοχές (Αρχική δημόσια προσφορά-IPO), επιτρέποντας στους μετόχους της να πουλήσουν τις μετοχές τους, ή να αγοραστεί ή συγχωνευθεί με μια άλλη εταιρεία και έτσι όσοι είχαν προηγουμένως επενδύσει σε αυτήν να λάβουν απευθείας μετρητά σε αντάλλαγμα για τις μετοχές τους.» Αμέσως παρατηρούμε ότι το φαινομενικά ίδιο πρόβλημα τροποποιείται σε ένα εντελώς διαφορετικό αφού πλέον ίδιες εταιρείες θα ανήκουν σε διαφορετικές κατηγορίες (επιτυχία-αποτυχία) σε σχέση με προηγουμένως. Οι τεχνικές μηχανικής που επιστράτευσε για την επίλυση του προβλήματος περιλαμβάνουν τη Λογιστική Παλινδρόμηση, τα Διανύσματα υποστήριξης και τα Τυχαία Δάση. Τα αποτελέσματα δε που έλαβε ήταν αρκετά ενθαρρυντικά και θα συζητηθούν στη συνέχεια.

1.4 Ορισμός προβλήματος με το οποίο θα ασχοληθούμε

Στη δική μας προσέγγιση, αποφασίσαμε να ορίσουμε μία πιο ξεκάθαρη διαφοροποίηση για τις επιτυχημένες - αποτυχημένες εταιρείες θεωρώντας ως αποτυχημένες όσες είναι κλειστές και ως επιτυχημένες εκείνες που είναι εν λειτουργία. Επιπλέον, ασχοληθήκαμε αποκλειστικά με τον τομέα της υγείας καθώς θελήσαμε να καθοδηγήσουμε τα μοντέλα μας ώστε να λάβουν έναν επιπλέον παράγοντα υπόψη που αφορά την ωριμότητα του εκάστοτε επιστημονικού πεδίου, το κατά πόσο δηλαδή είναι ικανή η υπάρχουσα επιστημονική γνώση να πλαισιώσει τις προσπάθειες των εταιρειών και να τις οδηγήσει σε επιτυχία ή αποτυχία. Παρακινούμενοι από τις απρόσμενες εξελίξεις

που σχετίζονται με την συνταρακτική αποτυχία της startup Theranos, εταιρείας που υποστήριξε ότι θα καταφέρει να απλοποιήσει δραματικά την διαδικασία αιματολογικού ελέγχου, προσπαθήσαμε να αποφανθούμε για το αν είναι ανθρωπίνως εφικτό, ανεξάρτητα από οικονομικούς - συγκυριακούς παράγοντες, για μια startup να πετύχει τους στόχους της.

Ένα ακόμα στοιχείο που μας προβλημάτισε κατά την παρούσα μελέτη αφορούσε τη χρονικότητα των συμβάντων. Αντλήσαμε πληροφορίες από δύο βάσεις δεδομένων όπου η μία αποτελείται από δεδομένα του 2013 και η άλλη του 2019. Αυτό στάθηκε αφορμή για να αντιληφθούμε τη σπουδαιότητα της επίγνωσης χρονοσειράς των γεγονότων που οδηγούν μια startup στην επιτυχία και την αποτυχία.

Με βάση τα παραπάνω συγκεκριμενοποιήσαμε το πρόβλημα μελέτης μας στο εξής: Έχοντας συγκεντρώσει πληροφορίες και χαρακτηριστικά για startup εταιρείες στον τομέα της υγείας μέχρι το 2013 προσπαθήσαμε να προβλέψουμε αν οι συγκεκριμένες εταιρείες θα λειτουργούν το 2019 ή όχι.

1.5 Δομή Διπλωματικής Εργασίας

Στο Κεφάλαιο 2, πρόκειται να περιγραφεί το τεχνικό υπόβαθρο, με τις αντίστοιχες παραπομπές στα πρωτότυπα κείμενα, που απαιτείται για την πλήρη κατανόηση της συγκεκριμένης διατριβής. Παρουσιάζουμε εδώ τις βασικές τεχνικές που θα χρησιμοποιηθούν για την προσέγγιση του ζητήματος με το οποίο ασχολούμαστε. Αρχικά, γίνεται αναφορά στην επιστήμη των δεδομένων και τη σημασία της για την πρόοδο των επιχειρήσεων. Στη συνέχεια αναλύεται εν συντομία η επιστήμη της μηχανικής μάθησης και ακολουθεί εκτενέστερη περιγραφή των τεχνικών μηχανικής μάθησης που επιστρατεύονται.

Στο κεφάλαιο 3, αποτυπώνεται η περιγραφή της μεθοδολογίας που επιλέχθηκε στα πλαίσια αυτής της εργασίας. Περιγράφεται αναλυτικά η διαδικασία απόκτησης και επεξεργασίας των δεδομένων. Δίνεται αρχικά, η περιγραφή του τρόπου απόκτησης των δεδομένων που χρησιμοποιήθηκαν καθώς και της μετέπειτα επεξεργασίας τους που κρίθηκε καθοριστική για την επιτυχία του εγχειρήματος μας. Έπειτα αναλύονται οι σχεδιαστικές επιλογές που κάναμε κατά την ανάπτυξη των μοντέλων μας και τα βήματα υλοποίησης που ακολουθήσαμε. Αναπτύσσονται προβλήματα που προέκυψαν, σχετικά με τη διαχείρηση των δεδομένων και η λογική επίλυσής τους. Καταλήγουμε τέλος στην ακριβή σκιαγράφηση της διαδικασίας που ακολουθήσαμε μέχρις ότου να έρθει η στιγμή της χρήσης των αλγόριθμων μηχανικής μάθησης.

Στο κεφάλαιο 4, δίνεται η παρουσίαση των τελικών αποτελεσμάτων και η αξιολόγηση τους. Ιδιαίτερη έμφαση δίνεται στη σύγκριση των χρησιμοποιούμενων μοντέλων με κριτήριο την καταλληλότητά τους για το συγκεκριμένο πρόβλημα. Επιπλέον γίνεται σύγκριση των διαφορετικών συνόλων δεδομένων που χρησιμοποιήσαμε και ακόμα επιχειρείται να γίνει μία ταξινόμηση των χαρακτηριστικών εκπαίδευσης με βάση τη σπουδαιότητά τους - επιρροή τους στο πρόβλημά μας. Ακολουθεί μια σύγκριση των αποτελεσμάτων μας με υπάρχουσες μελέτες και επιχειρείται να εξηγηθούν τυχόν αποκλίσεις και διαφορές. Τέλος, παρουσιάζεται μια σειρά από παρατηρήσεις, διαπιστώσεις και συμβουλές ώστε να προκύψει μια μοντελοποίηση της δουλειάς που υλοποιήσαμε στοχεύοντας στη διευκόλυνση της μελλοντικής επιστημονικής μελέτης.

Στο κεφάλαιο 5, καταλήγουμε στα συμπεράσματα καθώς και σε προτάσεις για μελλοντική εξέλιξη. Παρατηρούμε σε τι υπερτερεί και τι υστερεί η κάθε μέθοδος και παρέχονται προτάσεις για

εξέλιξη της παρούσας μελέτης και ιδέες για παρεμφερείς εφαρμογές που μπορεί να παρουσιάζουν ενδιαφέρον.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Επιστήμη Ανάλυσης Δεδομένων

Μία από τις συνέπειες τις high-tech εποχής στην οποία ζούμε αποτελεί η συσσώρευσή ασύλληπτων ποσοτήτων ακατέργαστης πληροφορίας, που είναι αδύνατον για κάποιον άνθρωπο να αφομοιώσει. Ωστόσο, η κατάλληλη επεξεργασία αυτών των δεδομένων κρίνεται απαραίτητη για την εξέλιξή και την επιβίωση σε έναν κόσμο που καθορίζεται από τις ψηφιακές τεχνολογίες. Η διαδικασία της αποθήκευσης, συντήρησης, επεξεργασίας των δεδομένων είναι απαραίτητη για την εκμαίευση ουσιαστικά χρήσιμης πληροφορίας.

Η παραπάνω ακολουθία ενεργειών που προκύπτει από την ανάγκη των ανθρώπων να καταλάβουν το νόημα και να αξιοποιήσουν τα συσσωρευμένα δεδομένα-πληροφορίες είναι εμφανής ήδη από την αρχαιότητα, από πρακτικές όπως η αποθήκευση πληροφοριών για τις πληρωμές των φόρων σε δισκία αργίλου στη Μεσοποταμία το 2000πχ ή η χρήση κοκάλων για την μέτρηση των αγαθών, τη σύγκρισή τους και την πρόβλεψη της διάρκειας των αποθεμάτων στην Ανατολή 9000 χρόνια πίσω.

Σήμερα, την ανάγκη της κάλυψης του αυξανόμενου χάσματος μεταξύ της υπερπαραγωγής δεδομένων και της ουσιαστικής κατανόησής τους καλείται να αντιμετωπίσει η επιστήμη της εξόρυξης δεδομένων. Πρόκειται για μια προσπάθεια του ανθρώπου να ανακαλύπτει μοτίβα στα δεδομένα ώστε να λύνει προβλήματα χρησιμοποιώντας ήδη υπάρχουσα πληροφορία. Witten and Frank (2011)

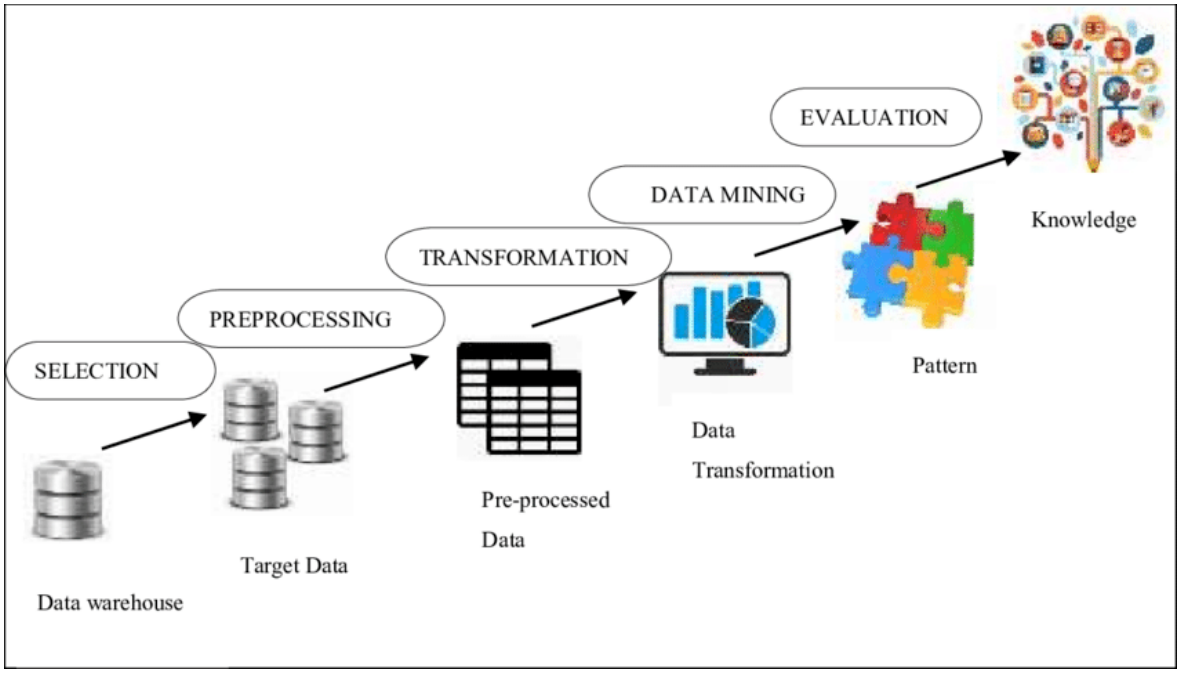
Στον τομέα των επιχειρήσεων όπου ο ανταγωνισμός και οι απαιτήσεις είναι υψηλές, η αξιοποίηση της υπάρχουσας πληροφορίας είναι μονόδρομος. Τα τελευταία χρόνια πρωτόγνωρα μεγάλες ποσότητες συσσωρευμένης και ακατέργαστης πληροφορίας περιγράφονται από τις λέξεις 'big data'. Σύμφωνα με τον Geoffrey Moore, επιχειρησιακό σύμβουλο και συγγραφέα, χωρίς 'big data' οι εταιρείες είναι τυφλές και κουφές στο διαδίκτυο σαν ένα ανυποψίαστο ζώο σε δρόμο ταχείας κυκλοφορίας. Συνεπώς η εξόρυξη δεδομένων είναι άρρηκτα συνδεδεμένη με την επιχειρησιακή έρευνα και καθοριστική για την επιβίωση των εταιρειών. Στα πλαίσια αυτά προέκυψε στις αρχές του αιώνα η ανάγκη για μια συγκεκριμενοποίηση της διαδικασίας αυτής καθαυτής και έναν αυστηρότερο καθορισμό των ενεργειών που περιλαμβάνει. Το 1996 οι Fayyad et al. (1996) ορίσαν την ανακάλυψη γνώσης σε βάση δεδομένων (KDD) ως τη nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Σύμφωνα με αυτούς η εξόρυξη δεδομένων αποτελεί απλά ένα βήμα στη συνολική διαδικασία της ανακάλυψης γνώσης. Συγκεκριμένα, τα κυριότερα στάδια της διαδικασίας αυτής είναι:

* Η συλλογή των δεδομένων

* Η προεπεξεργασία και ο καθαρισμός των δεδομένων

* Η επεξεργασία και κατάλληλη διαμόρφωση των χαρακτηριστικών των δεδομένων με στόχο την προσαρμογή της διαστατικότητάς τους

* Η εξόρυξη δεδομένων με τη χρήση κατάλληλου για το πρόβλημα αλγορίθμου



Σχήμα 2.1: Ανακάλυψη γνώσης από βάση δεδομένων

Η παραπάνω μεθοδολογία ενδέχεται να εμφανίζει κύκλους μεταξύ ενός ή παραπάνω βημάτων ή απουσία ορισμένων εξ'αυτών ανάλογα με τη μορφή του προβλήματος. Το συμπέρασμα στο οποίο καταλήγουμε εδώ είναι ότι το αποτέλεσμα της εξόρυξης δεδομένων που είναι ο βασικός στόχος όλης της διαδικασίας, πρόκειται για ένα πιο σύνθετο πρόβλημα σε κάθε βήμα του οποίο πρέπει να αφιερώσουμε τη δέουσα προσοχή για να λάβουμε και συνολικά τα βέλτιστα αποτελέσματα.

Ειδικά για το κομμάτι της εξόρυξης δεδομένων είθισται να χρησιμοποιείται κάποια τεχνική μηχανικής μάθησης , αναγνώρισης μοτίβων ή στατιστικής. Αξίζει τέλος να σημειωθεί ότι κανείς πρέπει να είναι ιδιαίτερα προσεκτικός σε όποια διαδικασία αφορά τη διαχείριση δεδομένων ώστε να μην κατευθύνει ακούσια την έρευνά του και οδηγηθεί σε λανθασμένα συμπεράσματα καθώς και να κατανοήσει πως ενδέχεται τα δεδομένα που έχει να μην 'κρύβουν' την πληροφορία που θέλει να αποσπάσει. Χαρακτηριστικά, ο οικονομολόγος Roland Coase ανέφερε:« Αν βασανίσεις τα δεδομένα αρκετά, θα ομολογήσουν.»

2.2 Μηχανική Μάθηση

Η μηχανική μάθηση αποτελεί έναν από τους πιο ραγδαία αναπτυσσόμενους τομείς της Τεχνητής Νοημοσύνης, ΤΝ. Οι δύο όροι συχνά συγχέονται όμως ουσιαστικά έχουν τη σχέση

υποσυνόλου – υπερσυνόλου. Ο Bellman το 1978, όρισε την Τεχνητή Νοημοσύνη ως «η αυτοματοποίηση των δραστηριοτήτων που συσχετίζουμε με την ανθρώπινη σκέψη, όπως η λήψη αποφάσεων, η επίλυση προβλημάτων , η μάθηση...» Bellman (1978). Σε αντίθεση με τον Bellman που χρησιμοποιεί ως μέτρο απόδοσης την εγγύτητα ως προς τις ανθρώπινες επιδόσεις, άλλοι μελετητές προσεγγίζουν Τεχνητή Νοημοσύνη, αξιολογώντας την επιτυχία σε σχέση με μία ιδανική έννοια νοημοσύνης, την ορθολογικότητα (rationality).

Από την άλλη ως μηχανική μάθηση ορίζεται από τον Hill η μελέτη των αλγορίθμων που επιτρέπουν στα υπολογιστικά προγράμματα να βελτιώνονται αυτόματα μέσω της εμπειρίας Mitchell (1997) .Το χαρακτηριστικό δηλαδή που διαχωρίζει τη μηχανική μάθηση από άλλες εφαρμογές της ΤΝ είναι η ικανότητα των αλγορίθμων να δέχονται νέα στοιχεία από το περιβάλλον τους, να προσαρμόζονται και να τροποποιούνται. Οι διαδικασίες με τις οποίες λειτουργούν οι αλγόριθμοι αυτοί μηχανικής μάθησης εισάγουν και μια έννοια κατηγοριοποίησης. Ακριβώς όπως υπάρχουν διαφορετικοί τρόποι με τους οποίους 'μαθαίνει' ένας άνθρωπος, το ίδιο ισχύει και για τους αλγορίθμους μηχανικής μάθησης που διαχωρίζονται στις κατηγορίες μάθηση με εκπαιδευτή , μάθηση χωρίς εκπαιδευτή.

1. Μάθηση με εκπαιδευτή ή επιβλεπόμενη μάθηση

   Στην περίπτωση αυτή ο αλγόριθμος-εκπαιδευτής έχει μια γνώση του περιβάλλοντός του, με την έννοια ότι δέχονται ένα σύνολο παραδειγμάτων εισόδου – εξόδου και έτσι γνωρίζει σε κάθε περίπτωση το επιθυμητό αποτέλεσμα. Το περιβάλλον αυτό είναι άγνωστο αρχικά στον αλγόριθμο και γίνεται γνωστό μέσω μιας διαδικασίας μάθησης όπου αρχίζει να τροφοδοτείται δεδομένα και να παράγει αποκρίσεις. Η επιθυμητή απόκριση, γνωστή στον εκπαιδευτή, αντιπροσωπεύει τη βέλτιστη ενέργεια που πρέπει να εκτελείται από τον αλγόριθμο και η διαφορά μεταξύ της επιθυμητής με την πραγματική απόκρισή του καθορίζει και την προσαρμογή των παραμέτρων του. Τελικά, ο αλγόριθμος αποθηκεύει τη γνώση που έλαβε με τη μορφή σταθερών συναπτικών βαρών τα οποία αντιπροσωπεύουν μία μακροπρόθεσμη μνήμη.

2. Μάθηση χωρίς εκπαιδευτή

   Στην περίπτωση αυτή δεν υπάρχει κάποιος εκπαιδευτής για να επιβλέπει τη διαδικασία μάθησης, δηλαδή δεν υπάρχουν χαρακτηρισμένα παραδείγματα σχετικά με τη γνώση που πρέπει να αποκτήσει το σύστημά μας. Εδώ, διακρίνουμε δύο κατηγορίες:

   (a) Ενισχυτική μάθηση

   Στην ενισχυτική μάθηση το feedback για την αξιολόγηση των εξόδων προκύπτει από την αλληλεπίδραση του αλγορίθμου με το περιβάλλον του. Η πρακτική που ακολουθείται έχει ως εξής: κάθε στιγμή μελετάται η τρέχουσα κατάσταση και λαμβάνεται μια απόφαση. Στη συνέχεια μέσω ενός συστήματος λαμβάνει ένα σήμα – ερέθισμα από το περιβάλλον του, το οποίο έχει μεταβεί σε μία νέα κατάσταση λόγω της απόφασης του προηγούμενου βήματος, και βάση αυτού αξιολογεί την ποιότητά της.

   (b) Μη επιβλεπόμενη μάθηση

   Εφόσον τα δεδομένα εισόδου δεν έχουν επιθυμητή έξοδο, εδώ οι αλγόριθμοι αξιολογούν τα αποτελέσματα της μάθησης με βάση κάποιας ανεξάρτητης από τη δια-

δικασία μετρικής και τροποποιούν έτσι προσαρμόζουν και τα βάρη τους. Χαρακτηριστικά παραδείγματα είναι η διαδικασία ομαδοποίησης δειγμάτων ή συσχέτισης.

Παρακάτω περιγράφεται συνοπτικά ο τρόπος λειτουργίας τεσσάρων αλγόριθμων επιβλεπόμενης μάθησης, των οποίων η πρακτική εφαρμογή είναι απαραίτητη και για την ολοκλήρωση της μελέτης μας. Τέλος, παρουσιάζουμε έναν αλγόριθμο μη επιβλεπόμενης μάθησης που θα χρησιμοποιηθεί για να ομαδοποιήσει τις εταιρείες μας σε κατηγορίες. Πρόκειται για μια τεχνική που μειώνει τη διαστατικότητα των δεδομένων μας κάνοντας με αυτόν τον τρόπο εφικτή τη δημιουργία τους σε κλάσεων στις οποίες τα κατηγοριοποιεί.

2.2.1 Λογιστική Παλινδρόμηση - Logistic Regression

Το μοντέλο λογιστικής παλινδρόμησης (logistic regression) αποτελεί ένα από τα πιο διαδεδομένα μοντέλα επιβλεπόμενης μηχανικής μάθησης. Πρόκειται ουσιαστικά για ένα μοντέλο πολυμεταβλητής στατιστικής ανάλυσης που περιγράφει τη σχέση μεταξύ μίας ή περισσότερων ανεξάρτητων μεταβλητών πρόβλεψης και μίας εξαρτημένης, μετρούμενης μεταβλητής. Η χρήση του επιτρέπει τον υπολογισμό της πιθανότητας που έχει ένα γεγονός να συμβεί, έχοντας ως κριτήριο την λογιστική συνάρτηση στην οποία αποτυπώνονται τα δεδομένα εισόδου. Η εξαρτημένη μεταβλητή Y είναι μια δυαδική μεταβλητή Bernoulli Y. Η προσέγγισή της γίνεται απαιτώντας την εξίσωση της συνάρτησης logit (log-odds) της μεταβλητής Y με ένα γραμμικό συνδυασμό των μεταβλητών εισόδου. Στο σημείο αυτό σημειώνουμε ότι οι πιθανότητες που συγκλίνουν υπέρ της εμφάνισης ενός γεγονότος εκφράζονται ως λόγος ζεύγους ακέραιων τιμών (odds) και αποτελούν ένα κλάσμα όπου ο αριθμητής προσδιορίζει την πιθανότητα που έχει το προσδοκώμενο γεγονός να συμβεί και ο παρονομαστής την πιθανότητα να μη συμβεί Peng et al. (2002). Για παράδειγμα, η πιθανότητα να ανασυρθεί ένα φύλλο σπαθί από μια τράπουλα 52 φύλλων είναι p=13/52 =0.25. Η τιμή της συνάρτησης logit της πιθανότητας αυτής είναι odds(y) = 0.25/0.75 = 1:3. Αυτό σημαίνει ότι η συμπληρωματική πιθανότητα του y, δηλαδή η πιθανότητα το χαρτί που τραβήξαμε να μην είναι σπαθί είναι 3 φορές πιο πιθανή από την πιθανότητα να είναι. Έτσι η συνάρτηση logit που αναφέραμε ουσιαστικά ισούται με το λογάριθμο του πηλίκου $\frac{P(Y)}{1-P(Y)}$. Θεωρούμε δηλαδή ότι:

$$log\frac{P(Y)}{1-P(Y)} = b_n X_n + b_{(n-1)} X_{(n-1)} \ + \ .... \ + \ b_1 X_1 \ + b_0 \tag{2.1}$$

και επιλύοντας ως προς P(Y) καταλήγουμε στην

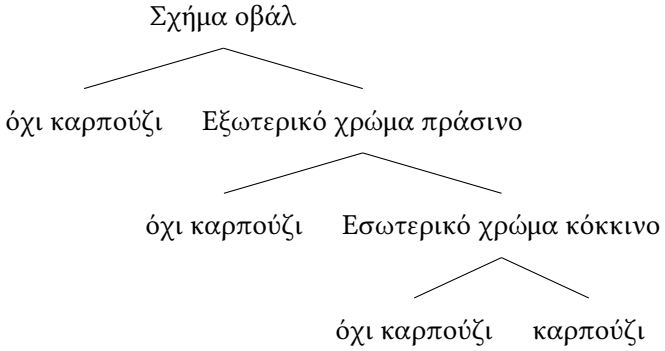
$$P(Y) = \frac{e^{b_n X_n + b_(n-1) X_(n-1) + ... + b_1 X_1 + b_0}}{1 + e^{b_n X_n + b_(n-1) X_(n-1) + ... + b_1 X_1 + b_0}} \tag{2.2}$$

Η επιλογή της χρήσης της συνάρτησης logit επιλύει προβλήματα που αντιμετωπίστηκαν κατά την απλή θεώρηση της πιθανότητας Y ως γραμμικό συνδυασμό των μεταβλητών εισόδου (γραμμική παλινδρόμηση) όπως ότι τα ακραία δείγματα συχνά δεν έδειχναν γραμμική συμπεριφορά και ότι τα σφάλματα δεν ήταν ούτε ενιαία κατανεμημένα ούτε σταθερά στο σύνολο των δεδομένων. Η σιγμοειδής μορφή της συνάρτησης 2.2 όμως που προκύπτει για τη δεσμευμένη πιθανότητα P(Y) επιλύει τέτοιου είδους προβλήματα ενώ παράλληλα διατη-

ρείται η γραμμικότητα που εισάγει ευκολία στη χρήση του μοντέλου σύμφωνα με τον τύπο 2.1. Στη μηχανική μάθηση, ο αλγόριθμος της λογιστικής παλινδρόμησης αναθέτει σε κάθε ανεξάρτητη μεταβλητή ένα συντελεστή ενδεικτικό της συνεισφοράς της στην προσέγγιση της τελικής μεταβλητής. Η ανάθεση αυτή γίνεται μέσω μιας επαναληπτικής διαδικασίας που αρχικά αναθέτει τυχαίες τιμές στους συντελεστές αυτούς και σε κάθε επανάληψη επιδιώκει τη βελτίωσή τους έχοντας ως μετρική μια συνάρτηση σφάλματος. Πλέον χρησιμοποιούμενη συνάρτηση για τη διαδικασία αυτή είναι η εκτίμηση μέγιστης πιθανοφάνειας -η μάλλον μια λογαριθμική παραλλαγή της- που υπονοεί ελαχιστοποίηση της απόκλισης Kullback–Leibler, η οποία παρουσιάστηκε παραπάνω. Η μέθοδος των ελαχίστων τετραγώνων αποτελεί επίσης μια επιλογή. Εφόσον ολοκληρωθεί η διαδικασία της μάθησης, που συνίσταται στον προσδιορισμό των συντελεστών b_i της εξίσωσης 2.2, το μοντέλο μας μπορεί να χρησιμοποιηθεί για να προβλέψει τιμές μεταβλητών Y. Στην παρούσα εργασία οι μεταβλητές Y_i , $i = 1, ..., N$ όπου N το σύνολο των εταιρειών μας, των οποίων την τιμή θέλουμε να προβλέψουμε, αντιστοιχούν σε μία από τις εταιρείες μας και είναι μεταβλητές Bernoulli δηλαδή μπορούν να πάρουν ως τιμές μόνο το 1 και το 0 ανάλογα με το αν αντιστοιχούν σε ανοιχτή ή κλειστή εταιρεία. Ο τύπος 2.2 υπολογίζει μία τιμή P(Y) που εκφράζει την πιθανότητα που έχει η μεταβλητή Y να είναι 1 και κατ'επέκταση την πιθανότητα της εταιρείας . Έτσι προβλέπεται 1 - γεγονός επιτυχίας- για την εταιρεία i αν η πιθανότητα $P(Y_i)$ είναι πάνω από 0.5 -η μια άλλη οριακή τιμή που επιλέγουμε- ή 0 -γεγονός αποτυχίας- διαφορετικά.

2.2.2 Δένδρα Αποφάσεων - Decision Trees

Πρόκειται για αλγόριθμους $\boldsymbol{h : X \longrightarrow Y}$, που προβλέπουν την τιμή (label) μιας εισόδου ακολουθώντας ένα μονοπάτι από τη ρίζα μιας δεντρικής δομής μέχρι κάποιο φύλλο. Κάθε εσωτερικός κόμβος της δομής αυτής, αναπαριστά μια συνθήκη που σχετίζεται με κάποιο από τα χαρακτηριστικά του διανύσματος εισόδου ενώ τα φύλλα αντιστοιχούν σε κάποιο label. Ας πάρουμε για παράδειγμα την απλή περίπτωση κατηγοριοποίησης ενός φρούτου στα σύνολα Καρπούζι και $\overline{\text{Καρπούζι}}$. Στην περίπτωση αυτή το σύνολο $\boldsymbol{X}$ περιέχει όλα τα φρούτα και το σύνολο $\boldsymbol{Y}$ τις τιμές {1,0} ανάλογα με το εάν ένα φρούτο είναι καρπούζι ή όχι. Δεδομένου ότι για τα φρούτα που θα ταξινομούμε γνωρίζουμε τις παρακάτω τρεις ιδιότητες {σχήμα , εξωτερικό χρώμα , εσωτερικό χρώμα} ένα δένδρο απόφασης για το πρόβλημά μας θα μπορούσε να έχει την ακόλουθη μορφή.



Ένας από τους πιο διαδεδομένους κανόνες - συνθήκες για τη κατασκευή του δέντρου είναι η χρήση κατωφλίου σε κάθε εσωτερικό κόμβο που επιλέγεται με κατάλληλο τρόπο και αφορά το σύνολο τιμών ενός εκ των χαρακτηριστικών του συνόλου δεδομένων. Κινούμαστε λοι-

πόν δεξιά αν η τιμή του συγκεκριμένου χαρακτηριστικού ενός δείγματος είναι μεγαλύτερη του κατωφλίου ή αλλιώς αριστερά. Ουσιαστικά στην περίπτωση αυτή, κάθε ταξινομητής δεντρικής δομής κατασκευάζεται με διαδοχικούς διαχωρισμούς (splits) του πεδίου τιμών της εισόδου σε υποσύνολα. Στους δυαδικούς ταξινομητές κάθε κόμβος εισάγει 2 υποσύνολα, όπως στο παράδειγμα που παρουσιάσαμε.

Γίνεται αντιληπτό ότι κατά τη διαδικασία αυτή, θα μπορούσε το ύψος του δένδρου να αυξηθεί τόσο που εν τελεί αυτό να αναπαριστά με απόλυτη πιστότητα όλα τα δείγματα του train set. Να υπάρχει δηλαδή μια διαδρομή στο δένδρο που να περιγράφει ακριβώς τις τιμές όλων των πεδίων για κάθε διάνυσμα εισόδου και τελικά να καταλήγει σε φύλλο που περιέχει και το σωστό Label. Είναι εμφανές ότι η περίπτωση αυτή είναι ενδεικτική της υπερεκπαίδευσης του μοντέλου μας και πρέπει να αποφευχθεί κατά την κατασκευή του δένδρου.

Υποθέτοντας τη δυαδική φύση του συνόλου τιμών των διανυσμάτων εισόδου δίνουμε στη συνέχεια μια σύντομη αλγοριθμική εκδοχή του σχηματισμού ενός δένδρου απόφασης σύμφωνα με το μοντέλο των CART δένδρων που προτείνουν οι Breiman et al. (1984) στο βιβλίο Classifiation and Regression Trees.

1. Ξεκίνα από τη ρίζα του δένδρου και για κάθε κόμβο που δημιουργείται επανάλαβε τα παρακάτω βήματα:
2. Για κάθε χαρακτηριστικό της εισόδου, βρες το καλύτερο κατώφλι / συνθήκη για το διαχωρισμό του συνόλου τιμών του σε δύο υποσύνολα. Στις περιπτώσεις που εξετάζεται αριθμητικό χαρακτηριστικό με σύνολο τιμών πλήθους $\boldsymbol{N}$ τότε έχουν να εξεταστούν N-1 πιθανοί διαχωρισμοί. Στις περιπτώσεις που έχουμε κατηγορηματικό χαρακτηριστικό με πλήθος διαφορετικών τιμών $\boldsymbol{R}$, οι πιθανοί συνδυασμοί υποσυνόλων που μπορούν να δημιουργηθούν είναι $2^{R-1} - 1$. Η βέλτιστη επιλογή θα κριθεί από μια συνάρτηση Κέρδους K.
3. Για τον κόμβο όπου βρίσκεσαι, βρες το χαρακτηριστικό που αν χρησιμοποιήσει το προηγούμενα υπολογισμένο κατώφλι θα ”σπάσει” το dataset μεγιστοποιώντας τη συνάρτηση Κέρδους K.
4. Έλεγξε αν ικανοποιείται η συνθήκη τερματισμού. Αν ναι η διαδικασία ολοκληρώνεται, διαφορετικά δημιούργησε 2 παιδιά για τον τρέχοντα κόμβο με βάση το χαρακτηριστικό που προσδιορίστηκε στο προηγούμενο βήμα.

Η επιλογή της συνάρτησης Κέρδους(X,i) για τον καθορισμό του κατάλληλου χαρακτηριστικού το οποίο θα περιγράψει τον εκάστοτε διαχωρισμό συνήθως γίνεται μεταξύ των συναρτήσεων που περιγράφουν α) το σφάλμα εκπαίδευσης β) το κέρδος πληροφορίας όπως πρότεινε ο Quinlan (1993) γ) την μετρική Gini κατά Breiman et al. (1984)

Σύμφωνα με τα παραπάνω αντιλαμβανόμαστε πώς τα δένδρα απόφασης κατασκευάζονται βαθμιαία χρησιμοποιώντας ευρετικές μεθόδους που λαμβάνουν υπόψη τοπικά βέλτιστες αποφάσεις. Έτσι δεν μπορούμε να είμαστε σίγουροι πως θα κατασκευάσουμε πάντα το βέλτιστο δένδρο. Παρόλα αυτά, η πρακτική εφαρμογή τους έχει αποδείξει ότι μπορούν να έχουν πολύ καλή απόδοση. Προσπάθειες για τη βελτιστοποίηση της δομής που περιγράψαμε επικεντρώ-

νονται στην επιλογή-διαμόρφωση κατάλληλης συνάρτησης κόστους και τον περιορισμό του ύψους του δένδρου που δημιουργείται.

2.2.3 Μάθηση Συνόλου - Ensemble learning

Από τα τέλη του 1990 παρατηρούνται προσπάθειες για τον συνδυασμό των αποτελεσμάτων περισσότερων ταξινομητών μηχανικής μάθησης με σκοπό την επιτυχία καλύτερου αποτελέσματος, τεχνική γνωστή ως μάθηση συνόλου. Η μάθηση συνόλου μπορεί να χωριστεί σε 2 βασικές κατηγορίες ανάλογα με τη σχέση που έχουν μεταξύ τους οι επιμέρους ταξινομητές που χρησιμοποιούνται, τη μάθηση συνόλου μη εξαρτώμενων ταξινομητών και τη μάθηση συνόλου εξαρτώμενων ταξινομητών. Οι μη εξαρτώμενοι αλγόριθμοι μάθησης συνόλου χρησιμοποιούν ταξινομητές που κατασκευάζονται ανεξάρτητα μεταξύ τους. Σε αυτή την περίπτωση το αρχικό σύνολο εκπαίδευσης που είναι διαθέσιμο χωρίζεται σε υποσύνολα, καθένα των οποίων χρησιμοποιείται για την εκπαίδευση ενός μόνο ταξινομητή. Ανάλογα με τον συγκεκριμένο αλγόριθμο, τα υποσύνολα αυτά μπορεί να είναι αμοιβαία αποκλειόμενα ή να έχουν κοινά στοιχεία. Ανάλογα με το είδος των ταξινομητών που χρησιμοποιούνται, οι έξοδοί τους συνδυάζονται με κατάλληλο τρόπο και προκύπτει η τελική έξοδος.

Αντιθέτως, στους εξαρτώμενους αλγόριθμους μάθησης συνόλου οι ταξινομητές κατασκευάζονται σειριακά και η εκπαίδευση του καθενός εξαρτάται από τα αποτελέσματα των ταξινομητών που έχουν ήδη εκπαιδευτεί.

Παρακάτω παρουσιάζουμε δύο αλγόριθμους - παραδείγματα μάθησης συνόλου που θα χρησιμοποιηθούν στη συνέχεια της εργασίας μας. Πρόκειται για αλγόριθμους που χρησιμοποιούν ως βασική δομή τα δένδρα αποφάσεων εκ των οποίων ο πρώτος ανήκει στην κατηγορία των μη εξαρτώμενων αλγόριθμων συνόλου ενώ ο δεύτερος στην κατηγορία των εξαρτώμενων.

2.2.3.1 Τυχαία Δάση - Random Forests

Στα πλαίσια αυτά, ο Ho το 1995 Ho (1995) εισήγαγε την έννοια του τυχαίου δάσους -random forest- που εκμεταλλεύεται τις προβλέψεις των τυχαίων δένδρων, τα οποία έχουν τον ρόλο των ασθενών ταξινομητών (weak learners), στοχεύοντας στην επιτυχία ακριβέστερων προβλέψεων για τις μεταβλητές εξόδου. Σύμφωνα με τον Breimen τυχαίο δάσος είναι μια συλλογή από δεντρικής δομής ταξινομητές με ανεξάρτητα σύνολα εκπαίδευσης Θκ τα οποία όμως ακολουθούν την ίδια κατανομή και μπορούν να συμβολιστούν ως $h(x,)$ οπού x είναι το σύνολο δειγμάτων εισόδου. Κάθε ταξινομητής κ αποφασίζει πια είναι η πιο πιθανή κλάση για την ταξινόμηση της δεδομένης εισόδου και ο συνδυασμός των αποφάσεων αυτών καθορίζει και το τελικό αποτέλεσμα.

Σε αυτή τη διαδικασία συνδυασμού των δένδρων αποφάσεων παράμετρο για την επιτυχία των προβλέψεων αποτελεί η κατάλληλη επιλογή των . Προφανώς η εκπαίδευση δένδρων στα ίδια δεδομένα θα είχε ως αποτέλεσμα και την παραγωγή ίδιας εξόδου οπότε πρέπει να εισαχθεί τυχαιότητα στην επιλογή των διανυσμάτων εισόδου του εκάστοτε δένδρου. Η τυχαιότητα αυτή αποτυπώνεται στην αλγοριθμική διαδικασία με τους εξής τρόπους: τυχαία

επιλογή χαρακτηριστικών για κάθε δένδρο αποφάσεων καθώς και τυχαία επιλογή υποσυνόλου του συνόλου εκπαίδευσης.

Αναλυτικότερα, για την επιλογή των διανυσμάτων εκπαίδευσης χρησιμοποιείται η τεχνική bagging κατά την οποία για την εκπαίδευση κάθε επιμέρους ταξινομητή επιλέγεται τυχαία ένα υποσύνολο του train set που αποτελείται από μέρος των αρχικών δειγμάτων. Οι υπολειπόμενες θέσεις, που αντιστοιχούν στα δείγματα που δεν χρησιμοποιήθηκαν, περιέχουν διπλότυπα των επιλεγέντων δειγμάτων. Αν δηλαδή το train set περιέχει N διανύσματα, κάθε δένδρο θα χρησιμοποιήσει N διανύσματα για την εκπαίδευσή του τα οποία όμως θα περιέχουν M < N αυτούσια δείγματα από το train set και άλλα N-M δείγματα που είναι αντίγραφα αυτών που έχουν επιλεγεί. Επιπλέον, κάθε δένδρο αποφάσεως χρησιμοποιεί για να κατασκευαστεί διαφορετικό συνδυασμό χαρακτηριστικών, ο οποίος επίσης αποτελεί υποσύνολο των συνολικών χαρακτηριστικών εισόδου. Το πλήθος των χαρακτηριστικών που χρησιμοποιούνται για την κατασκευή κάθε δένδρου παραμένει σταθερό κατά τη διαδικασία δημιουργίας του δάσους.

Η διαδικασία που ακολουθείται για την πρόβλεψη με χρήση του τυχαίου δάσους όπως περιγράφεται από τον Breiman αναφέρει πως για την κατηγοριοποίηση ενός αντικειμένου εισάγουμε το διάνυσμα εισόδου αυτό ως είσοδο σε όλα τα δέντρα. Καθένα από τα δέντρα αυτά μεγαλώνει απεριόριστα (χωρίς περιορισμό στις διακλαδώσεις) και δίνει μία «ψήφο» ως προς το αν ανήκει σε μια κατηγορία ή όχι. Στη συνέχεια, τα αποτελέσματα ομαδοποιούνται αθροίζοντας τους εν λόγω ψήφους (ή βάσει κάποιας άλλης συνάρτησης για την ομαδοποίηση) και εξάγεται το τελικό αποτέλεσμα.

Το σφάλμα γενίκευσης που προκύπτει στον τελικό ταξινομητή επηρεάζεται από τη συσχέτιση μεταξύ των δένδρων και την ισχύ τους (strength). Με τον τελευταίο όρο περιγράφουμε τη διαφορά από την πιθανότητα ενός δένδρου να προβλέψει το σωστό αποτέλεσμα έναντι οποιουδήποτε άλλου, δεδομένης της εισόδου. Μείωση της συσχέτισης και αύξηση της ισχύος των δένδρων οδηγούν και σε μικρότερα σφάλματα γενίκευσης. Η επιλογή του αριθμού χαρακτηριστικών που θα ληφθούν υπόψη για τη κατασκευή κάθε δένδρου επηρεάζει και τους δύο προηγούμενους παράγοντες, μάλιστα αυξάνοντας και τους δύο όσο και ο ίδιος αυξάνεται. Τέλος ο Breiman, απόδειξε ότι η αύξηση του αριθμού των δένδρων δεν οδηγεί σε καταστάσεις υπερεκπαίδευσης - overfitting - αλλά εισάγει τελικά μια οριακή τιμή του σφάλματος γενίκευσης, αφού όπως προκύπτει από το νόμο των μεγάλων αριθμών η πιθανότητα του να υπάρχει απόφαση πλειοψηφίας υπέρ της μίας κλάσης έναντι οποιασδήποτε άλλης συγκλίνει.

Η ευρεία χρήση των τυχαίων δασών οφείλεται στο γεγονός ότι είναι εύκολα και γρήγορα να χρησιμοποιηθούν για μεγάλα datasets, τα αποτελέσματά τους είναι αποδεκτά και σε περιπτώσεις καλύτερα από όλες μεθόδους μάθησης συνόλου, διαχειρίζονται αποτελεσματικά τις απουσιάζουσες τιμές χαρακτηριστικών και μπορούν να δώσουν μια εκτίμηση για τη σπουδαιότητα των χαρακτηριστικών εκπαίδευσης.

2.2.3.2 Δέντρα λήψης αποφάσεων ενίσχυσης κλίσης

Στην δεύτερη κατηγορία των αλγόριθμων μάθησης συνόλου ανήκουν οι ταξινομητές που χρησιμοποιούν την τεχνική της ενδυνάμωσης-ενίσχυσης (boosting). Ένας από αυτούς υλοποιεί ενίσχυση κλίσης σε δένδρα αποφάσεων, διαδικασία που θα περιγραφεί στη συνέχεια. Στην παρούσα διπλωματική θα χρησιμοποιήσουμε τη βιβλιοθήκη Catboost που υλοποιεί την παραπάνω διαδικασία. Ο Michael Kearns περιέγραψε την τεχνική του boosting ως έναν αποτελεσματικό αλγόριθμο για τη μετατροπή σχετικά ασθενών μοντέλων μάθησης σε ένα ισχυρό τελικό μοντέλο Kearns (1988) . Η βασική ιδέα είναι να χρησιμοποιούνται «ασθενείς μέθοδοι μάθησης» πολλές φορές, με σκοπό να λάβουμε μία σειρά από διαδοχικές υποθέσεις -συναρτήσεις που μοντελοποιούν το πρόβλημά μας-, κάθε μία από τις οποίες θα επικεντρώνεται στη σωστή διαχείριση των στοιχείων που προηγουμένως είχαν ταξινομηθεί-προβλεφθεί-υπολογιστεί με μεγάλο σφάλμα. Αν και πρόκειται βέβαια για μια απλή ιδέα δεν είναι εμφανής ο τρόπος με τον οποίο μπορούμε να καταφέρουμε κάτι τέτοιο

Η τελική μορφή που παίρνει το μοντέλο σύμφωνα με τα παραπάνω περιγράφεται από τους τύπους:

$$F_{t+1}(x) \leftarrow F_t(x) + \alpha_t f_t(x) \tag{2.3}$$

$$F(x) = \sum_{1}^{T} \alpha_i f_i(x) \tag{2.4}$$

Στη βιβλιογραφία ο Friedman (2000) βασιζόμενος στη διαδικασία προσέγγισης συναρτήσεων μέσω υπολογισμού της κλίσης του σφάλματος μεταξύ των εκτιμόμενων τιμών και των πραγματικών και τη χρήση του για την επαναληπτική ενημέρωση της προσεγγιστικής συνάρτησης πρότεινε ένα παρόμοιο μοντέλο για τον προσδιορισμό συναρτήσεων που επιδιώκουν να περιγράψουν τα μοντέλα μηχανικής μάθησης. Πρόκειται για μια διαδικασία που χρησιμοποιεί την κλίση g_m μιας συνάρτησης σφάλματος της επιλογής μας μεταξύ των εκτιμόμενων και των πραγματικών τιμών του δείγματός μας, ώστε να διαμορφώσει, σύμφωνα με την 2.5, τη συνάρτηση πρόβλεψης του επόμενου βήματος. Όπως εξηγεί και στο σχετικό paper αφού υπολογιστεί η κλίση

$$-g_t(x) = \left[\frac{\partial L(y_i, F(x_i)}{\partial F(x_i)}\right]_{F(x)=F_{t-1}(x)} \tag{2.5}$$

της συνάρτησης L που υπολογίζει το σφάλμα μεταξύ των εκτιμόμενων και των αναμενόμενων τιμών της F, επιλέγουμε τη συνάρτηση $f_t(x)$ η οποία αντιπροσωπεύει έναν αδύναμο ταξινομητή, έτσι ώστε να έχει κλίση όσο το δυνατόν πιο κοντά στην κλίση $-g_t(x)$που υπολογίσαμε. Οι συντελεστές a_i, όπου $i = 1, ...T$ αντιστοιχούν στις φορές τις οποίες επαναλαμβάνουμε τη διαδικασία επαναπροσδιορισμού της συνάρτησης F, μπορούν σε απλές περιπτώσεις να επιλεχθούν ως σταθερές είτε να προσδιοριστούν μέσα από μια line search διαδικασία. Η τεχνική λοιπόν των GDDT συνίσταται στη δημιουργία μιας σειράς δένδρων $f_t(x)$, όπου τα

αποτελέσματα που προκύπτουν από την εκπαίδευση του ενός επηρεάζουν την εκπαίδευση του επόμενου σύμφωνα με μία προσαρμογή του τύπου 2.4 .

2.2.4 Non-negative matrix factorization

Πρόκειται για μια τεχνική πολυπαραμετρικής ανάλυσης με βάσεις στην γραμμική άλγεβρα, της οποίας η ικανότητα να παραγοντοποιεί και να συσταδοποιεί δεδομένα υψηλής διαστατικότητας μελετήθηκε από τα μέσα της δεκαετίας του 1990 από τους Paatero and Tapper (1994). Οι ίδιοι παρουσίασαν τα σημεία υπεροχής της συγκεκριμένης μεθόδου (επονομαζόμενη τότε ως Positive matrix factorization) έναντι της ευρέως διαδεδομένης Ανάλυσης ανεξάρτητων συνιστωσών τεχνικής τεκμηριώνοντας τη μελέτη τους με πρακτικές εφαρμογές. Το 1999 οι Lee and Seung (1999) σε δημοσίευσή τους παρουσίασαν σημαντικά αποτελέσματα χρήσης της μεθόδου αυτής, που αναφερόταν πλέον ως Non-negative matrix factoriztion, NMF, σε προβλήματα ανάλυσης εικόνας και κειμένου και έκτοτε οι αναφορές στην τεχνική αυτή καθώς και τα πεδία εφαρμογών της αυξάνονται συνεχώς.

Η βασική λογική λειτουργίας του μοντέλου στηρίζεται στην προσέγγιση ενός μη αρνητικού πίνακα ως το γινόμενο δύο εξίσου μη αρνητικών πινάκων. Δεδομένου δηλαδή ενός μη αρνητικού πίνακα V η τεχνική συνίσταται στην εύρεση δύο μη αρνητικών πινάκων H,W τ.ω. V≈WH.

Ας σκεφτούμε ότι έχουμε για παράδειγμα ένα σύνολο από m n-διάστατα διανύσματα δεδομένων τα οποία χρησιμοποιούμε για τη σύσταση του πίνακα V_{nxm} .

Σύμφωνα με την παραπάνω υπόθεση κάθε διάνυσμα μπορεί να προσεγγιστεί από το γινόμενο Wh όπουh είναι η αντίστοιχη στήλη του πίνακα H. Η παραπάνω εξίσωση περιγράφει το διάνυσμα v ως γραμμικό συνδυασμό ενός πίνακα βάσης επί μια στήλη βαρών. Η απόκλιση της παραπάνω προσέγγισης σχετίζεται με την ύπαρξη ή απουσία κάποιας αλληλοσυσχέτισης μεταξύ των δεδομένων εισόδου και στην ικανότητα του αλγορίθμου να την εντοπίσει και αποκωδικοποιήσει.

Ο αλγόριθμος συνίσταται από μια επαναληπτική διαδικασία όπου δεδομένων κάποιων αρχικών πινάκων V_0,W_0,H_0 σε κάθε επανάληψη i ενημερώνεται ο πίνακας V_i ως γινόμενο των W_i,H_i , υπολογίζεται το σφάλμα σε σχέση με τον αρχικό πίνακα V και προσαρμόζονται αναλόγως οι τιμές των W_{i+1},H_{i+1}.

Η συνήθης συνάρτηση κόστους για τον υπολογισμό της απόκλισης του V από το W_iH_i είναι το τετραγωνικό σφάλμα με τη χρήση του οποίου το πρόβλημα μεταφράζεται ως εξής : Δεδομένου ενός πίνακα V να βρεθούν 2 μη αρνητικοί πίνακες , W τ.ω. να ελαχιστοποιούν τη συνάρτηση

$$C(H,W) = \|V - WH\|^2 \tag{2.6}$$

Άλλη ευρέως χρησιμοποιούμενη συνάρτηση σφάλματος αποτελεί η απόκλιση Kullback–Leibler για θετικούς πίνακες που ορίζεται ως

$$D_{KL}(A||B) = \sum_{ij}^{\infty} A_{ij} log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \tag{2.7}$$

και εφόσον επιλεχθεί μετατρέπει το πρόβλημά μας στο: Δεδομένου ενός πίνακα V να βρεθούν 2 μη αρνητικοί πίνακες H, W τ.ω. να ελαχιστοποιούν τη συνάρτηση

$$C(H,W) = \sum_{ij}^{\infty} V_{ij} log \frac{V_{ij}}{WH_{ij}} - V + WH_{ij} \tag{2.8}$$

Τα κοινά στοιχεία των συναρτήσεων που τις καθιστούν και αποδεκτές για τον NMF είναι ότι είναι φραγμένες από το μηδέν, τιμή την οποία παίρνουν μόνο όταν V =WH. Όσον αφορά τον τρόπο που γίνεται η ενημέρωση των τιμών των W,H επίσης έχουν εξεταστεί διαφορετικοί αλγόριθμοι. Μια από τις πιο κοινές υλοποιήσεις, την οποία χρησιμοποιούμε και στη μεθοδολογία της παρούσας διπλωματικής εργασίας„ περιγράφει μια πολλαπλασιαστική διαδικασία όπου ισχύει $H_{i+1} \leftarrow H_i\eta$ και $W_{i+1} \leftarrow W_i\lambda$ όπου τα η,λ καθορίζονται με βάση τη συνάρτηση κόστους. Θεωρώντας λοιπόν συνάρτηση κόστους τετραγωνικού σφάλματος και πολλαπλασιαστική μέθοδο ενημέρωσης καταλήγουμε στα:

$$H_{i+1} \leftarrow H_i \frac{W_i^T V}{W_i^T W_i H_i} \tag{2.9}$$

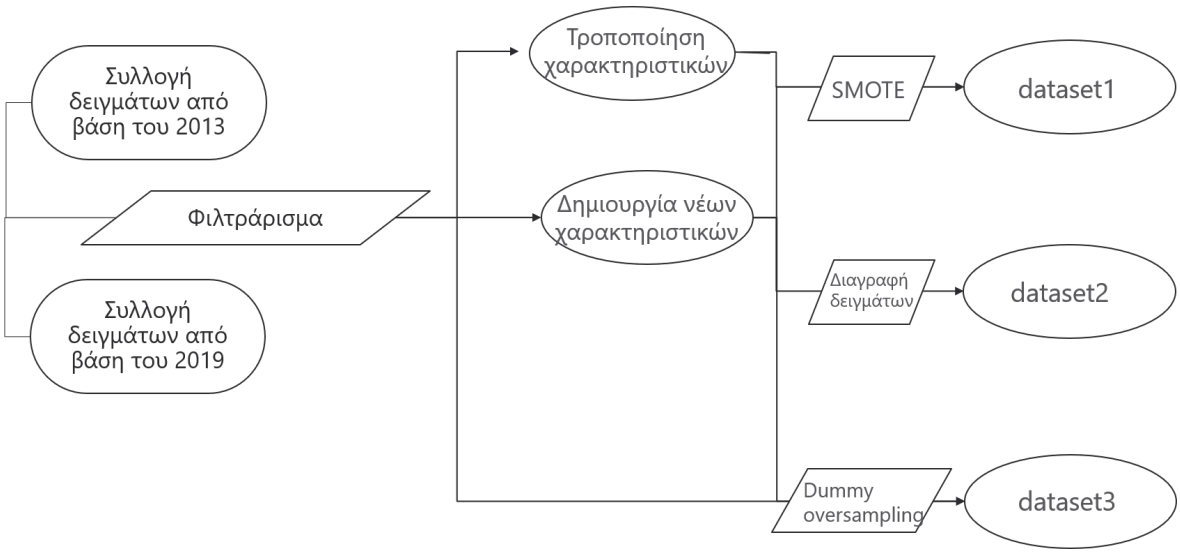
$$W_{i+1} \leftarrow W_i \frac{W_i H_i^T}{W_i H_i H_i^T} \tag{2.10}$$

Η μέθοδος που περιγράφηκε καθιστά δυνατή την εξαγωγή θεματικών ενοτήτων από μια ομάδα κειμένων και την κατηγοριοποίηση των τελευταίων σε αυτά. Εισάγοντας ως στήλες του πίνακα V τα διαφορετικά κείμενα και ακολουθώντας την παραπάνω διαδικασία προκύπτουν πίνακες W,H που περιγράφουν ο πρώτος τις κατηγορίες που προκύπτουν από τα παραπάνω κείμενα και ο δεύτερος το πόσο μεγάλη είναι συσχέτιση του κάθε κειμένου με την εκάστοτε κατηγορία. Προκύπτει έτσι μια κατηγοριοποίηση των δοσμένων κειμένων συγκρίνοντας τις τιμές των στηλών του H και συμπεραίνοντας ποιο θέμα αναπαριστά πιο πιστά το εκάστοτε κείμενο.

Κεφάλαιο 3

Μεθοδολογία

Αυτή η ενότητα επικεντρώνεται στην διαδικασία που ακολουθήσαμε για να συγκεντρώσουμε και να επεξεργαστούμε το σύνολο δεδομένων μας. Στην εικόνα περιγράφεται σχηματικά όλη η διαδικασία που θα αναλύσουμε στη συνέχεια.



Σχήμα 3.1: Σχηματική περιγραφή δημιουργίας συνόλων δεδομένων

3.1 Συλλογή Δεδομένων

Για τη συλλογή δεδομένων χρησιμοποιήθηκε ως επί των πλείστων η διεπαφή προγραμματισμού εφαρμογών (API) του ιστότοπου crunchbase.com. Πρόκειται για μια πλατφόρμα με επιχειρησιακές πληροφορίες, για ιδιωτικές και δημόσιες εταιρείες, τις οποίες προσθέτουν και ενημερώνουν οι χρήστες. Για την παρούσα μελέτη χρησιμοποιήσαμε ένα στιγμιότυπο της βάσης από τις 10/11/2013. Παράλληλα, όπου κρινόταν απαραίτητο, έγινε διασταύρωση πληροφοριών με επιπλέον ιστοσελίδες όπως οι angel.co και businessinsider.com. Τέλος, όπως θα αναφέρουμε και στη συνέχεια, χρησιμοποιήσαμε επίσης μερικές συμπληρωματικές εταιρείες που ανήκουν στη βάση δεδομένων του crunchbase.com και προστέθηκαν μετέπειτα του 2013.

3.2 Επεξεργασία Δεδομένων

3.2.1 Επιλογή Δεδομένων

Η βάση δεδομένων που χρησιμοποιήθηκε αρχικά, αποτελείτο από τους παρακάτω πίνακες: cb_funding_rounds, cb_degrees, cb_objects, cb_relationships ,

Η των εταιρειών βασίστηκε σε εγγραφές του πίνακα cb_objects και ακολούθησε την διαδικασία που περιγράφεται στη συνέχεια.

- Φιλτράρισμα αντικειμένων του πίνακα επιλέγοντας αυτά που είχαν χαρακτηρισμό «Company» στο πεδίο entity_type
- Εκ νέου φιλτράρισμα αντικειμένων με βάση το πεδίο 'categories' του ίδιου πίνακα ώστε να επιλεχούν μόνο σχετικές με τον τομέα της υγείας εταιρείες.Οι λέξεις κλειδιά που χρησιμοποιήθηκαν για την επιλογή αυτή είναι οι «healthcare» , «biotechnology» , «medical device».
- Διαγραφή των εταιρειών με ημερομηνία δημιουργίας πριν από το 2000. Με τον τρόπο αυτό επιδιώξαμε να εξασφαλίσουμε την εγκυρότητα της χρήσης του χαρακτηρισμού startup για τα αντικείμενα της μελέτης μας. Επιπλέον κρίναμε πως οι εταιρείες αυτές θα ανταποκρίνονταν πιο πιστά στο μοντέλο της εταιρείας που προσπαθήσαμε να μελετήσουμε, εκείνης δηλαδή της οποίας η λειτουργία σχετίζεται με κάποια επιστημονική-τεχνολογική καινοτομία. Τελικά, λόγω μεγάλης ανισορροπίας στο dataset αποφασίσαμε να κρατήσουμε και 5 εταιρείες με ημερομηνία δημιουργίας : 1996 < ημερομηνία <2000.
- Διαγραφή εταιρειών που δεν είχαν πληροφορία σε κάποιο από τα πεδία: Όνομα , ημερομηνία δημιουργίας και κατηγορία. Με την τακτική αυτή έγινε προσπάθεια να αποκλειστούν εταιρείες που παρέπεμπαν σε εταιρείες-φαντάσματα. Οι προαναφερθέντες ιστότοποι επιτρέπουν στους χρήστες να αποθηκεύουν και να επεξεργάζονται δεδομένα χωρίς την ύπαρξη συστήματος αξιολόγησης της εγκυρότητάς τους και για το λόγο αυτό οδηγηθήκαμε στην παραπάνω επιλογή.
- Για να εξασφαλιστεί όσο το δυνατόν περισσότερο ότι οι εταιρείες του δείγματός μας έχουν υπάρξει απαιτήθηκε επίσης να διαθέτουν ελάχιστη περιγραφή για το αντικείμενο ενασχόλησής τους. Στη συγκεκριμένη περίπτωση απαιτήθηκε τουλάχιστον μία από τις στήλες tags, description , short_description της βάσης να μην είναι κενή.

Αφού σχηματίστηκε αυτό το πρώτο σύνολο εταιρειών παρατηρήσαμε ότι η αναλογία εταιρειών που θα κατηγοριοποιούσαμε στις αποτυχημένες σε σχέση με αυτές οι οποίες θεωρήθηκαν πετυχημένες δεν ήταν ικανοποιητική και σε συνδυασμό με το μικρό μέγεθος του dataset θεωρήσαμε ότι δε θα είχαμε ικανοποιητικά αποτελέσματα. Αποφασίσαμε για το λόγο αυτό να χρησιμοποιήσουμε και ένα δεύτερο στιγμιότυπο της βάσης του crunchbase το οποίο αντιστοιχούσε στα δεδομένα της βάσης στις 20/12/2019. Για να συγκεντρώσουμε καινούρια δεδομένα χρησιμοποιήσαμε τα παρακάτω φίλτρα:

- 2000 ≤ Ημερομηνία ίδρυσης ≤ 2013
- Status: closed
- Η λίστα κατηγοριών της εταιρείας να περιλαμβάνει κάποιον από τους όρους biotechnology, medical, therapeutics.

- Η εταιρεία να μην έχει συμπεριληφθεί στο προηγούμενο dataset
- Ένα από τα πεδία description, short_description να είναι μη κενό.

Επίσης, καθώς συγκεντρώθηκαν οικονομικά χαρακτηριστικά κατεγράφησαν στο τελικό dataset μόνο πληροφορίες με χρονοσφραγίδα μέχρι το 2013. Όπως αναφέραμε και στην εισαγωγή στόχος μας είναι από ένα δείγμα - εταιρεία του 2013 να αποφασίσουμε για την κατάστασή του στο 2019. Σημειώνεται ότι η μορφή της βάσης ήταν πλέον διαφορετική εξ'ού και η τροποποίηση των φίλτρων επιλογής μας.

3.2.1.1 Επιλογή χαρακτηριστικών από τη βάση

Αφού σχηματίστηκε το σύνολο των εταιρειών του αποτελούν το δείγμα μας έγινε η επιλογή των χαρακτηριστικών που θα κρατήσουμε από τους πίνακες της βάσης. Τα πεδία που επιλέχθηκαν αρχικά διαφαίνονται στον ακόλουθο πίνακα 3.1 .

Πίνακας Βάσης	Όνομα Πεδίου	Περιγραφή	Τύπος
cb_objects.sql	uuid	Αναγνωριστικό για αναφορά στην εταιρεία	varchar(32)
cb_objects.sql	Status	Πληροφορίες για την κατάσταση Λειτουργίας	varchar(32)
cb_objects.sql	Founded_at	Ημερομηνία ίδρυσης	date
cb_objects.sql	Closed_at	Ημερομηνία παύσης λειτουργίας	date
cb_objects.sql	Country_code	Χώρα Ίδρυσης	varchar(32)
cb_objects.sql	Twitter_username	Όνομα χρήστη twitter της εταιρείας, εφόσον υπάρχει	varchar(64)
cb_objects.sql	Short_description	Σύντομη περιγραφή	varchar(max)
cb_objects.sql	Description	Αναλυτική περιγραφή	varchar(max)
cb_objects.sql	first_funding_at	Ημερομηνία πρώτης χρηματοδότησης	date
cb_objects.sql	Funding_rounds	Γύροι χρηματοδότησης	int
cb_objects.sql	Total_funding	Συνολική χρηματοδότηση σε USD	int
cb_objects.sql	Tags	Ετικέτες, λέξεις-κλειδία	varchar(64)
cb_funding_rounds.sql	Description	Τύπος γύρου χρηματοδότησης	varchar(32)
cb_relationships.sql	person_object_id	Αναγνωριστικό ατόμων που συμμετέχουν σε μία σχέση	varchar(64)
cb_relashionships.sql	relationship_object_id	Αναγνωριστικό εταιρειών που συμμετέχουν σε μία σχέση	varchar(64)
cb_relationships.sql	title	Περιγραφή θέσης εργαζομένου στην εταιρεία	varchar(64)
cb_degrees.sql	object_id	Αναγνωριστικό ατόμου που συνδέεται με κάποιο πτυχίο	varchar(64)
cb_degrees.sql	subject	Αντικείμενο σπουδών	varchar(64)

Πίνακας 3.1: Πίνακας πεδίων που συλλέχθηκαν από τη βάση

Η μορφή της βάσης του ιστότοπου crunchbase την δεύτερη φορά που αναζητήσαμε επιπλέον δεδομένα είχε τροποποιηθεί. Στη συνέχεια παρουσιάζουμε την αντιστοίχηση που έχουν οι πίνακες και τα χαρακτηριστικά που αποθηκεύτηκαν της πρώτης με τη δεύτερη εκδοχή της βάσης.3.2.

Τέλος, τα νέα δεδομένα στα οποία αποκτήσαμε πρόσβαση, περιείχαν πολύ περισσότερες πληροφορίες για την κάθε οντότητα της βάσης. Από αυτές επιλέξαμε να συγκεντρώσουμε επιπλέον τον συνολικό αριθμό των επενδυτών που συνδέονται με την κάθε startup ως ένα χαρακτηριστικό τύπου int και ονόματος number_of_investors καθώς και την ύπαρξη πληροφορίας για το site της εταιρείας που αποθηκεύσαμε στο πεδίο has_site τύπου int.

Οι συνολικές εταιρείες που συλλέχθηκαν ήταν 1201 από τις οποίες +++

3.2.2 Επεξεργασία χαρακτηριστικών ή Δημιουργία νέων χαρακτηριστικών

3.2.2.1 Τροποποιημένα Χαρακτηριστικά

Τα παραπάνω πεδία της βάσης είτε εισήχθησαν αυτούσια στα μοντέλα μας, είτε τροποποιήθηκαν προτού χρησιμοποιηθούν, είτε χρησιμοποιήθηκαν για τη δημιουργία νέων μεταβλητών.

Πίνακας Βάσης 2013	Όνομα Πεδίου	Πίνακας Βάσης 2019	Όνομα Πεδίου
cb_objects.sql	id	organizations	uuid
cb_objects.sql	Founded_at	organizations	founded_at
cb_objects.sql	Closed_at	organizations	closed_on
cb_objects.sql	Country_code	organizations	country_code
cb_objects.sql	Twitter_username	organizations	twitter_url
cb_objects.sql	Short_description	organizations	short_description
cb_objects.sql	Description	organization_descriptions	description
cb_objects.sql	first_funding_at	funds	announced_on
cb_objects.sql	Funding_rounds	funding_rounds	announced_on uuid
cb_objects.sql	Total_funding	funding_rounds	announced_on rasied_ammount_usd
cb_objects.sql	Tags	-	-
cb_funding_rounds.sql	Description	funding_rounds	type
cb_people.sql	Company	people	featured_job_organization_uuid
cb_people.sql	Degree	degrees	name

Πίνακας 3.2: Πίνακας αντιστοίχησης

Αρχικά, σχετικά με την τροποποίηση των ήδη υπαρχόντων χαρακτηριστικών, επιχειρήσαμε τον περιορισμό των αριθμητικών πεδίων και τη χρήση κατηγορηματικών μεταβλητών με δυαδικό σύνολο τιμών για να εξετάσουμε ποια ανταποκρίνονται καλύτερα στη διαδικασία μάθησης των αλγόριθμων. Συνεπώς, για τα επιλεγμένα κατηγορηματικά χαρακτηριστικά, δημιουργήσαμε νέα δυαδικά πεδία όπου η τιμή 1 είναι ενδεικτική της παρουσίας ενός συγκεκριμένου προσδιοριστικού και 0 της απουσίας του. Αντίστοιχα, για τα αριθμητικά χαρακτηριστικά, δημιουργήσαμε συστάδες τμηματοποιώντας το σύνολο τιμών τους, στις οποίες και τα εντάξαμε. Στον παρακάτω πίνακα 3.3 παρουσιάζεται η διαδικασία δημιουργίας των καινούριων αυτών μεταβλητών καθώς και τα χαρακτηριστικά της βάσης από τα οποία προέκυψαν.

Όνομα παλιού χαρακτηριστικού	Όνομα νέου χαρακτηριστικού	Περιγραφή	Τύπος
Total_funding	funds>30mil	funding$\in [30.000.000, inf)$	binary
Total_funding	funds>10mil	funding$\in [10.000.000, 30.000.000)$	binary
Total_funding	funds>5mil	funding$\in [5.000.000, 10.000.000)$	binary
Total_funding	funds>5mil	funding$\in [0, 5.000.000)$	binary
Total_funding	has_funding	Ύπαρξη πληροφοριών χρηματοδότησης	binary
Funding_rounds	funding_round>4.5	funding$_rounds \in [5, inf)$	binary
Funding_rounds	funding_round>2.5	funding$_rounds \in [3, 4]$	binary
Funding_rounds	funding_round>0	funding$_rounds \in [1, 2]$	binary
Funding_rounds	has_funding_rounds	Ύπαρξη πληροφοριών γύρων χρηματοδότησης.	binary
Founded_at	Years1	Χρονικό διάστημα λειτουργίας$\in [0, 3]$	binary
Founded_at	Years2	Χρονικό διάστημα λειτουργίας$\in [4, 8]$	binary
Founded_at	Years3	Χρονικό διάστημα λειτουργίας$\in [9, inf)$	binary

Πίνακας 3.3: Πίνακας τροποποιημένων χαρακτηριστικών

Η επιλογή των διαστημάτων αρχικά έγινε με στόχο τη δημιουργία ισάριθμων συνόλων. Ωστόσο, κατά την εκπαίδευση των αλγορίθμων, μελετώντας τα αποτελέσματα που λαμβάναμε καταλήξαμε στην τμηματοποίηση που παρουσιάζεται, η οποία δε διαφέρει πολύ από την αρχική μας προσέγγιση.

3.2.2.2 Νέα χαρακτηριστικά

Επιδιώξαμε να συγκεντρώσουμε σε ένα χαρακτηριστικό πληροφορία ενδεικτική της ωριμότητας του επιστημονικού πεδίου απασχόλησης της κάθε επιχείρησης. Θελήσαμε να αξιολογήσουμε με αυτόν τον τρόπο την πιθανότητα που θα είχε η εταιρεία να δημιουργήσει το προϊόν ή να φέρει εις πέρας την υπηρεσία που είχε ως στόχο κατά την ίδρυσή της. Η προσπάθεια αυτή συνάδει με την ευρύτερη κατεύθυνση που θελήσαμε να δώσουμε στην εργασία μας ως προς την βαρύτητα που θα έχει η ύπαρξη γνώσης σε σχέση με άλλα κριτήρια για να κατατάξουμε τις εταιρείες. Για τον λόγο αυτό επιχειρήσαμε να βρούμε πόσα επιστημονικά άρθρα, που αφορούν το αντικείμενο ενασχόλησης της εκάστοτε startup, έχουν δημοσιευτεί. Έτσι, για κάθε μία από αυτές, χρησιμοποιήσαμε το πεδίο περιγραφής που υπήρχε στη βάση, κάναμε μικρές τροποποιήσεις όπως η αφαίρεση αναφορών και ονομάτων, και στη συνέχεια υλοποιήσαμε μία αναζήτηση δημοσιεύσεων στο google scholar. Στα χαρακτηριστικά της αναζήτησης, θέσαμε ως κείμενο το παραπάνω πεδίο περιγραφής και ως χρονικό πλαίσιο το διάστημα ενός χρόνου που έχει ως αρχή τη χρονιά δημιουργίας της startup. Το νέο αυτό χαρακτηριστικό το ονομάσαμε publications. Κατά την εκπαίδευση του αλγορίθμου το διασπάσαμε σε δύο δυαδικά χαρακτηριστικά για την καλύτερη αξιοποίησή του από το μοντέλο μας.

Επιπλέον, τα πεδία με τα αναγνωριστικά χρησιμοποιήθηκαν για να συντάξουμε ερωτήματα και συνδυάζοντας τους παραπάνω πίνακες να συγκεντρώσουμε:

a) Τον αριθμό των ιδρυτών των εταιρειών

b) Τον αριθμό των μελών του διοικητικού συμβουλίου των εταιρειών

c) Τον συνολικό αριθμό των πτυχίων των ιδρυτών που αφορούν επιστήμες στον κύκλο της ιατρικής

d) Τον συνολικό αριθμό των πτυχίων των ιδρυτών που αφορούν επιστήμες σχετικές με την τεχνολογία

e) Αν υπήρξε γύρος χρηματοδότησης A

f) Αν υπήρξε γύρος χρηματοδότησης B

g) Αν υπήρξε γύρος χρηματοδότησης C

h) Αν υπήρξαν κεφάλαια επιχειρηματικών συμμετοχών (VC)

i) Το χρονικό διάστημα λειτουργίας της εταιρείας εως το 2013

j) Το χρονικό διάστημα μεταξύ της ίδρυσης της εταιρείας και του πρώτου γύρου χρηματοδότησής της.

Όπως και παραπάνω, για τα δύο τελευταία αριθμητικά πεδία δημιουργήσαμε νέα χαρακτηριστικά που προέκυψαν από την τμηματοποίηση του συνόλου τιμών τους.

3.2.2.3 Χρήση της μεθόδου NMF για την εξαγωγή των κατηγοριών απασχόλησης των εταιρειών

Για να χωρίσουμε τις εταιρείες σε πιο εξειδικευμένες κατηγορίες χρησιμοποιήσαμε μια διαδεδομένη τεχνική topic extraction γνωστή ως Non negative matrix factorization η οποία έχει περιγραφεί προηγουμένως. Η επιλογή μας βασίστηκε στην ευκολία υλοποίησής της με την προγραμ-

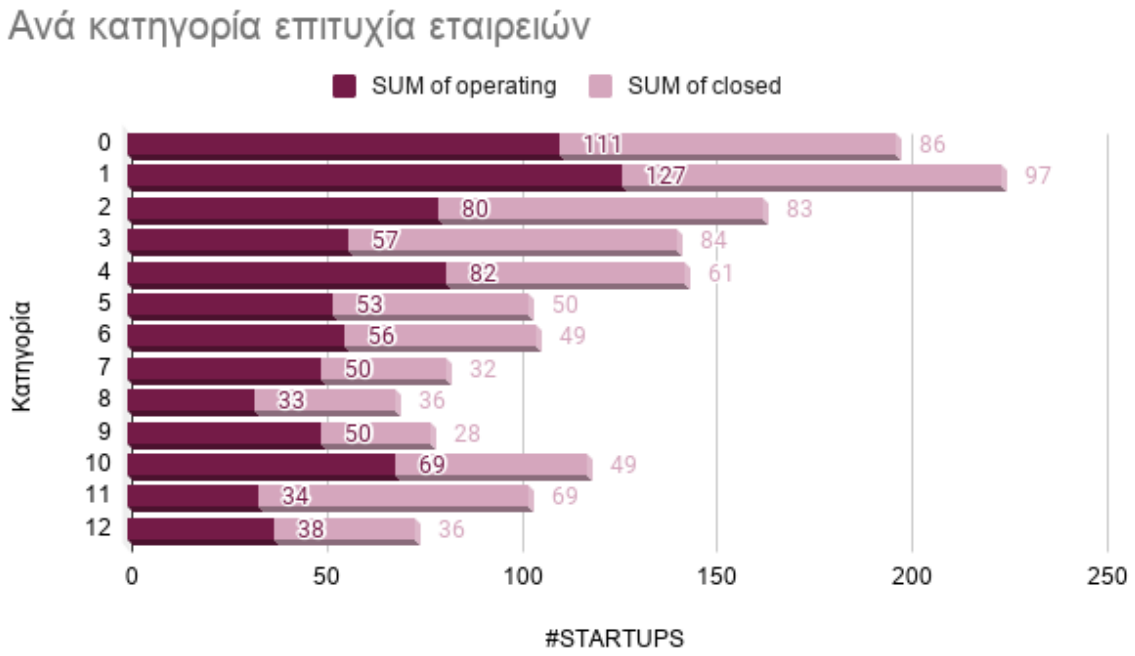
ματιστική γλώσσα python καθώς και στις μελέτες που την παρουσιάζουν ως πιο αποτελεσματική μεταξύ άλλων, στην εξαγωγή θεμάτων από μικρά κείμενα.

Για τον αλγόριθμό μας, χρησιμοποιήσαμε τα κείμενα που περιέχονταν στα πεδία description – short-description και tags. Συγκεκριμένα, σχηματίσαμε μια περιγραφή από το συνδυασμό των short description – description που θεωρήθηκε ότι αντιπροσώπευε καλύτερα το αντικείμενο ενασχόλησης της κάθε startup. Στα κείμενα αυτά που σχηματίστηκαν προστέθηκαν, όπου αυτά ήταν διαθέσιμα, τα tags του κάθε δείγματος. Για την εξαγωγή των κατηγοριών και των αντιπροσωπευτικών λέξεων κάθε κατηγορίας χρησιμοποιήσαμε μόνο το σύνολο των εταιρειών που στη συνέχεια θα αποτελούσε το σύνολο εκπαίδευσής ώστε το μοντέλο μας να έχει μηδενική γνώση του συνόλου επαλήθευσης πριν την τελική αξιολόγηση.

Στη συνέχεια έγιναν κάποιες δοκιμές για την εύρεση του βέλτιστου αριθμού κατηγοριών που θα έπρεπε να δημιουργηθούν. Χρησιμοποιήσαμε ένα word2vec μοντέλο με το οποίο μετρούσαμε την απόδοση του αλγόριθμου για κάθε αριθμό κατηγοριών από 1 έως 15. Να σημειωθεί ότι η βέλτιστη επιλογή μοντέλου βάση αυτής της μετρικής δε συνεπάγεται και βελτιστοποίηση χαρακτηριστικών εισόδου του τελικού αλγόριθμου ταξινόμησης. Τελικά ο αλγόριθμος κατέταξε τις εταιρείες σε 13 κατηγορίες των οποίων οι θεματικές ενότητες ορίζονται από τις ακόλουθες λέξεις – κλειδιά:

- pain, metabol , inflammatori , chronic, cardiovascular , prevent , disorder , drug ,disease , treatment
- vascular , image , minim , invas , surgi , technolog , tissue , medic device , medic , devic
- platform, deliveri, candid, target, small molecul, small, drug discoveri, molecul, discoverie, drug
- biotechnolog, applic, industri, research, medicin, servic, antibodi, chemic, technolog, product
- platform, clinic, genet, assay, diagnosi, technolog, detect, molecular, test, diagnostic
- proprietari, molecul, biolog, natur, target, autoimmun, protein, cell
- oncolog, cancer, agent, infecti diseas, target, autoimmun, protein, disease, therapeut
- effect, biotechnolog, industri, medic, pain, patient, research, high, clinic, pharmaceut
- drug discoveri , cell, manufactor, drug, discover, pain, vascular, chronic, diseas, therapi
- molecul, biotech, treatment, target, person, patient, antibodi, inhibitor, activ,
- data, softwar, servic, manag, medic, platform, onlin, patient, health, healthcar
- implant, technolog, diabet, design, heart, drug deliveri, deliveri system, monitor, deliveri system
- test, protein, biotech, innov, mannufactur, organ, research, servic, clinic, solut

Στη συνέχεια, αντιστοιχήθηκε ένας αριθμός σε κάθε κατηγορία και κατ' επέκταση σε κάθε εταιρεία. Με τον τρόπο αυτό δημιουργήθηκε ένα καινούριο χαρακτηριστικό εισόδου του αλγορίθμου μας με ονομασία category και σύνολο τιμών [0,12]. Μετά από δοκιμές στον αλγόριθμό μας καταλήξαμε στα συμπεράσματα ότι λόγω και της δενδρικής δομής των λύσεών μας θα ήταν προτιμότερο να δημιουργηθούν διαφορετικά δυαδικά πεδία με όνομα tag_i, ένα για την κάθε κατηγορία κάθε ένα από τα οποία θα έχει την τιμή 1 αν και μόνο αν μια εταιρεία ανήκει στη συγκεκριμένη κατηγορία. Στη συνέχεια, εφόσον λόγω του μικρού μεγέθους του dataset θέλαμε να περιορίσουμε τον αριθμό των features, επιλέξαμε να χρησιμοποιήσουμε στην είσοδο μόνο συγκεκριμένες από τις κατηγορίες. Παρακάτω παρουσιάζεται ένα διάγραμμα που εμφανίζει το μέσο όρο των κλειστών και των εν ενεργεία εταιρειών για την κάθε κατηγορία 3.2 .



Σχήμα 3.2: Ανά κατηγορία επιτυχία εταιρειών

Η πρώτη παρατήρησή μας είναι ότι δεν υπάρχει κάποιος σαφής κανόνας που καθιστά τις επιχειρήσεις που λειτουργούν σε συγκεκριμένα πεδία πιο επιρρεπείς στην επιτυχία ή την αποτυχία. Για τις περισσότερες κατηγορίες, η κατανομή των κλειστών-ανοιχτών εταιρειών είναι ισομοιρασμένη. Αντιλαμβανόμαστε ότι αυτές οι κατηγορίες πιθανότατα θα πρέπει να απουσιάζουν από το σύνολο εισόδου τουλάχιστον του αλγόριθμου της λογιστικής παλινδρόμησης, καθώς δεν προσφέρουν σε αυτόν καμία επιπλέον πληροφορία.

3.2.3 Τελική Παρουσίαση dataset – Προβλήματα που κληθήκαμε να αντιμετωπίσουμε

Το σύνολο των τελικών χαρακτηριστικών τα οποία δοκιμάσαμε για να εκπαιδεύσουμε το μοντέλο μας παρουσιάζονται στον επόμενο πίνακα.

Όνομα Χαρακτηριστικού	Περιγραφή	Τύπος
Status	Πληροφορίες για την κατάσταση Λειτουργίας	bin
is_usa	Χώρα Ίδρυσης είναι η Αμερική	bin
Twitter	Ύπαρξη λογαριασμού twitter της εταιρείας	bin
no_info_first_funding	Ανυπαρξία πληροφοριών για την ημερομηνία πρώτης χρηματοδότησης	bin
Funding_rounds	Γύροι χρηματοδότησης	int
Before2013funding	Συνολική χρηματοδότηση σε USD	int
hasRoundA	Ύπαρξη γύρου χρηματοδότησης A	bin
hasRoundB	Ύπαρξη γύρου χρηματοδότησης B	bin
hasRoundC	Ύπαρξη γύρου χρηματοδότησης C	bin
has_VC	Ύπαρξη κεφαλαίου επιχειρηματικών συμμετοχών	bin
years_between_first_funding	Χρόνια μεταξύ ίδρυσης και πρώτου γύρου χρηματοδότησης	int
years_between_first_funding1	Χρόνια μεταξύ ίδρυσης και πρώτου γύρου χρηματοδότησης $\in [0, 1]$	bin
years_between_first_funding2	Χρόνια μεταξύ ίδρυσης και πρώτου γύρου χρηματοδότησης $\in [2, 3]$	bin
years_between_first_funding3	Χρόνια μεταξύ ίδρυσης και πρώτου γύρου χρηματοδότησης $\in [4, inf)$	bin
has_board	Ύπαρξή πληροφορίας σχετικά με μέλη διοικητικού συμβουλίου	bin
has_founders	Ύπαρξή πληροφορίας σχετικά με ιδρυτικά μέλη	bin
Number_of_board	Αριθμός μελών συμβουλίου	int
Number_of_founders	Αριθμός ιδρυτών	int
NO_FUNDING_INFO	Συνολική χρηματοδότηση : NULL	bin
fundingrounds<2.5	Γύροι χρηματοδότησης < 2.5	bin
fundingrounds<5	Γύροι χρηματοδότησης $\in [3, 4]2.5$	bin
fundingrounds>5	Γύροι χρηματοδότησης > 5	bin
more_than_30mil	Before2013FUNDS	binary
10-30_mil	YEARS1	binary
5-10_mil	YEARS2	binary
less-than-5	PUBLICATIONS>2000000	binary
Years_open	Συνολικά χρόνια λειτουργίας μέχρι το 2013 ή μέχρι το κλείσιμο της εταιρείας αν συνέβη νωρίτερα [1]	int
YEARS1	Χρόνια λειτουργίας _in [0,4]	bin
YEARS2	Χρόνια λειτουργίας _in [5,8]	bin
YEARS3	Χρόνια λειτουργίας _in [9,inf)	bin
Years_open	Συνολικά χρόνια λειτουργίας μέχρι το 2013 ή μέχρι το κλείσιμο της εταιρείας αν συνέβει νωρίτερα	int
PUBLICATIONS>2000000	Αριθμός publications > 2000000	bin
PUBLICATIONS<1000000	Αριθμός publications < 1000000	bin
Number_of_investors	Συνολικός αριθμός επενδυτών	int
number_of_science_degrees	Συνολικός αριθμός πτυχίων των ιδρυτών που σχετίζονται με ιατρικές - φυσικές επιστήμες	int
number_of_technology_degrees	Συνολικός αριθμός πτυχίων των ιδρυτών που σχετίζονται με την τεχνολογία	int
isbiotech	Ανήκει σε κατηγορία με Tag biotechnology	int
has_site	Υπάρχει επίσημη ιστοσελίδα της εταιρείας στο διαδίκτυο	int
category	Αναγνωριστικό της κατηγορίας στην οποία ανήκει η εταιρεία	$\in [0, 12]$
tag0	Η εταιρεία ανήκει στην κατηγορία tag0	bin
tag1	Η εταιρεία ανήκει στην κατηγορία tag1	bin
tag2	Η εταιρεία ανήκει στην κατηγορία tag2	bin
tag3	Η εταιρεία ανήκει στην κατηγορία tag3	bin
tag4	Η εταιρεία ανήκει στην κατηγορία tag4	bin
tag5	Η εταιρεία ανήκει στην κατηγορία tag5	bin
tag6	Η εταιρεία ανήκει στην κατηγορία tag6	bin
tag7	Η εταιρεία ανήκει στην κατηγορία tag7	bin
tag8	Η εταιρεία ανήκει στην κατηγορία tag8	bin
tag9	Η εταιρεία ανήκει στην κατηγορία tag9	bin
tag10	Η εταιρεία ανήκει στην κατηγορία tag10	bin
tag11	Η εταιρεία ανήκει στην κατηγορία tag11	bin
tag12	Η εταιρεία ανήκει στην κατηγορία tag12	bin

Πίνακας 3.4: Πίνακας πεδίων που συλλέχθηκαν από τη βάση

Βασικά προβλήματα που αντιμετωπίσαμε κατά την υλοποίηση αυτό του project αφορούν την «ποιότητα» των δεδομένων που λάβαμε. Συγκεκριμένα, όπως αναφέρθηκε ήδη, τα δεδομένα των

[1] Ορισμένες από τις προς παρατήρηση εταιρείες, που ιδρύθηκαν μετά από το 2000 είχαν κλείσει πριν το 2013. Ωστόσο, εξαιτίας της απουσίας πολλών δεδομένων, τις κρατήσαμε, καταγράψαμε όλα τα χαρακτηριστικά που είχαν στη βάση του 2013 και στο συγκεκριμένο πεδίο καταγράψαμε τα συνολικά χρόνια για τα οποία λειτούργησαν.

site με επιχειρησιακές πληροφορίες μπορούν να ενημερώνονται από τους επισκέπτες του εκάστοτε ιστότοπου. Συνεπώς, δεν διαμεσολαβεί κάποια διαδικασία εξακρίβωσης της εγκυρότητάς τους. Υπήρξαν λοιπόν περιπτώσεις που αντιληφθήκαμε ότι συγκεκριμένα πεδία περιείχαν λανθασμένες πληροφορίες με χαρακτηριστικό παράδειγμα startups που ενώ εμφανίζονταν εν λειτουργία ήταν τελικά κλειστές. Όπου αυτή ή παρόμοιες παρατυπίες έγιναν αντιληπτές διορθώθηκαν.

Επιπλέον , προβλήματα αντιμετωπίσαμε και με τις απουσιάζουσες τιμές χαρακτηριστικών. Το dataset που λάβαμε από τη βάση ήταν πολύ «αραιό». Στις περιπτώσεις λοιπόν που θελήσαμε να «σπάσουμε» αριθμητικά δεδομένα σε clusters χαρακτηριστικών, δημιουργήσαμε ένα νέο πεδίο το οποίο είχε τη δυαδική τιμή 1 αν και μόνο αν το αντίστοιχο πεδίο της εισόδου είχε κάποια τιμή ως δεδομένο. Αντιθέτως, αν το πεδίο είχε τιμή NULL, τότε και το καινούριο χαρακτηριστικό, για τη συγκεκριμένη εγγραφή, θα έπαιρνε τιμή 0. Μετά από αυτήν τη διαδικασία αντικαταστήσαμε όλες τις απουσιάζουσες τιμές χαρακτηριστικών με 0 για να μπορέσει να τις χειριστεί αποδοτικά ο αλγόριθμος.

Τέλος, αξίζει να αναφερθούμε στη δυσκολία που αντιμετωπίσαμε κατά τη συλλογή δεδομένων και αφορά ένα φαινόμενο γνωστό ως προκατάληψη επιβίωσης. Το φαινόμενο αυτό αναφέρεται στην «προκατάληψη» που εισάγει στο μοντέλο μας ο τρόπος συγκέντρωσης των δειγμάτων και πρόκειται για ένα σύνηθες χαρακτηριστικό μελετών που ασχολούνται κυρίως με οικονομικά – επιχειρησιακά αντικείμενα. Brown et al. (2015) . Αναλυτικότερα, υποστηρίζουμε ότι οι startup για τις οποίες υπάρχουν διαθέσιμες πληροφορίες στο διαδίκτυο αποτελούν ένα υποσύνολο των startup που ιδρύονται και μάλιστα προνομιούχο υποσύνολο με την έννοια ότι έχουν ξεπεράσει το πρώιμο και ασταθές στάδιο που επέρχεται λίγο μετά τη δημιουργία μιας εταιρείας και έχουν αποκτήσει δημοσιότητα.

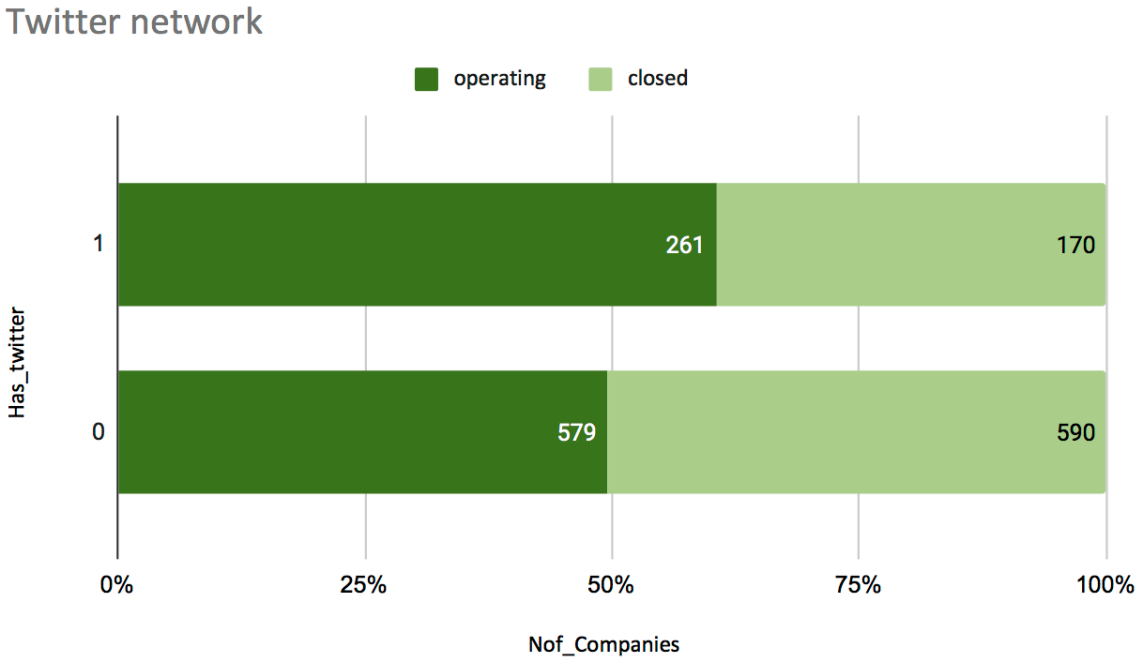
Ρίχνοντας μια πρώτη ματιά στο σύνολο δεδομένων, παρατηρούμε ότι δεν υπάρχουν ξεκάθαρες διακυμάνσεις στα περισσότερα χαρακτηριστικά ανάλογα με τη λειτουργία ή μη της εκάστοτε εταιρείας. Αντιλαμβανόμαστε λοιπόν ότι η διαδικασία της ταξινόμησης θα είναι δύσκολο να επιτευχθεί από τους ταξινομητές. Ενδεικτικά ακολουθούν κάποια διαγράμματα που αφορούν χαρακτηριστικά του dataset μας και δίνουν μια εικόνα των δεδομένων που θα χρησιμοποιήσουμε. Οριακά περισσότερες εταιρείες που έχουν twitter παρουσιάζονται να είναι εν λειτουργία.

Σχετικά με τα οικονομικά στοιχεία έχουμε:

Εδώ παρουσιάζονται οι γύροι χρηματοδότησης κάθε εταιρείας. Στα διαγράμματα με έντονο κόκκινο σημειώνονται οι εν λειτουργία εταιρείες ενώ με ανοιχτό οι κλειστές. Παρατηρούμε και πάλι ότι ο διαχωρισμός τους με βάση αυτό το κριτήριο δεν είναι πολύ σαφής.3.5

Παράλληλα, ενδιαφερθήκαμε να εξετάσουμε αν ο διαχωρισμός σε πεδία ενασχόλησης εισήγαγε κάποια λογική τμηματοποίηση σε ορισμένα οικονομικά χαρακτηριστικά όπως είναι για παράδειγμα η χρηματοδότηση. Μελετήσαμε λοιπόν τις κατανομές για κάθε διαφορετική κατηγορία. Ο περιορισμένος αριθμός όμως των δειγμάτων δεν μας βοήθησε να εξάγουμε με ασφάλεια συγκεκριμένα αποτελέσματα. Ενώ στις μέσες τιμές των κατανομών παρατηρήσαμε μεγάλες διαφορές της τάξης του 70% η διακύμανση που μετρήσαμε σε κάθε κατανομή υπήρξε τόσο μεγάλη που το confidence level σε όποια παρατήρηση επιχειρούσαμε να κάνουμε θα ήταν μηδενικό.

Όπως αντιλαμβανόμαστε και από τα διαγράμματα 3.5, 3.4, 3.2, ο διαχωρισμός των εταιρειών δεν ακολουθεί κάποιο εύκολα αντιληπτό μοτίβο. Ο συνδυασμός κατάλληλων χαρακτηριστικών πιθανόν να κρύβει περισσότερη πληροφορία.



Σχήμα 3.3: Δίκτυο twitter ως δείκτης επιτυχίας

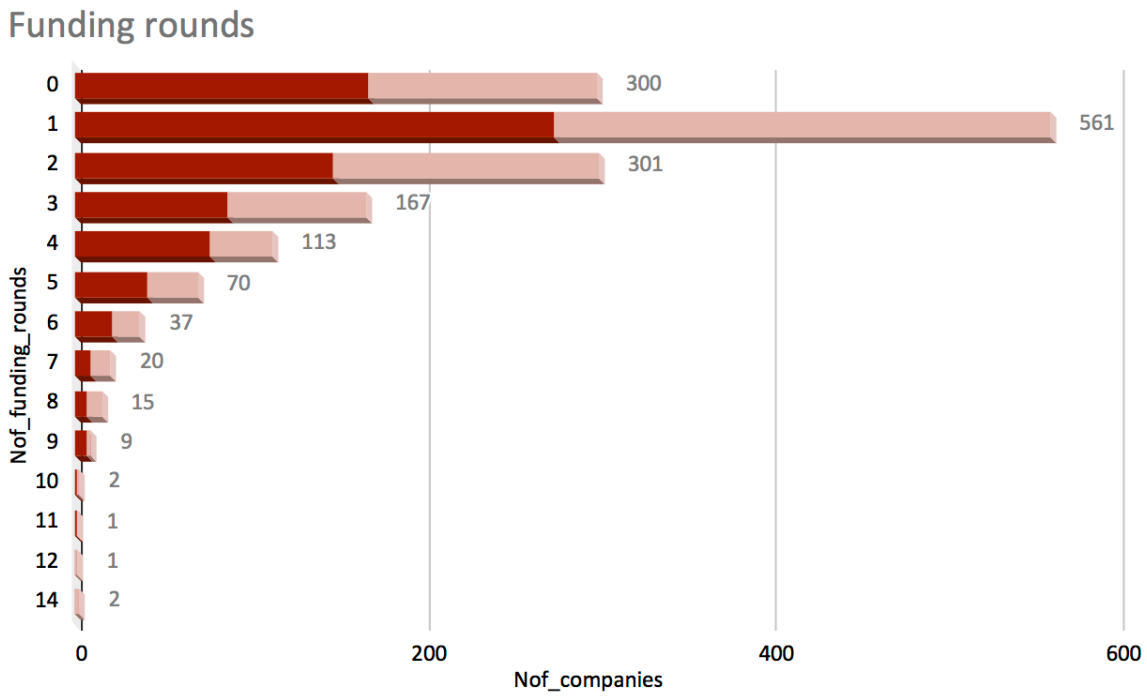
3.2.4 Προετοιμασία συνόλων δεδομένων πριν τη χρήση των αλγόριθμων μηχανικής μάθησης

Προτού εισάγουμε τα δεδομένα στους αλγορίθμους, έπρεπέ να αντιμετωπίσουμε την ανισορροπία των κλάσεων που αναφέρθηκε και προηγουμένως. Συγκεκριμένα, στο αρχικό μας dataset το 75% των δειγμάτων είχαν label 1 (επιτυχία) και το 25% μόνο είχαν label 0 (αποτυχία). Γίνεται αντιληπτό πως μια startup για να φτάσει στο σημείο να εγγραφεί σε κάποιο πληροφοριακό ιστότοπο θα έχει περάσει προ πολλού το αρχικό στάδιο λειτουργίας στο οποίο διατρέχει και το μεγαλύτερο κίνδυνο αποτυχίας. Τα ποσοστά που αναφέραμε παραπάνω ήταν αποτρεπτικά για να κάνουμε προβλέψεις στο συγκεκριμένο σύνολο δεδομένων, δεδομένου του μικρού μεγέθους δείγματος που κατείχαμε. Ακολουθήσαμε λοιπόν, τρεις διαφορετικές πρακτικές.

1. Υλοποιήσαμε oversampling με τη βοήθεια της συνάρτησης της python SMOTE ώστε να δημιουργηθούν επιπλέον 'εικονικές' εταιρείες που θα χρησιμοποιηθούν,μαζί με μερικές από τις εταιρείες που λάβαμε από τη βάση, στους αλγορίθμους μόνο κατά τη διαδικασία εκπαίδευσης . Οι προβλέψεις και τα τελικά αποτελέσματα θα προκύψουν χρησιμοποιώντας υπαρκτές εταιρείες. Θα αναφερόμαστε στο σύνολο αυτό ως dataset1.

2. Μειώσαμε τον αριθμό των θετικών δειγμάτων του dataset ώστε να προκύψουν ίσες αριθμητικά κλάσεις. Πρόκειται για το dataset2 με δείγμα μεγέθους 778 και απόλυτη ισορροπία των δύο κλάσεων.

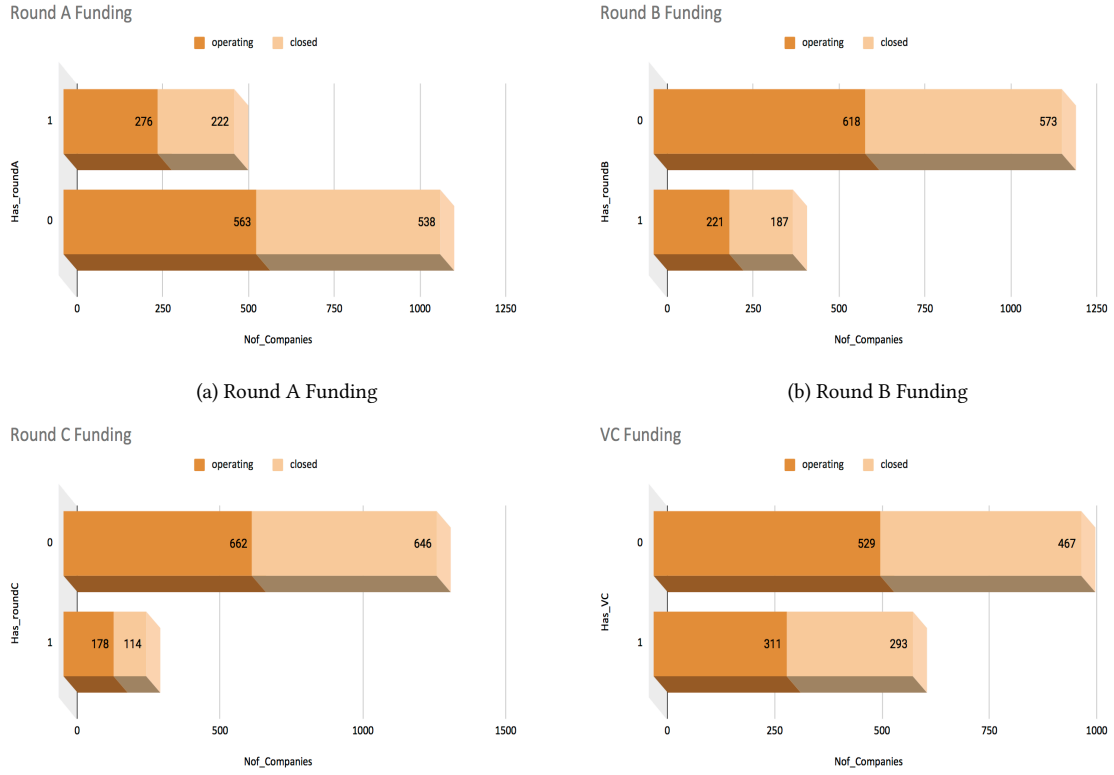
3. Πραγματοποιήσαμε ένα είδος dummy oversampling όπου επιλέξαμε να δημιουργήσουμε αντίγραφα των κλειστών startup και να εισάγουμε κάποιον λευκό θόρυβο ώστε να προκύψουν νέες, παραπλήσιες με τις κλειστές, εταιρείες. Η διαδικασία που ακολουθήσαμε περιγράφεται ως εξής. Αρχικά χωρίσαμε τις εταιρείες ανά κατηγορία, την οποία και κρατήσαμε αμετά-



Σχήμα 3.4: Number of funding rounds and success rate

βλητη. Για κάθε κατηγορία, και κάθε ένα από τα πεδία funding , funding_rounds, years_bt_funding , publications , years_operting , υπολογίσαμε την τυπική απόκλιση f. Στη συνέχεια, δημιουργήσαμε ένα αντίγραφο για κάθε υπάρχουσα εταιρεία που η τιμή της για τα πεδία αυτά επιλέχθηκε τυχαία από το διάστημα [-f + initial_val , initial_val +f]. Τέλος, πριν την οριστική εγγραφή της παραπάνω τιμής φροντίσαμε να διατηρήσουμε το ποσοστό απουσιάζουσων τιμών ανά χαρακτηριστικό ανά κατηγορία και έτσι την εισάγαμε με πιθανότητα ίση με το εκάστοτε sparsity ποσοστό, αλλιώς εισάγαμε NULL. Όσον αφορά τα υπόλοιπα δυαδικά χαρακτηριστικά για κάθε νέα εταιρεία αντιστρέψαμε το καθένα σε σχέση με την αρχική με μία πολύ μικρή πιθανότητα. Έτσι προέκυψε το dataset3 με συνολικό μέγεθος 1600 εταιρείες από τις οποίες οι 840 είναι εν ενεργεία και οι υπόλοιπες 760 είναι κλειστές. Τα ποσοστά των δύο κλάσεων κυμαίνονται δηλαδή στα 52,5% και 47,5% αντίστοιχα.

Στη συνέχεια δοκιμάσαμε διαφορετικούς συνδυασμούς χαρακτηριστικών για κάθε αλγόριθμο. Ο λόγος που δεν εισάγαμε σαν είσοδο όλα τα χαρακτηριστικά που συλλέξαμε έχει να κάνει με το μέγεθος του dataset το οποίο ήταν πολύ μικρό σε σχέση με το πλήθος των χαρακτηριστικών. Έτσι, έπειτα από μια σειρά δοκιμών καταλήξαμε στην κατάλληλη - βέλτιστη είσοδο για κάθε αλγόριθμο. Στη διαδικασία αυτή φροντίσαμε να υπάρχει όσο το δυνατόν μικρότερη συσχέτιση μεταξύ των χαρακτηριστικών ώστε να συμβάλουμε στην αποδοτικότερη λειτουργία των ταξινομητών, καθώς όπως έχει μελετηθεί για τις δομές που χρησιμοποιούμε η ύπαρξη συσχετιζόμενων χαρακτηριστικών επιδρά αρνητικά στη διαδικασία της μάθησης Toloşi and Lengauer (2011). Τα αποτελέσματα μας αναλύονται στην επόμενη ενότητα.



(a) Round A Funding

(b) Round B Funding

(c) Round C Funding

(d) VC Funding

Σχήμα 3.5: Kind of funding rounds and success rate

Κεφάλαιο 4

Αποτελέσματα και Συγκρίσεις

Στην παρούσα ενότητα θα αξιολογήσουμε τα αποτελέσματα που προκύπτουν κατά την εκπαίδευση των τριών αλγόριθμων μηχανικής μάθησης που παρουσιάστηκαν στο Κεφάλαιο 3. Εκπαιδεύουμε τους αλγορίθμους αυτούς 3 φορές χρησιμοποιώντας κάθε φορά ένα από τα dataset που περιγράψαμε στην προηγούμενη ενότητα 3.2.4. Προκύπτουν έτσι 9 διαφορετικά μοντέλα τα οποία θα αξιολογηθούν με τη βοήθεια ορισμένων μετρικών που αναλύονται στην πρώτη υποενότητα του παρόντος κεφαλαίου 4.1. Θα αξιολογήσουμε έτσι τόσο τους διαφορετικούς αλγορίθμους σχετικά με την καταλληλότητά τους για το συγκεκριμένο πρόβλημα, όσο και τα διαφορετικά dataset αλλά και τα συλλεχθέντα χαρακτηριστικά.

4.1 Μετρικές Αξιολόγησης

Για τη μέτρηση της αποδοτικότητας του μοντέλων μηχανικής μάθησης χρησιμοποιούνται συνήθως οι ακόλουθες μετρικές:

- Ακρίβεια (precision) : Είναι ο λόγος των σωστών προβλέψεων για τη θετική κλάση ως προς όλες τις προβλέψεις που έγιναν για τη θετική κλάση, δηλαδή ισούται με:

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

  ,όπου TP οι σωστές προβλέψεις για την θετική κλάση και FP οι λανθασμένες προβλέψεις για τη θετική κλάση

- Ανάκληση (recall) : Είναι ο λόγος των σωστών προβλέψεων για τη θετική κλάση ως προς το άθροισμα των σωστών προβλέψεων για τη θετική κλάση με τις λάθος προβλέψεις για την αρνητική κλάση και έχει ως εξής:

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

  ,όπου FN οι λανθασμένες προβλέψεις για τη αρνητική κλάση

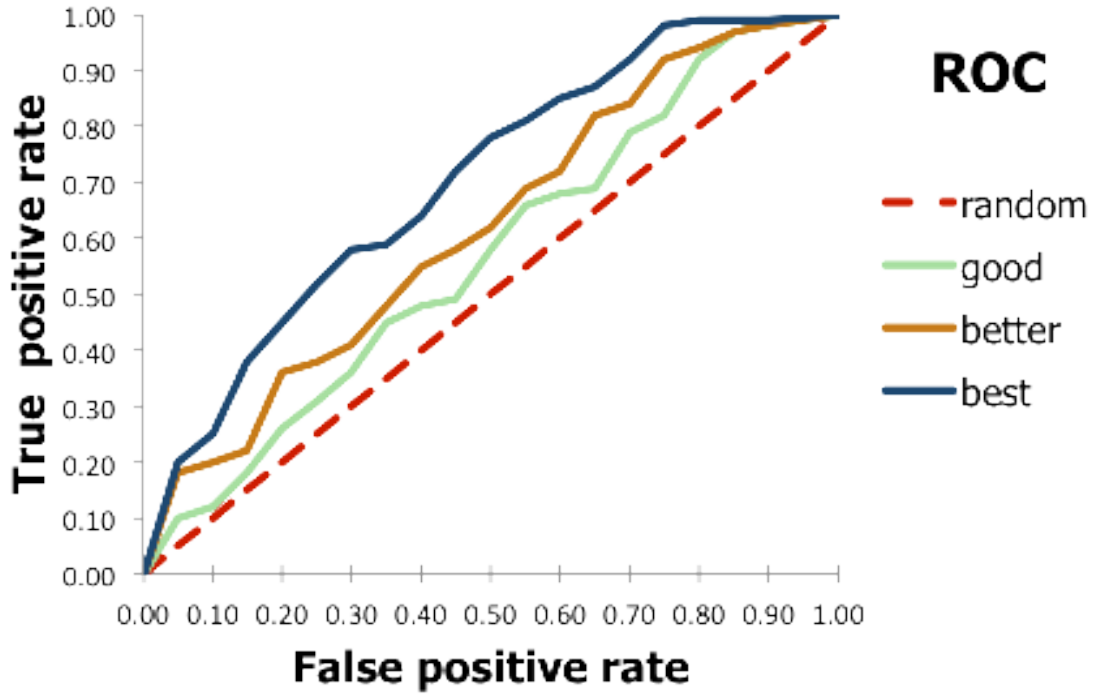
- f1-score : Είναι μια μετρική ακρίβειας που ισούται με τον αρμονικό μέσο της ακρίβειας και της ανάκλησης, προκύπτει δηλαδή από την παρακάτω μαθηματική έκφραση:

  $$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Ακόμα, μια σημαντική πληροφορία για την περιγραφή ενός συνόλου χαρακτηριστικών είναι ο συντελεστής συσχέτισης τους. Ως συσχέτιση ορίζεται η μέτρηση της γραμμικής σχέσης μεταξύ δύο ποσοτικών μεγεθών. Όπως περιγράφεται στο άρθρο Why Feature Correlation Matters...A Lot! θετική συσχέτιση σημαίνει ότι όταν το χαρακτηριστικό Α αυξάνεται, τότε και το Β αυξάνεται και

αντίστροφα. Κατ' αναλογία, αρνητική συσχέτιση δείχνει ότι η αύξηση του ενός χαρακτηριστικού οδηγεί στη μείωση του άλλου.

Τέλος, μια καθοριστική μετρική για την αξιολόγηση μιας μεθόδου ταξινόμησης είναι η μετρική AUC_score. Η μετρική αυτή, σύμφωνα με το Useful properties of ROC curves and AUC scoring λειτουργεί ως εξής: Για κάθε παρατήρηση υπολογίζεται μια πιθανότητα, στην συνέχεια ταξινομούνται τα δεδομένα με βάση την τιμή της προβλεπόμενης πιθανότητας κι αξιολογείται το πόσο κοντά είναι αυτή η διάταξη στην βέλτιστη σειρά που μπορούν να διαταχθούν τα δεδομένα. Η περιγραφή αυτή δείχνει ότι η μετρική αυτή είναι ανεξάρτητη της κλίμακας ή οποιασδήποτε μεταμόρφωσης που διατηρεί σχετική διάταξη. Πρόκειται για μια χρήσιμη μετρική σε προβλήματα ταξινόμησης με ανισορροπία κλάσεων όπως φαίνεται μέσα από το παρακάτω απλό παράδειγμα. Ας πούμε ότι θέλουμε να προβλέψουμε το ενδεχόμενο ένας άνθρωπος να έχει άνοια χρησιμοποιώντας ως κριτήριο τον τρόπο ομιλίας του. Γνωρίζουμε ότι 99% των ατόμων δεν έχουν άνοια, πρόκειται δηλαδή για μία ασθένεια που έχει το 1% του πληθυσμού. Αν δημιουργήσουμε έναν ταξινομητή που θα προβλέπει πάντα πως οι άνθρωποι δεν΄έχουν άνοια τότε θα έχει πιστότητα - accuracy 99%. Ωστόσο, το μοντέλο μας ουσιαστικά δεν θα μπορεί να διακρίνει καθόλου μεταξύ των δύο κλάσεων για αυτό και η τιμή της AUC θα είναι 0,5. Για τον υπολογισμό της AUC χρησιμοποιείται το διάγραμμα της καμπύλης Λειτουργικού Χαρακτηριστικού Δέκτη (ROC) 4.1 . Πρόκειται για ένα συνεχές γράφημα που ορίζουν τα σημεία (FP,TP) για όλα τα δυνατά σημεία απόφασης στο μοναδιαίο τετράγωνο. Η τιμή της AUC καθορίζεται από το εμβαδό που ορίζει το γράφημα αυτό και ο άξονας x που αντιστοιχεί στο βαθμό των εσφαλμένα ταξινομημένων θετικά παρατηρήσεων (FPR) Calì and Longobardi (2015).



Σχήμα 4.1: Παράδειγμα διαγράμματος ROC

4.2 Αποτελέσματα

Παρόλο που στη διάθεσή μας είχαμε αρκετά χαρακτηριστικά , το σύνολο εκπαίδευσης όπως ήδη αναφέραμε ήταν περιορισμένο. Για το λόγο αυτό δε θελήσαμε να τα εισάγουμε όλα ως είσοδο στους αλγορίθμους καθώς θα δυσχέραιναν τη διαδικασία της μάθησης. Επιπλέον, αντιληφθήκαμε σε ορισμένες περιπτώσεις ότι η χρήση των χαρακτηριστικών με τη μορφή ακεραίου, χωρίς να χρησιμοποιηθούν δηλαδή τα δυαδικά πεδία που προέκυψαν από την τμηματοποίηση του συνόλου τιμών, οδηγούσε σε πιο ακριβή κατηγοριοποίηση και έτσι προτιμήθηκαν.

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα που προέκυψαν από την εκπαίδευση των μοντέλων με σύνολα εκπαίδευσης (train set) 67% επί του συνολικού δείγματος και επαλήθευσης (test set) 33% αντίστοιχα.

4.2.1 Λογιστική Παλινδρόμηση

Θα ξεκινήσουμε από την παρουσίαση των αποτελεσμάτων που προέκυψαν από τον αλγόριθμο της λογιστικής παλινδρόμησης. Ξεκινάμε με το dataset1 όπου εφαρμόσαμε συνάρτηση oversampling για να εξισορροπήσουμε το σύνολο εισόδου:

	Precision	Recall	F1_score	Samples
0	0.54	0.51	0.53	125
1	0.79	0.81	0.80	278
weighted AVG	0.71	0.71	0.71	403

Πίνακας 4.1: Σύνολο δεδομένων από oversampling LR

Συνεχίζουμε με το ισορροπημένο dataset2 με την προσθήκη λευκού θορύβου για τη γέννηση νέων αποτυχημένων εταιρειών:

	Precision	Recall	F1_score	Samples
0	0.78	0.69	0.73	239
1	0.77	0.84	0.80	290
weighted AVG	0.77	0.77	0.77	529

Πίνακας 4.2: Σύνολο δεδομένων από δημιουργία εταιρειών με προσθήκη θορύβου LR

Στην προσπάθεια αυτή λάβαμε AUC 76,6% στο test set ενώ στο training set είχαμε accuracy 75% και AUC 74,4%

	Precision	Recall	F1_score	Samples
0	0.66	0.82	0.73	122
1	0.79	0.61	0.69	135
weighted AVG	0.73	0.71	0.71	257

Πίνακας 4.3: Ισορροπημένο σύνολο δεδομένων LR

Εδώ, είχαμε 71,7% AUC στο test ενώ στο train οι τιμές των AUC και accuracy ήταν 70,1% και 71% αντίστοιχα.

4.2.2 Τυχαία Δάση

Αντίστοιχα, για τα τυχαία δάση λάβαμε τα ακόλουθα αποτελέσματα. Dataset1:

	Precision	Recall	F1_score	Samples
0	0.49	0.66	0.56	125
1	0.82	0.70	0.75	278
weighted AVG	0.72	0.68	0.69	403

Πίνακας 4.4: Σύνολο δεδομένων από oversampling

Για τις μετρικές test και train AUC λάβαμε τιμές 67,1% και 68,3% αντίστοιχα. Dataset2:

	Precision	Recall	F1_score	Samples
0	0.84	0.71	0.77	239
1	0.79	0.89	0.84	290
weighted AVG	0.81	0.81	0.81	529

Πίνακας 4.5: Σύνολο δεδομένων με θόρυβο

Η AUC στην περίπτωση αυτή είχε τιμή 80% ενώ στο train set είχαμε 85,5% AUC και 86% Accuracy. Τέλος, το μικρό σύνολο δεδομένων, dataset3 μας έδωσε τα ακόλουθα αποτελέσματα:

	Precision	Recall	F1_score	Samples
0	0.72	0.82	0.77	122
1	0.82	0.72	0.76	135
weighted AVG	0.77	0.77	0.77	257

Πίνακας 4.6: Ισορροπημένο σύνολο δεδομένων RF

Η AUC στην περίπτωση αυτή είχε τιμή 76,9% ενώ στο train set είχαμε 82,1% AUC και 82% Accuracy.

4.2.3 Catboost

Τέλος συγκεντρώνουμε στους επόμενους πίνακες τα αποτελέσματα που μας έδωσε ο αλγόριθμος catboost για τις διαφορετικές εισόδους. Dataset1:

	Precision	Recall	F1_score	Samples
0	0.63	0.60	0.61	125
1	0.82	0.84	0.83	278
weighted AVG	0.76	0.77	0.77	403

Πίνακας 4.7: Σύνολο δεδομένων από oversampling catboost

AUC:72,1% TEST: ACC 81% auc 81,3%

Dataset2:

	Precision	Recall	F1_score	Samples
0	0.84	0.79	0.82	239
1	0.83	0.89	0.86	290
weighted AVG	0.84	0.84	0.84	529

Πίνακας 4.8: Σύνολο δεδομένων με θόρυβο catboost

Στην περίπτωση αυτή η ακρίβεια του train set ήταν 86% και η περιοχή AUC 85,3% σε αντίθεση με το test όπου είχαμε AUC 83,8%

Dataset3:

	Precision	Recall	F1_score	Samples
0	0.76	0.85	0.80	122
1	0.85	0.76	0.80	135
weighted AVG	0.80	0.80	0.80	257

Πίνακας 4.9: Ισορροπημένο σύνολο δεδομένων Catboost

Η AUC στην περίπτωση αυτή είχε τιμή 80,4% ενώ στο train set είχαμε 87,2% AUC και 86% Accuracy. [1]

4.3 Ερμηνεία

Από τα αποτελέσματα της Ενότητας 4.2 είναι διακριτό πως κάθε αλγόριθμος συμπεριφέρεται διαφορετικά βάσει του συνόλου δεδομένων (dataset) που εφαρμόζεται. Η Ενότητα αυτή προσπαθεί να αποτυπώσει και να αξιολογήσει τα αποτελέσματα αυτά ως εξής: αρχικά, η υποενότητα 4.3.1 επικεντρώνεται στην αξιολόγηση του συνόλου δεδομένων (dataset) προκειμένου να υποδείξει την επίδραση των δεδομένων εισόδου στο τελικό αποτέλεσμα. Στη συνέχεια, 4.3.2 αξιολογού-

[1] Στο σημείο αυτό θα θέλαμε να τονίσουμε το εξής: Το catboost χρησιμοποιεί την GPU για την εκτέλεση του κώδικά του. Τα προβλήματα λοιπόν αντιμετωπίζονται με μία μη ντετερμινιστική προσέγγιση λόγω του πολυεπίπεδου παραλληλισμού. Οι ατομικές ενέργειες που εντοπίζονται στον κώδικα γίνονται κάθε φορά με διαφορετική σειρά, με αποτέλεσμα διαφορετική σειρά στις πράξεις μεταξύ δεκαδικών και τελικά διαφορετικό αποτέλεσμα. Για το λόγο αυτό κατά την εκπαίδευση στα ίδια δεδομένα λαμβάναμε πάντα διαφοροποιημένο αποτέλεσμα του οποίου όμως η διακύμανση υπήρξε μικρή της τάξης του 5%.

νται λεπτομερώς τα επιμέρους χαρακτηριστικά, έτσι ώστε να αναγνωριστούν αυτά που παίζουν σημαντικότερο ρόλο στο μοντέλο καθώς επίσης και της συσχετίσεις μεταξύ τους. Έπειτα 4.3.3, αξιολογούνται οι ίδιοι οι αλγόριθμοι, τόσο ως προς την απόδοσή τους, όσο και ως προς τις παραμέτρους που απαιτήθηκαν για την εκτέλεσή τους. Τέλος, η υποενότητα 4.3.4 παρουσιάζει μία σύγκριση των αποτελεσμάτων της παρούσας διπλωματικής εργασίας με άλλες, παρόμοιες μελέτες της βιβλιογραφίας.

4.3.1 Αξιολόγηση συνόλου δεδομένων (dataset)

Όπως είναι εμφανές και από τους 3 παραπάνω αλγόριθμους που χρησιμοποιήσαμε η χρήση του DATASET1 αποδείχτηκε ισοπεδωτική για την σωστή ταξινόμηση των δειγμάτων και ιδίως όσων ανήκουν στην ασθενέστερη κατηγορία, δηλαδή τις κλειστές εταιρείες. Η υλοποίηση του oversampling με τις βιβλιοθήκες της python που χρησιμοποιήσαμε στο σύνολο δεδομένων αυτό, στηρίζεται στην δημιουργία νέων δειγμάτων αξιοποιώντας υπάρχοντα δείγματα και τους κοντινότερος γείτονές τους που ανήκουν στην ίδια κλάση. Η ουσία της τεχνικής αυτής είναι ότι δημιουργεί νέα δείγματα στο n-διάστατο χώρο που ορίζεται από το δείγμα "στόχο", κάποιον γείτονά του και κάποιον τυχαίο συντελεστή στο (0,1). Όμως, μέσω της διαδικασίας αυτής, οποιουδήποτε είδους συσχετίσεις έχουμε στα δεδομένα είναι πιθανό να εξασθενήσουν. Για παράδειγμα, η ένταξη των δειγμάτων σε κατηγορίες απαιτεί την ύπαρξη του 1 σε ένα μόνο από τα πεδία tag_i και 0 σε όλα τα υπόλοιπα. Μετά από μελέτη του αποτελέσματος του oversampling παρατηρήσαμε ότι η συνθήκη αυτή δεν ικαναποιούταν. Ακόμα, συσχετίσεις που εντοπίσαμε μεταξύ κατηγορίας και ύψους χρηματοδότησης δεν αποτυπώνονται πάντα στα νέα δείγματα καθώς ενδέχεται για παράδειγμα ως κοντινότερος γείτονας, σύμφωνα με την ευκλείδεια νόρμα, ενός δείγματος να θεωρηθεί μια εταιρεία που ανήκει σε άλλη κατηγορία στην οποία η μέση χρηματοδότηση των εταιρειών είναι μακράν υψηλότερη της πρώτης. Τότε θα επιλεγεί μια τιμή χρηματοδότησης μεταξύ των δύο τιμών των δειγμάτων που ενδέχεται να είναι πολύ υψηλή σε σχέση με την κατηγορία που θα αποδοθεί στη νέα εταιρεία.

Στον τρόπο που δημιουργήσαμε νέες εταιρείες (DATASET3) αποτυπώσαμε τη λογική αυτή που υπήρχε πίσω από τα δεδομένα του train set, όπως περιγράφηκε στην προηγούμενη ενότητα, δημιουργώντας ένα πιο αντιπροσωπευτικό του αρχικού dataset, γεγονός που γίνεται αντιληπτό και στα αποτελέσματά μας.

Τέλος, θα αξιολογήσουμε το DATASET2, ένα dataset στο οποίο δεν διακρίνονται ασθενής και ισχυρή κλάση καθώς έχουμε ίσο αριθμό πραγματικών κλειστών και ανοιχτών επιχειρήσεων. Τα αποτελέσματα που λαμβάνουμε είναι αρκετά ικανοποιητικά και μάλιστα βέλτιστα στις προβλέψεις που αφορούν αποκλειστικά τις κλειστές εταιρείες. Πρόκειται για την πιο δύσκολη από τις δύο κλάσεις (ανοιχτές - κλειστές) σε ό,τι αφορά την ακρίβεια στην πρόβλεψη καθώς στο αρχικό μας dataset έχουμε λιγοστά δείγματα σε σχέση με τις επιτυχημένες startup. Στο dataset2 έχουμε κρατήσει ισάριθμο αριθμό ανοιχτών και κλειστών εταιρειών και παρατηρώντας πως τα μοντέλα μας προβλέπουν καλύτερα την κατηγορία των startup που δεν επιτυγχάνουν κατανοούμε ότι το προφίλ, τα χαρακτηριστικά τους, σκιαγραφούνται από τα δεδομένα που επιλέξαμε με μεγαλύτερη ακρίβεια και πληρότητα. Η απουσία όμως επαρκούς αριθμού δειγμάτων είναι υπαίτια για τα χειρότερα αποτελέσματα που λαμβάνουμε σχετικά με την ταξινόμηση αποκλειστικά των αποτυχημένων startup κατά τη χρήση των δύο άλλων dataset.

4.3.2 Αξιολόγηση χαρακτηριστικών

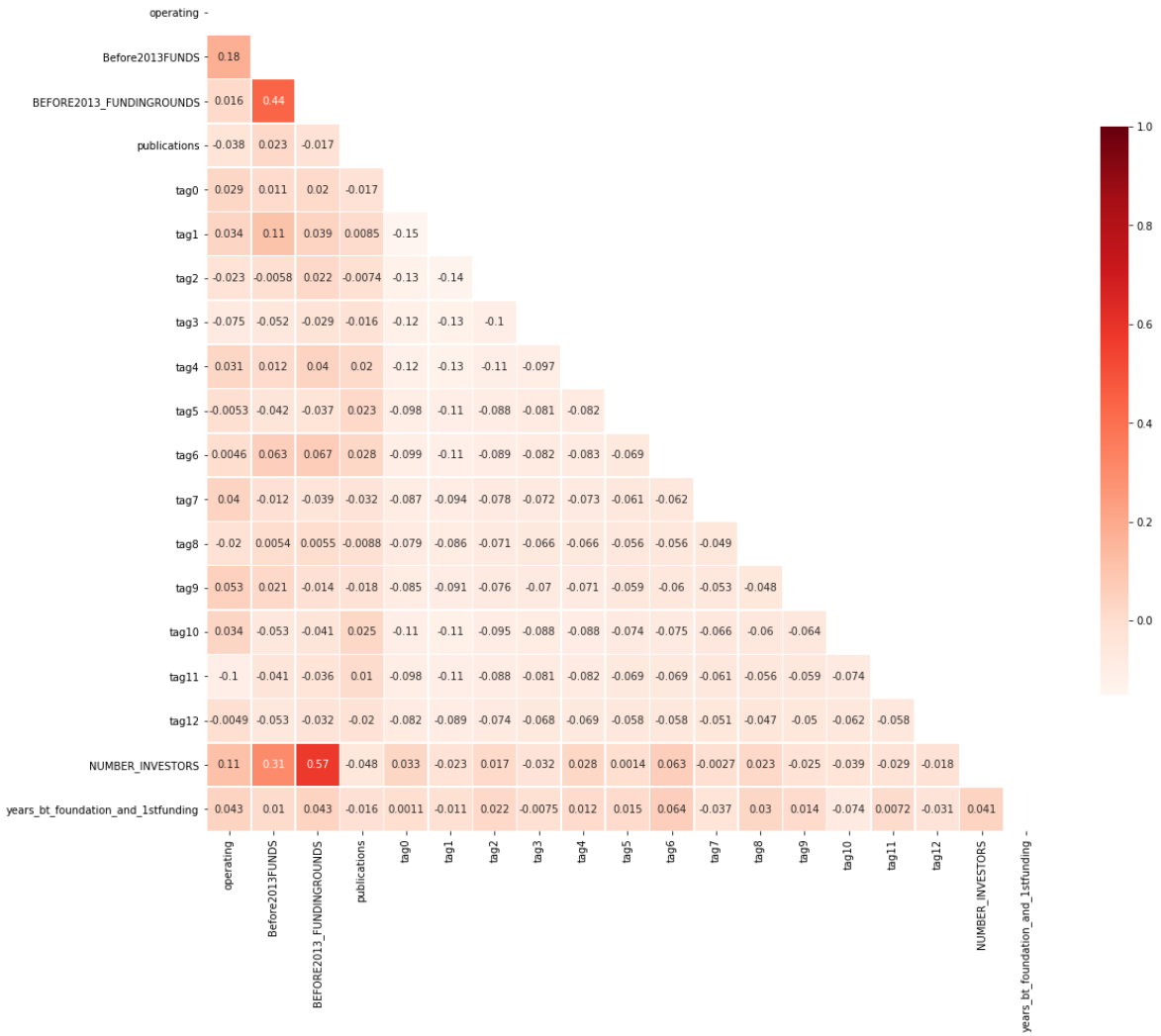
Στην παράγραφο αυτή θα προσπαθήσουμε να αξιολογήσουμε ποιοτικά τα χαρακτηριστικά που επιλέξαμε για τη διαδικασία της εκπαίδευσης. Θα επικεντρωθούμε στην είσοδο των αλγορίθμων που επιλέχθηκε για τα dataset2 και dataset3 καθώς από αυτά λάβαμε τα πιο έγκυρα αποτελέσματα. Τα χαρακτηριστικά που χρησιμοποιήσαμε για την εκπαίδευση κάθε αλγόριθμου παρουσιάζονται στη συνέχεια στον πίνακα 4.10.

Logistic Regression	Random Forests	Catboost
isbiotech	isbiotech	isbiotech
hasRoundA	hasRoundA	hasRoundA
hasRoundB	hasRoundB	hasRoundB
has_VC	has_VC	has_VC
has_board	has_board	has_board
No_info_about_1stfunding	No_info_about_1stfunding	NO_info_about_1stfunding
years_bt_foundation_and_1stfunding	years_bt_foundation_and_1stfunding	years_bt_foundation_and_1stfunding
HASTWITTER	HASTWITTER	HASTWITTER
NO_FUNDING_INFO	NO_FUNDING_INFO	NO_FUNDING_INFO
fundingrounds<2.5	fundingrounds<2.5	fundingrounds<2.5
fundingrounds<3.5	fundingrounds<3.5	fundingrounds<3.5
fundingrounds>5	fundingrounds>5	fundingrounds>5
is_usa	is_usa	is_usa
more_than_30mil	Before2013FUNDS	Before2013FUNDS
10-30_mil	YEARS1	YEARS1
5-10_mil	YEARS2	YEARS2
less-than-5	PUBLICATIONS>2000000	PUBLICATIONS>2000000
YEARS1	PUBLICATIONS<1000000	PUBLICATIONS<1000000
YEARS2	Number_of_investors	Number_of_investors
YEARS3	tag0	tag0
UBLICATIONS>2000000	tag2	tag2
UBLICATIONS<1000000	tag4	tag4
Number_of_investors	tag5	tag5
	tag9	tag9
	tag11	tag11
	tag12	tag12
		tag10
		tag8
		tag1

Πίνακας 4.10: Χαρακτηριστικά εκπαίδευσης

Αρχικά θα αναφερθούμε σε ορισμένες γενικές παρατηρήσεις σχετικά με τα χαρακτηριστικά που χρησιμοποιήσαμε. Προκειμένου να εξετάσουμε την ύπαρξη ή απουσία συσχέτισης μεταξύ των χαρακτηριστικών και πως επηρεάζει την επιτυχία ή αποτυχία αντίστοιχα, δημιουργήσαμε τους πίνακες συσχέτισης που παρουσιάζονται στη συνέχεια 4.2 , 4.3 .

Στο διάγραμμα 4.2 φαίνεται η συσχέτιση των κατηγοριών που δημιουργήσαμε με ορισμένα οικονομικά μεγέθη, με τον αριθμό των δημοσιεύσεων και με την κατάσταση λειτουργίας των εταιρειών. Θεωρήσαμε πως αυτά θα ήταν τα πιο πιθανά χαρακτηριστικά που θα μπορούσαν ίσως να εμφανίσουν κάποια σημαντική συσχέτιση με κάποια από τις κατηγορίες. Σύμφωνα με το διάγραμμα όμως, τα πεδία της χρηματοδότησης, BEFORE2013FUNDS και οι οι γύροι χρηματοδότησης, BEFORE2013_FUNDINGROUNDS, φαίνεται να έχουν σημαντική συσχέτιση, ενώ τα υπόλοιπα χαρακτηριστικά δεν έχουν κάποια σημαντική αλληλεπίδραση. Σχετικά με τις κατηγορίες και την επιτυχία των εταιρειών οι πιο ισχυρές, αν και ακόμα πολύ μικρές, διασυνδέσεις που παρατηρούμε

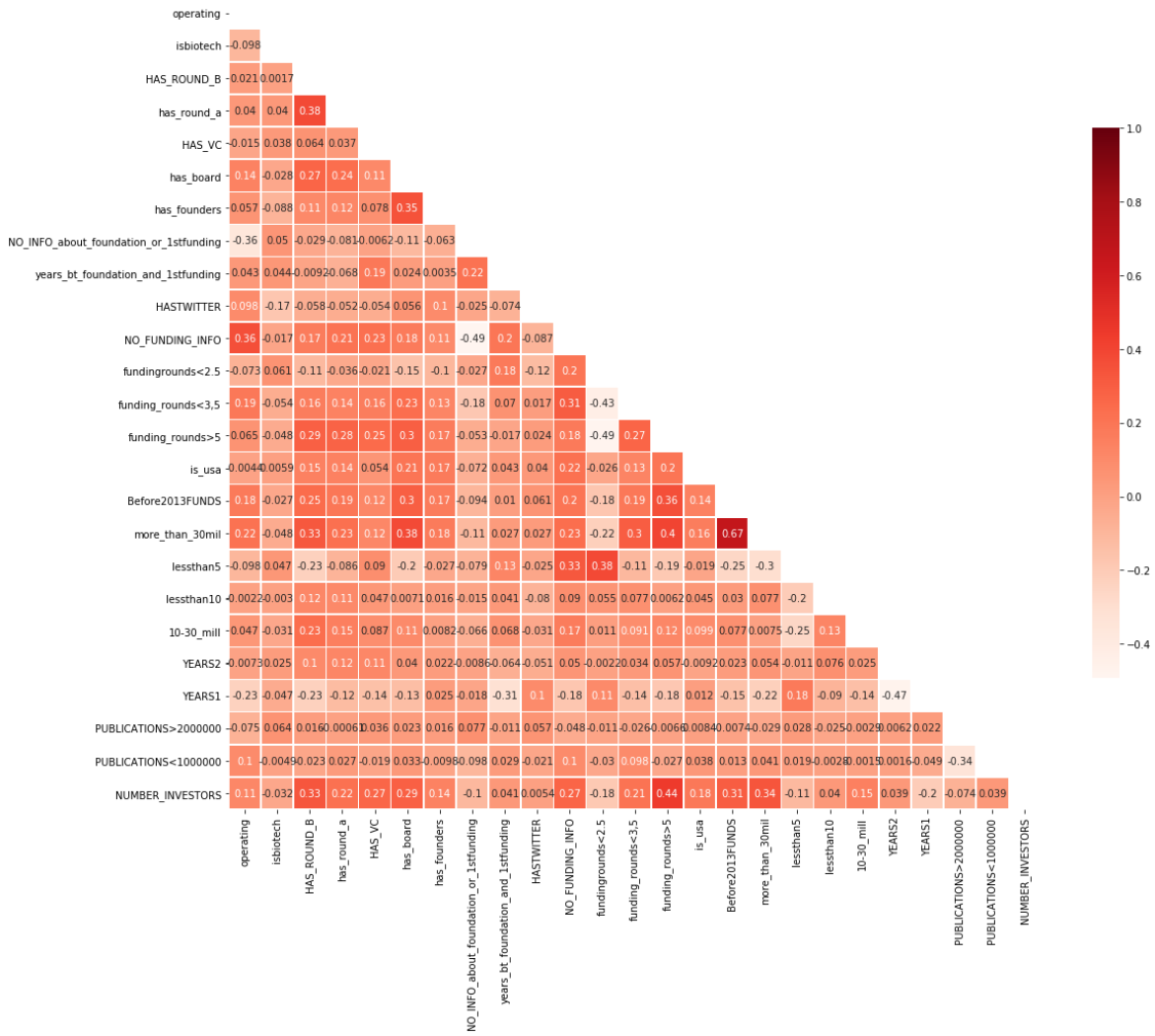


Σχήμα 4.2: Πίνακας συσχέτισης κατηγοριών

αφορούν το tag11 που φαίνεται να μειώνει τις πιθανότητες επιτυχίας όταν μία εταιρεία σχετίζεται με αυτό. Επίσης αξιόλογο είναι το πόσο μικρή σε σχέση με αυτό που θα περιμέναμε είναι η αλληλοσυσχέτιση της επιτυχίας μιας επιχείρησης με το ύψος της χρηματοδότησης που λαμβάνει κατά τη λειτουργία της.

Όσον αφορά τα υπόλοιπα χαρακτηριστικά εκπαίδευσης στον ακόμα πίνακα συσχέτισης 4.3 εξετάζονται οι μεταξύ τους σχέσεις .

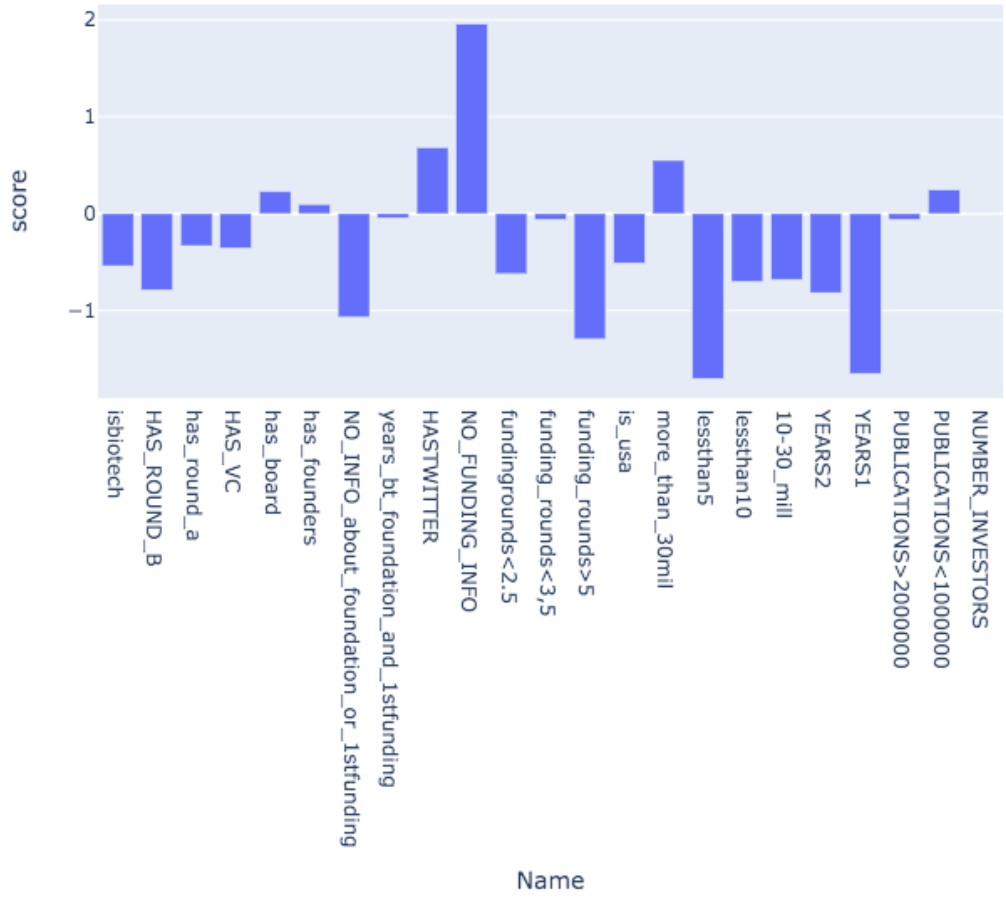
Τα οικονομικά χαρακτηριστικά και εδώ φαίνονται να συσχετίζονται μεταξύ τους. Συγκεκριμένα μεγάλες χρηματοδοτήσεις συνοδεύονται συνήθως από πλήθος γύρων χρηματοδότησης, περισσότερους χρηματοδότες, ύπαρξη γύρων A/B/C. Μια γενικότερη παρατήρηση που μπορούμε να κάνουμε είναι ότι η μεταβλητή πρόβλεψης Y -επιτυχία/αποτυχία- φαίνεται να μην εμφανίζει μεγάλες συσχετίσεις με τα χαρακτηριστικά μας, γεγονός που επιβεβαιώνεται από τη δυσκολία εκπαίδευσης των μοντέλων μας. Μεγαλύτερη συσχέτιση φαίνεται να παρουσιάζει με τα οικονομικά χαρακτηριστικά, όχι όμως στο βαθμό όπου το ένα χαρακτηριστικό θα μπορούσε να μας βοηθήσει να εξάγουμε έγκυρα συμπεράσματα για το άλλο. Τέλος, θα πρέπει να αναφέρουμε ότι η ανυπαρξία υπολογίσιμων συσχετίσεων μεταξύ των δεδομένων μας αποτελεί ένα θεμιτό χαρακτηριστικό



Σχήμα 4.3: Πίνακας συσχέτισης χαρακτηριστικών

των σύνολων δεδομένων που χρησιμοποιούνται σαν είσοδος για αλγορίθμους μηχανικής μάθησης. Τόσο στην περίπτωση της παλινδρόμησης όσο και στα δένδρα αποφάσεων υψηλή συσχέτιση μεταξύ των χαρακτηριστικών μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα και εσφαλμένα συμπεράσματα εξαιτίας της πολυσυγγραμικότητας των χαρακτηριστικών. Στην περίπτωση του πρώτου αλγόριθμου επηρεάζεται ο υπολογισμός των συντελεστών που δημιουργούνται για κάθε χαρακτηριστικό με αποτέλεσμα πολύ μικρές αλλαγές στα δεδομένα να δημιουργούν αδικαιολόγητα σοβαρές αλλαγές στους συντελεστές, ενώ στη δεύτερη, αν επιλεγούν και τα δύο συσχετιζόμενα χαρακτηριστικά για την κατασκευή του δένδρου, είναι πιθανό να ευνοηθούν από το αποτέλεσμα τα δείγματα που ανήκουν στην κλάση με τα λιγότερα στοιχεία.

Για την αξιολόγηση της σημασίας κάθε χαρακτηριστικού που χρησιμοποιήσαμε, σε ό,τι αφορά τον αλγόριθμο της λογιστικής παλινδρόμησης, οι συντελεστές που αντιστοιχούν σε κάθε χαρακτηριστικό είναι ενδεικτικοί της βαρύτητάς του σχετικά με την επιτυχία ή αποτυχία μιας επιχείρησης. Συγκεκριμένα, αποτελούν συντελεστές του λογαρίθμου της συνάρτησης odds με μεταβλητή την πιθανότητα ένα δείγμα να ανήκει στην θετική κλάση. Ένα διάγραμμα των βαρών που προέκυψαν παρατίθεται στη συνέχεια. 4.4

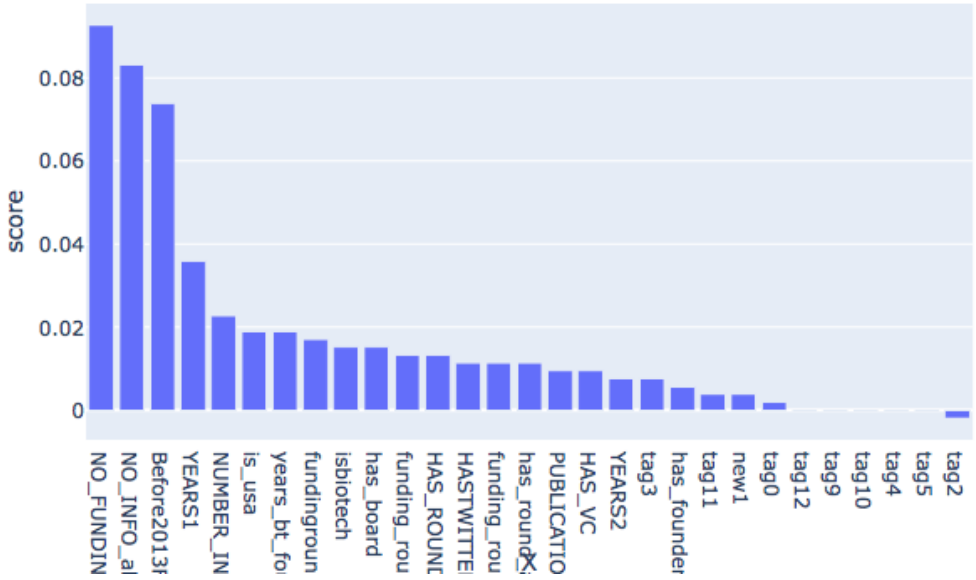


Σχήμα 4.4: Βάρη εκπαίδευσης αλγόριθμου λογιστικής παλινδρόμησης

Όπως φαίνεται από την εικόνα τα βάρη είναι σε μεγάλο βαθμό αρνητικά. Ωστόσο, η σταθερά b που καθορίζει την απόκριση του δικτύου σε μηδενική είσοδο είναι υπολογισμένη στο 2.79. Δεδομένου ότι το σύνολο τιμών των περισσότερων χαρακτηριστικών που έχουμε εισάγει είναι το 0,1, οπότε οι μικροί συντελεστές που φαίνονται στην εικόνα πολλαπλασιασμένοι με τις τιμές των χαρακτηριστικών είναι συγκρίσιμοι με τη σταθερά b, αντιλαμβανόμαστε ότι είναι λογικό να έχουμε αρνητικούς παράγοντες. Με διαφορά μεγαλύτερη σημασία για την εκπαίδευσή μας φαίνεται να έχει η ύπαρξη πληροφορίας για τη χρηματοδότηση που έχει λάβει η εταιρεία, πόρισμα το οποίο ήταν αναμενόμενο. Επιπλέον, ο αλγόριθμος φαίνεται να αδυνατεί να χειριστεί αποτελεσματικά το μοναδικό χαρακτηριστικό το οποίο δεν είναι δυαδικό.

Για τους υπόλοιπους αλγόριθμους η διερεύνηση της σπουδαιότητας των χαρακτηριστικών είναι μια διαδικασία λιγότερο ξεκάθαρη. Έχουν προταθεί διάφορες μέθοδοι αξιολόγησης ώστε να προκύψει μια κατάταξη των χαρακτηριστικών μετά τη βαθμολόγησή τους. Η default συνάρτηση που χρησιμοποιεί η python για να κατατάξει τα χαρακτηριστικά περιγράφει τη συνολική μείωση του δείκτη gini από όλους τους κόμβους των δένδρων του δάσους στους οποίους έχει επιλεγεί να γίνει διάσπαση με βάση το προς μελέτη χαρακτηριστικό. Ωστόσο, έχει διαπιστωθεί πως η τεχνική αυτή συχνά οδηγεί σε εσφαλμένα συμπεράσματα πόσο μάλλον αν η ίδια η μετρική gini έχει χρησιμοποιηθεί ως κριτήριο για την κατασκευή του δάσους οπότε και εισάγεται προκατάληψη στην

όλη διαδικασία. Μια άλλη τεχνική περιγράφεται από της εξής διαδικασία: Υπολογίζουμε για ένα σύνολο εισόδου μία μετρική όπως ακρίβεια, f1 , accuracy κοκ. Στη συνέχεια επιλέγουμε ένα χαρακτηριστικό, του οποίου τη σημασία επιθυμούμε να μετρήσουμε και μεταθέτουμε τυχαία τις τιμές που παίρνει σε όλα τα δείγματα. Στη συνέχεια υπολογίζουμε εκ νέου τη μετρική που επιλέξαμε αρχικά. Η διαφορά μεταξύ των δύο τιμών αποτελεί στην περίπτωσή μας το μέτρο της σημασίας του χαρακτηριστικού. Παρακάτω φαίνεται η σπουδαιότητα των χαρακτηριστικών, όπως προέκυψε με τη χρήση της συγκεκριμένης μεθόδου, για τον αλγόριθμο των τυχαίων δασών με τη χρήση του dataset2 4.5 .



Σχήμα 4.5: Σπουδαιότητα χαρακτηριστικών

Όπως φαίνεται, τα οικονομικά χαρακτηριστικά διαμορφώνουν τους πιο ισχυρούς παράγοντες για την επιτυχία των αλγόριθμων. Πιο σημαντική φαίνεται να είναι η πληροφορία της γνώσης της ημερομηνίας της πρώτης χρηματοδότησης της εταιρείας ενώ μεγάλο βάρος έχει και η ύπαρξη πληροφορίας αναφορικής με τη χρηματοδότηση. Επιπλέον, ο αριθμός των επενδυτών καθώς και οι γύροι χρηματοδότησης φαίνεται να συμβάλουν σημαντικά στη διαμόρφωση του αποτελέσματος. Αναφορικά με τη συνολική χρηματοδότηση, όπως είναι αναμενόμενο στους αλγόριθμους που χρησιμοποιούν δένδρα έπαιξε πολύ σημαντικό ρόλο στη διαμόρφωση του αποτελέσματος. Μάλιστα φάνηκε ότι οι αλγόριθμοι λειτουργούσαν καλύτερα όταν δίναμε ως είσοδο το πεδίο με τη συνολική χρηματοδότηση, χωρίς να το διαιρέσουμε σε επιμέρους clusters. Λόγω του μικρού μεγέθους του dataset το οποίο μας προέτρεψε να χρησιμοποιήσουμε και λιγότερα χαρακτηριστικά και δένδρα μικρότερου ύψους μπορούμε να εξηγήσουμε τη συμπεριφορά αυτή. Ενδιαφέρον είναι το πως το πεδίο αυτό φάνηκε να έχει μηδαμινή βαρύτητα στον αλγόριθμο της λογιστικής παλινδρόμησης. Πράγματι το πεδίο είχε ένα πολύ διευρυμένο σύνολο τιμών που όπως αναφέραμε συσχετιζόταν και με την κατηγορία κάθε εταιρείας και το οποίο θα ήταν δύσκολο να διαχειριστεί ένα μοντέλο που απλά καταχωρεί μια μοναδική τιμή βαρύτητας σε κάθε χαρακτηριστικό. Στην περίπτωση αυτού λοιπόν του μοντέλου χρησιμοποιήσαμε τα δυαδικά χαρακτηριστικά που περιγράφουν τη χρηματοδότηση τα οποία φάνηκε να αποφέρουν πολύ πιο έγκυρα αποτελέσματα.

Περνώντας στις κατηγορίες όπου αντιστοιχήθηκαν στις εταιρείες. Πρόκειται για το μόνο καθαρά κατηγορηματικό χαρακτηριστικό που έπρεπε να διαχειριστούμε. Εφόσον τα δένδρα αποφάσεων σαν δομές δεν δέχονται κατηγορηματικά χαρακτηριστικά δημιουργήσαμε όπως αναφέρθηκε 12 δυαδικά χαρακτηριστικά ένα από τα οποία είχε την τιμή 1 για κάθε δείγμα. Παρόλο που δοκιμάσαμε και άλλους αριθμούς σχετικά με το πλήθος των κατηγοριών όπου θα χωρίζονταν οι εταιρείες και σε μερικές περιπτώσεις προέκυπταν υπολογίσιμες συσχετίσεις με το πεδίο της επιτυχίας - αποτυχίας οι αλγόριθμοι 2 και 3 φάνηκαν να λειτουργούν καλύτερα με την προς παρουσίαση κατηγοριοποίηση. Μάλιστα κατά τη διαδικασία της εκπαίδευσης με τον αλγόριθμο του catboost καθώς σχεδιάζαμε ορισμένα από τα δένδρα που προέκυπταν οδηγούσαν σε ικανοποιητική ταξινόμηση, παρατηρήσαμε ότι σε αρκετές περιπτώσεις επιλέγονταν από τον αλγόριθμο χαρακτηριστικά που αφορούν κατηγορία για την πραγματοποίηση ενός split, ακολουθούμενα από κάποιο κατάλληλα χωρισμένο αριθμητικό συνήθως χαρακτηριστικό. Αναφορικά, η κλάση του classifier του catboost που χρησιμοποιήσαμε δέχεται και διαχειρίζεται με κωδικοποίηση τα κατηγορηματικά πεδία. Αρχικά λοιπόν δοκιμάσαμε να αφήσουμε τον αλγόριθμο να διαχειριστεί μόνος του το χαρακτηριστικό αυτό. Τα αποτελέσματα που προέκυψαν από αυτή τη διαδικασία φάνηκαν να εμφανίζουν μεγαλύτερο overfitting οπότε επιλέξαμε να χρησιμοποιήσουμε και στους δύο αλγορίθμους τα δυαδικά χαρακτηριστικά. Όσον αφορά τον αλγόριθμο λογιστικής παλινδρόμησης δεν χρησιμοποιήσαμε ως είσοδο καθόλου τη συγκεκριμένη πληροφορία καθώς το συγκεκριμένο μοντέλο δεν μπορεί να αξιοποιήσει πληροφορία που του παρέχεται από κατηγορηματικά χαρακτηριστικά. Νόημα θα είχε μόνο να εισάγουμε κάποια δυαδικά πεδία που αντιστοιχούν σε κατηγορίες αν μπορούσαμε να εντοπίσουμε ότι οι εταιρείες της αντίστοιχης κατηγορίας τείνουν να έχουν μεγαλύτερα ποσοστά επιτυχίας - αποτυχίας. Κάτι τέτοιο όμως όπως εξηγήσαμε δεν φαίνεται να συμβαίνει για αυτό και τα εν λόγω πεδία δεν χρησιμοποιήθηκαν.

Όσον αφορά των αριθμό των δημοσιεύσεων που συγκεντρώσαμε και αφορά το επιστημονικό πεδίο - αντικείμενο κάθε εταιρείας, δεν φαίνεται να έχει ιδιαίτερη σημασία για την επιτυχή εκπαίδευση των μοντέλων μας. Αρχικά, γνωρίζουμε πως ο τρόπος με τον οποίο έγινε η αναζήτηση των δημοσιεύσεων εισάγει αρκετό θόρυβο και ενδεχομένως οδηγεί σε υπερεκτίμηση του αριθμού που αναζητούμε. Για να εκτιμήσουμε την ωριμότητα ενός επιστημονικού πεδίου απαιτείται αρχικά να κατανοήσουμε το αντικείμενο ενασχόλησης της κάθε startup σε μεγαλύτερο βάθος από την απλή γνώση του είδους της ασθένειας με την οποία καταπιάνεται ή της ιατρικής λειτουργίας - διαδικασίας που επιχειρεί να βελτιστοποιήσει. Ο τρόπος με τον οποίο κάθε εταιρεία προσεγγίζει το αντικείμενό της, οι μέθοδοι που επιστρατεύει για να επιφέρει επιθυμητά αποτελέσματα, αποτελούν στοιχεία που θα έπρεπε να γνωρίζουμε και να κατανοούμε. Ωστόσο, κάτι τέτοιο θα απαιτούσε ενδελεχή έρευνα για την κάθε επιχείρηση ξεχωριστά και όχι απλώς στα πλαίσια της ανασκόπησης της περιγραφής της από το crunchbase ή ακόμα και από την ίδια την ιστοσελίδα της εταιρείας. Επιπλέον μια ακόμα προϋπόθεση θα ήταν η ύπαρξη των απαραίτητων γνώσεων από τη μεριά μας ώστε να συμπεράνουμε από τις διαθέσιμες πληροφορίες την επιστημονική "γνώση" που θα έπρεπε να υπάρχει ώστε να είναι σε θέση η εταιρεία να εκπληρώσει τους επιχειρησιακούς της στόχους. Πέρα από αυτό σίγουρα μια αναζήτηση στο google scholar όπου δεν έχουμε τη δυνατότητα να φιλτράρουμε τα αποτελέσματα που μας επιστρέφει δεν ανταποκρίνεται απόλυτα στις ανάγκες θα θέλαμε να μας καλύψει. Από την άλλη πλευρά, είναι εύλογο να θεωρήσουμε ότι τα παραπάνω αποτελέσματα είναι ενδεικτικά του γεγονότος ότι για η ίδρυση μιας εταιρείας που καταπιάνεται με την

εισαγωγή μίας καινοτομίας, συνήθως, προηγείται επαρκή έρευνα και μελέτη που διασφαλίζει σε ένα βαθμό το ότι η επιτυχία είναι εφικτή.

4.3.3 Αξιολόγηση αλγορίθμου

Ένα σημαντικό στοιχείο που προέκυψε από αυτή τη μελέτη αφορά την απόδοση του ταξινομητή της λογιστικής παλινδρόμησης η οποία ήταν περισσότερο ικανοποιητική από ότι θα περιμέναμε. Αρχικά είχαμε την εντύπωση ότι το πρόβλημά μας δεν θα μπορούσε να αποτυπωθεί με γραμμικές σχέσεις εξαιτίας της πολυπλοκότητας των διασυνδέσεων που των χαρακτηριστικών μας. Ωστόσο, αυτό φάνηκε να μην αποτελεί μεγάλο εμπόδιο και τα αποτελέσματα που λάβαμε ήταν αρκετά ικανοποιητικά. Θελήσαμε να δούμε αν η γραμμικότητα που εισάγεται στην πρόβλεψη ευνοεί τις εταιρείες που έχουν τιμή 1 στα δυαδικά πεδία και έτσι επιτυγχάνει υψηλά ποσοστά επιτυχίας. Πράγματι αφαιρώντας όλα τα χαρακτηριστικά αυτής της μορφής είχαμε μία πτώση τουλάχιστον 10 μονάδων στην τιμή της accuracy. Ωστόσο, το ίδιο παρατηρήσαμε και στα άλλα δύο μοντέλα οπότε δεν μπορούμε να χρησιμοποιήσουμε τη συγκεκριμένη επιχειρηματολογία ως εξήγηση για την απόδοση του ταξινομητή. Τέλος, αξίζει να αναφέρουμε ότι ο αλγόριθμος της λογιστικής παλινδρόμησης, όντας και ένα λιγότερο προσαρμοστικό μοντέλο, δεν εμφάνισε σε καμία περίπτωση δείγματα υπερεκπαίδευσης.

Ο αλγόριθμος του gradient boosting μας έδωσε και τα καλύτερα αποτελέσματα και οριακά χειρότερη επίδοση φαίνεται να είχαν τα τυχαία δάση. Στην πραγματικότητα, πρόκειται για μία συμπεριφορά που περιμέναμε αφού αξιολογήσαμε το σύνολο δεδομένων που συγκεντρώσαμε. Αρχικά είχαμε να ασχοληθούμε με ένα μικρού μεγέθους dataset, γεγονός που μας απέτρεψε από τη χρήση αλγόριθμων βαθιάς μηχανικής μάθησης, με αρκετά μεγάλο μέγεθος απουσιάζουσων τιμών, με μία μίξη από κατηγορηματικά και αριθμητικά χαρακτηριστικά, ενδεχομένως σχέσεις μεταξύ των υποσυνόλων των σύνολων τιμών ορισμένων χαρακτηριστικών, στοιχεία που θεωρήσαμε πως μπορούσε να περιγράψει αρκετά πιστά μια δενδρική δομή. Το πρόβλημα που είχαμε να αντιμετωπίσουμε και στις δύο περιπτώσεις ήταν το ενδεχόμενο της υπερεκπαίδευσης που αντιληφθήκαμε να παρουσιάζεται σε ορισμένες περιπτώσεις όπου αφήναμε μεγάλη ευελιξία στις παραμέτρους εκπαίδευσης.Πρόκειται για ένα πρόβλημα που γίνεται ιδιαίτερα αισθητό σε τέτοιες μορφές αλγοριθμικών μοντέλων καθώς, αν το επιτρέψουμε, το μοντέλο μας μπορεί να καταλήξει να αναπαριστά με απόλυτη πιστότητα όλο το σύνολο των δεδομένων εισόδου χωρίς όμως να διαθέτει κάποια σημαντική ικανότητα απόφασης για άγνωστα προς αυτό δεδομένα. Στην περίπτωσή μας το φαινόμενο αυτό αποφεύχθηκε με κατάλληλο χειρισμό των παραμέτρων των αλγόριθμων.

4.3.4 Σύγκριση αποτελεσμάτων

Στην παράγραφο αυτή θα αναφέρουμε ορισμένα αποτελέσματα που παρουσιάστηκαν στις περιορισμένες παρόμοιες μελέτες που έχουν γίνει και θα τα συγκρίνουμε με την απόδοση των δικών μας μοντέλων. Στον πίνακα που ακολουθεί μπορούμε να δούμε ορισμένα αποτελέσματα που συγκεντρώσαμε.

Το 2018 ο R.Bento παρουσίασε εκπληκτικά θα έλεγε κανείς αποτελέσματα κάνοντας μια παρόμοια προσπάθεια να μοντελοποιήσει την επιτυχία των startups da Silva Ribeiro Bento (2018).

[2] Παρουσιάστηκαν στη μελέτη αποτελέσματα ανα κατηγορία. Εδω αναγράφονται τα αποτελέσματα που αντιστοιχούν στην κατηγορία biotech

	Accuracy	AUC	method	Recall
S.Bento, 2018	93,2%	93,2%	Random Forests	94,1%
Xiang et al. 2012 [2]	-	88,7%	Baysian Networks	62,2%
Sharlchilev et al. 2018	-	86,4%	WBSSP	62,6%
C.Pan el, 2018	73,3%	79,9%	KNN	74,1%

Πίνακας 4.11: Σύγκριση αποτελεσμάτων

Παρατηρούμε ότι η ακρίβεια των αποτελεσμάτων του είναι αρκετά υψηλότερη από αυτήν που παρουσιάσαμε. Ωστόσο, υπάρχει μια σημαντική διαφορά που τελικά καθιστά τα δύο προβλήματα εντελώς διαφορετικά και δεν είναι άλλη από τη μεταβλητή που προσπαθούμε να προβλέψουμε. Έχοντας αντιστοιχίσει διαφορετικές εταιρείες με το label 1 και διαφορετικές με το label 0 από ότι εμείς το πρόβλημα που αποσκοπεί να εξεταστεί γίνεται εντελώς διαφορετικό με αυτό της παρούσας εργασίας. Θα μπορούσαμε να παρομοιάσουμε τη συσχέτιση των δύο μελετών με εκείνη που θα είχαν δυο μελέτες που, έχοντας ως βάση ένα κοινό dataset με στοιχεία για την υγεία ασθενών, προσπαθούν να απαντήσουν, η πρώτη το ερώτημα αν οι ασθενείς έχουν καρκίνο και η δεύτερη αν οι ασθενείς έχουν διαβήτη. Πέρα από αυτήν τη διαφορά καθώς και την διαφορά στο μέγεθος του dataset εντοπίζουμε και έναν επιπλέον παράγοντα διαφοροποίησης σχετικά με τη χρονικότητα των παρατηρήσεων. Όπως αναφέραμε προσπαθήσαμε να παρατηρήσουμε τις εταιρείες σε βάθος χρόνου και έτσι η πρόβλεψή μας αφορά την κατάσταση των εταιρειών σε βάθος 6 χρόνων και όχι τη στιγμή στην οποία αναφέρονται τα δεδομένα. Ακόμα, είναι γνωστό πως σε οποιοδήποτε πρόβλημα μηχανικής μάθησης το πλήθος των δεδομένων καθορίζει και την ποιότητα του αποτελέσματος. Τα δεδομένα της συγκρινόμενης μελέτης είχαν μέγεθος σχεδόν 72 φορές μεγαλύτερο από το σύνολο που συγκεντρώσαμε εμείς καθώς δεν περιοριζόντουσαν από τις κατηγορίες των εταιρειών ή από λοιπούς περιορισμούς που θέσαμε και αναφέραμε προηγουμένως 3.2.1 . Η συγκεκριμένη μελέτη παρέχει και αποτελέσματα για ορισμένες ξεχωριστές κατηγορίες εταιρειών. Στην περίπτωση των εταιρειών της κατηγορίας healthcare(tech) που φαίνεται να συμπίπτουν αρκετά με το δικό μας χώρο μελέτης και με μέγεθος μόλις 5 φορές μεγαλύτερο από το δικό μας, οι μετρικές είναι χαμηλότερες με ανάκληση στο 87%, εμφανώς χαμηλότερη από τη βέλτιστη που παρουσιάστηκε παραπάνω. Τέλος, αξίζει να αναφέρουμε πως σύμφωνα με τα ευρήματα της παραπάνω μελέτης τα οικονομικά κριτήρια επηρεάζουν σαφώς περισσότερο την έκβαση του αποτελέσματος σημειώνοντας μάλιστα ότι από τα 20 πιο σημαντικά χαρακτηριστικά μόνο 2 δε σχετίζονταν με επενδυτικά μεγέθη.

Τις ίδιες περίπου χαρακτηριστικές διαφορές εντοπίζουμε και στη μελέτη των Xiang et al. (2012) η οποία όμως σε κάποια σημεία βρίσκεται πιο κοντά στη δική μας προσέγγιση. Αρχικά ο προσδιορισμός των επιτυχημένων εταιρειών ορίζεται πιο αυστηρά (όπως αυστηρά ορίζεται σε εμάς αντίστοιχα ο προσδιορισμός των αποτυχημένων) γεγονός που καθιστά δυσκολότερη την ταξινόμηση. Επιπλέον, οι ερευνητές εδώ έχουν επιχειρήσει να αξιοποιήσουν και άρθρα που συγκέντρωσαν για τις εξεταζόμενες εταιρείες. Συγκεκριμένα, συγκέντρωσαν άρθρα για εταιρείες πριν αυτές 'επιτύχουν' και προσπάθησαν να δημιουργήσουν, με έναν αλγόριθμο εξαγωγής θεμάτων, μοτίβα στα κείμενα που να προμηνύουν επιτυχία. Ωστόσο, τα άρθρα που συγκέντρωσαν σε σχέση με το σύνολο των εταιρειών ήταν λίγα και η επίδρασή τους στο αποτέλεσμα δεν ήταν τόσο σημαντική. Επίσης, και σε αυτήν την περίπτωση η μελέτη έχει γίνει με τέτοιο τρόπο ώστε να μην υπάρχει

χρονική συνέχεια μεταξύ της διαμόρφωσης των χαρακτηριστικών και του προβλεπόμενου αποτελέσματος, με το τελευταίο συχνά να προηγείται του πρώτου. Να αναφερθεί ότι στη συγκεκριμένη προσπάθεια χρησιμοποιείται ο Πιθανοτικός αλγόριθμος των δικτύων bayes για την πρόβλεψη της κατάστασης της εταιρείας.

Μία πιο ολοκληρωμένη δουλειά παρουσίασαν οι Sharchilev et al. (2018) , που επικεντρώθηκαν στην πρόβλεψη γύρων χρηματοδότησης ύστερα από ένα αρχικό συμβάν χρηματοδότησης σποράς (seed funding) που απαιτήθηκε να υπάρχει. Στη συγκεκριμένη μελέτη οι ερευνητές χρησιμοποίησαν μια σταθερή ημερομηνία για να κάνουν την πρόβλεψη και χρησιμοποίησαν ως δείγματα snapshots των εταιρειών στο χρονικό διάστημα μεταξύ του seed funding και της ημερομηνίας πρόβλεψης. Με τον τρόπο αυτό ενίσχυσαν το μέγεθος του dataset εξασφαλίζοντας και την αντικειμενικότητά του. Επιπλέον επιχείρησαν να υπολογίσουν τη συνολική παρουσία των εταιρειών στο διαδίκτυο συγκεντρώνοντας το πλήθος των αναφορών για αυτές από διάφορες ιστοσελίδες. Τα χαρακτηριστικά αυτά , όπως και ορισμένα ακόμα 'αραιά' χαρακτηριστικά ανάμεσα στα οποία και οι περιγραφές των εταιρειών, προεπεξεργάζονται από ένα νευρωνικό δίκτυο προτού εισαχθούν σαν είσοδο σε έναν catboost ταξινομητή. Τα αποτελέσματα όπως φαίνεται είναι αρκετά ικανοποιητικά. Ωστόσο, δεν παρουσιάζονται στη δημοσίευση τα αντίστοιχα αποτελέσματα για το train set ώστε να μπορεί να διαπιστώσει κανείς το ενδεχόμενο υπερεκπαίδευσης. Στην περίπτωση αυτή οι μελετητές χρησιμοποίησαν ένα σύνθετο δίκτυο όπου αρχικά εκπαίδευσαν ένα μοντέλο λογιστικής παλινδρόμησης και ένα νευρωνικό δίκτυο χρησιμοποιώντας ορισμένα από τα χαρακτηριστικά τους και στη συνέχεια έκαναν προβλέψεις με τη χρήση του αλγόριθμου catboost όπου έβαλαν σαν είσοδο, πέρα από τα υπόλοιπα χαρακτηριστικά, και τις εξόδους των προηγούμενων μοντέλων.

Τέλος, στην προσπάθειά τους να αναπαράγουν τη μελέτη των Xiang et al. (2012) , οι Pan et al. (2018), δοκίμασαν να κάνουν παρόμοιες προβλέψεις χρησιμοποιώντας την ίδια βάση δεδομένων, χωρίς την ανάλυση των κειμένων και με τη χρήση της μεθόδου των k-κοντινότερων- γειτόνων. Η χρήση της τεχνικής αυτής μας προϊδεάζει για την όχι και τόσο ικανοποιητική μοντελοποίηση του προβλήματος καθώς αναπαράγει τα προβλήματα που εμφανίστηκαν στη δική μας μελέτη κατά τη διαδικασία του Oversampling. Πράγματι, πρόκειται για τα λιγότερο ικανοποιητικά αποτελέσματα.

4.4 Μεθοδολογία ανάλυσης ως baseline για επερχόμενες μελέτες

Μέσα από τη διαδικασία ανάπτυξης μιας μεθοδολογίας για την πραγματοποίηση του παρόντος project εστιάσαμε την προσοχή μας σε ορισμένα σημεία τα οποία θα μπορούσαν να συγκεντρωθούν ώστε να λειτουργήσουν βοηθητικά σε επερχόμενες μελέτες. Το σημαντικότερο ζήτημα που θα θέλαμε να θίξουμε αφορά το χειρισμό των ανισόρροπων κλάσεων. Παρατηρήσαμε ότι οι κοινές βιβλιοθήκες oversampling της Python αδυνατούν να χειριστούν σωστά περιπτώσεις όπου α) διαθέτουμε πολλαπλά δυαδικά χαρακτηριστικά που προκύπτουν από τη διαίρεση ενός κατηγορηματικού χαρακτηριστικού και β) το dataset μας περιέχει πολλαπλές διασυνδέσεις μεταξύ χαρακτηριστικών οι οποίες πρέπει να διατηρηθούν στα νέα δείγματα. Στις περιπτώσεις που η χρήση κάποιου αλγόριθμου παλινδρόμησης 'ταιριάζει' στο πρόβλημά μας, το oversampling δε θα επιφέρει σοβαρά αρνητικά αποτελέσματα γιατί η ίδια η μέθοδος χειρίζεται τα χαρακτηριστικά ως ανεξάρτητα. Πιο σύνθετοι όμως αλγόριθμοι όπως αλγόριθμοι που βασίζονται σε τυχαία δένδρα η ακόμα και βαθιά νευρωνικά δίκτυα θα επηρεαστούν σημαντικά.

Επιπλέον, ένα ακόμα σημαντικό στοιχείο είναι η επιλογή κατάλληλων διαστημάτων χωρισμού των αριθμητικών χαρακτηριστικών. Συγκεκριμένα, είναι γνωστό ότι αλγόριθμοι παλινδρόμησης αδυνατούν να χειριστούν χαρακτηριστικά που το σύνολο τιμών τους είναι διευρυμένο και εμφανίζει σοβαρές διακυμάνσεις. Για το λόγο αυτό συνηθίζεται είτε να προηγείται μια κανονικοποίηση των χαρακτηριστικών, είτε να γίνεται διάσπαση του χαρακτηριστικού σε περισσότερα δυαδικά κάθε ένα από τα οποία αντιστοιχεί σε ένα μικρό σύνολο τιμών. Ωστόσο, τόσο στο τυχαία δάση όσο και στο boosting δένδρων αποφάσεων παρατηρήσαμε ότι η τεχνική αυτή δεν αποτελεί πάντα τη βέλτιστη επιλογή. Το να δώσουμε στον αλγόριθμο τη δυνατότητα να επιλέξει ανάλογα με τη συνθήκη του κόμβου του προηγούμενου επιπέδου, την τιμή κατωφλίου για να διαχωρίσει τα δείγματα βάση του εν λόγω χαρακτηριστικού, προσφέρει μια ευελιξία που έχει συνήθως θετικό αντίκτυπο στα αποτελέσματα.

Παράλληλα, βρήκαμε χρήσιμη την τεχνική δημιουργίας επιπλέον δυαδικών πεδίων για να δηλώσουμε την απουσία τιμών χαρακτηριστικών από το σύνολο δεδομένων μας. Όπως αναφέραμε, σε ορισμένα πεδία είχαμε πολύ υψηλά ποσοστά απουσιάζουσων τιμών. Όταν τα πεδία αυτά αφορούσαν αριθμητικά χαρακτηριστικά με μεγάλα σύνολα τιμών όπως για παράδειγμα η συνολική χρηματοδότηση, τα επιπλέον μηδενικά που εισάγονταν από τις τιμές null σίγουρα θα έβλαπταν την απόδοση του αλγόριθμου. Μία μέθοδος που χρησιμοποιείται συχνά για την αντιμετώπιση αυτού του προβλήματος είναι η αντικατάσταση των τιμών αυτών με έναν μέσο όρο που υπολογίζεται για το συγκεκριμένο πεδίο. Ωστόσο θεωρήσαμε ότι η πρακτική αυτή θα μπορούσε να αλλοιώσει τη μορφή των δεδομένων μας. Σε αυτή λοιπόν την περίπτωση παρατηρήσαμε ότι η προσθήκη ενός επιπλέον δυαδικού χαρακτηριστικού, ενδεικτικού της ύπαρξης η απουσίας πληροφορίας για το συγκεκριμένο χαρακτηριστικό παρουσίασε τα βέλτιστα αποτελέσματα στην έρευνά μας.

Ένα τελευταίο ζήτημα που θα θέλαμε να θίξουμε αφορά τις μελέτες που ασχολούνται με προβλέψεις σε βάθος χρόνου. Η ορθότητα των αποτελεσμάτων έγκειται, στις περιπτώσεις αυτές, στην σωστή κατανόηση και τον συνυπολογισμό της χρονικής συνέχειας των γεγονότων και τον καθαρισμό των δεδομένων ώστε όλα τα χαρακτηριστικά να αφορούν χρονικές στιγμές προ της πρόβλεψης.

Κεφάλαιο 5

Επίλογος

5.1 Σύνοψη

Στην παρούσα διπλωματική μελετήθηκε το πρόβλημα της πρόβλεψης της επιτυχίας μιας νέας εταιρείας. Πρόκειται για ένα ζήτημα υψηλής σημασίας για όσους δραστηριοποιούνται καθημερινά στο χώρο της οικονομίας και των επενδύσεων το οποίο επηρεάζει έμμεσα και το υπόλοιπο κομμάτι του πληθυσμού. Επικεντρώσαμε το ενδιαφέρον μας στις εταιρείες που δραστηριοποιούνται στον τομέα της υγείας - βιοτεχνολογίας, ο οποίος τα τελευταία χρόνια έχει γνωρίσει σημαντική ανάπτυξη.

Συλλέξαμε πληροφορίες που βρήκαμε στο διαδύκτιο με βασική πηγή δεδομένων την ιστοσελίδα crunchbase.com που μας έδωσε πρόσβαση στη βάση δεδομένων της. Σε πολλές περιπτώσεις αναγκαστήκαμε να ελέγξουμε την εγκυρότητά των δεδομένων και να κάνουμε τις δέουσες αλλαγές όπου αυτό κρίθηκε αναγκαίο. Η γενικότερη εποπτεία και βαθύτερη κατανόηση του dataset που χρησιμοποιήσαμε έπαιξε καταλυτικό ρόλο στον τρόπο επεξεργασίας και χρήσης του και για άλλη μια φορά αποδείχτηκε ότι αποτελεί βασική προϋπόθεση για την επιτυχία μιας λύσης μηχανικής μάθησης.

Άμεση συνέπεια της κατανόησης του συνόλου δεδομένων αποτελεί και η επιλογή του αλγόριθμου εκπαίδευσης. Υποστηρίξαμε και επιβεβαιώσαμε με τα αποτελέσματά μας ότι πολυπαραγοντικά προβλήματα, με χαρακτηριστικά τα οποία φαίνεται να μη συσχετίζονται εκ πρώτης όψεως αλλά πιθανόν φέρουν βαθύτερες διασυνδέσεις, και για τα οποία διαθέτουμε ένα αρκετά αραιό σύνολο δεδομένων, μοντελοποιούνται ικανοποιητικά από αλγόριθμους που βασίζονται σε δένδρα αποφάσεων.

Η χρήση gradient boosting σε δένδρα αποφάσεων μας έδωσε και τα καλύτερα αποτελέσματα με ποσοστά επιτυχίας που περιγράφονται στην ενότητα 4.2.3 . Για να προκύψουν τα συγκεκριμένα αποτελέσματα έπρεπε να βρούμε έναν αποτελεσματικό τρόπο να διαχειριστούμε την μεγάλη ανισορροπία των κλάσεων σε συνδυασμό με το μικρό, για προβλήματα μηχανικής μάθησης συνδυασμό του dataset. Αποδείχθηκε ότι οι συνήθεις τεχνικές υπερδειγματοληψίας μπορούν να αλλοιώσουν σημαντικά τα χαρακτηριστικά των δεδομένων μας υπονομεύοντας την ποιότητα του εκπαιδευτή μας. Ακολουθήσαμε λοιπόν, για να πετύχουμε τα παραπάνω αποτελέσματα, μία διαδικασία δημιουργίας εικονικών δειγμάτων για να χρησιμοποιήσουμε ώστε να εξισορροπήσουμε το σύνολο δεδομένων μας.

Κατά το πειραματικό μέρος, δοκιμάσαμε να χωρίσουμε αριθμητικά χαρακτηριστικά σε clusters καθένα από τα οποία παίρνει μια τιμή από τις 0,1 και αντιπροσωπεύει ένα σύνολο τιμών. Παρατηρήσαμε ότι αυτή η πρακτική περιορίζει την ευελιξία στα μοντέλα που βασίζονται σε δένδρα

αποφάσεων ενώ σε αλγόριθμους παλινδρόμησης, όπως στη λογιστική παλινδρόμηση στην περίπτωσή μας, βοηθάει το μοντέλο στη αποδοτική μεταχείριση χαρακτηριστικών με πολύ μεγάλα σύνολα τιμών.

Τα αποτελέσματα όλων των προσπαθειών μας υπόδειξαν την ύπαρξη πληροφοριών χρηματοδότησης ως το σημαντικότερο χαρακτηριστικό ενώ στη συνέχεια ερχόντουσαν τα πεδία που αποθηκεύονται πληροφορίες για τα χρόνια λειτουργίας καθώς και για το χρονικό διάστημα μεταξύ της πρώτης χρηματοδότησης και της ίδρυσης της εταιρείας. Εξετάζοντας τις εταιρείες που είναι αποτυχημένες και παρόλα αυτά προβλέφυηκαν ως επιτυχείς αντιλαμβανόμαστε την επίδραση που έχουν τα παραπάνω χαρακτηριστικά στον ταξινομητή μας καθώς, οι περισσότερες από αυτές συγκέντρωναν πολύ νωρίς σημαντική χρηματοδότηση μέσω κεφαλαίων επιχειρηματικ΄ψν συμμετοχών. Τα ευρήματα αυτά συνάδουν με την εμπειρική παρατηρήσεις των ειδικών που υποστηρίζουν ότι εταιρείες με πολύ υψηλή χρηματοδότηση επιχειρηματικών κεφαλαίων αποτυγχάνουν συχνότερα από τις υπόλοιπές, αν όμως επιτύχουν τα κέρδη των εμπλεκομένων επενδυτών είναι πολύ υψηλά. Οι προβλέψεις που έγιναν συνολικά, ήταν άκρως ικανοποιητικές.

Η σύνδεση της επιτυχίας των εταιρειών με την ύπαρξη 'γνώσης' γύρω από το αντικείμενο ενασχόλησής τους φάνηκε ασθενής. Ωστόσο, πιστεύουμε ότι το κομμάτι αυτό θα μπορούσε να μελετηθεί περαιτέρω, κυρίως για τις ερευνητικού χαρακτήρα εταιρείες, με συγκέντρωση πληροφοριών από επίσημες ιατρικές ιστοσελίδες και κατάλληλη επεξεργασία τους.

Ολοκληρώνοντας, η μελέτη της επιτυχίας των εταιρειών είναι ιδιαίτερα απαιτητική καθώς επηρεάζεται από αστάθμητους παράγοντες που προκύπτουν από την άρρηκτη σύνδεσή της με την κοινωνία και τους ταχύτατους ρυθμούς εξέλιξής της. Ωστόσο, η παραπάνω μελέτη παρέχει μια μεθοδολογία επεξεργασίας και αξιοποίησης λιγοστών, σε σχέση με τις συνολικές, μετρικών ώστε να προκύψει ένα ικανοποιητικό αποτέλεσμα με όσο το δυνατόν περισσότερο έγκυρα αποτελέσματα.

5.2 Μελλοντικές Κατευθύνσεις Επιστημονικής Μελέτης

Η μελέτη και πρόβλεψη της επιτυχίας οργανισμών, εταιρειών, φορέων αποτελεί ένα πεδίο που έχει απασχολήσει πολλούς ερευνητές αλλά δεν έχει τεκμηριωθεί επιστημονικά από ικανοποιητικό αριθμό μελετών. Καθώς η έρευνα προς αυτήν την κατεύθυνση φαίνεται μόνο να εντατικοποιείται τα επόμενα χρόνια μπορούμε να συγκεντρώσουμε ορισμένες προτάσεις που θεωρούμε πως θα βοηθήσουν την αντίστοιχη επιστημονική κοινότητα.

Αρχικά κρίνουμε πως μια ολοκληρωμένη έρευνα θα απαιτούσε περαιτέρω έρευνα για συγκέντρωση δεδομένων ώστε να υπάρχει μεγαλύτερη ποικιλία και αξιοπιστία. Τα δεδομένα αυτά θα μπορούσαν να κυμανθούν από τη συνολική παρουσία της εταιρείας στο διαδίκτυο, τις σχέσεις, το χαρακτήρα, την εμπειρία των εργαζομένων εώς την ανάλυση συναισθημάτων του κοινού απέναντι στην ίδια την εταιρεία ή στην υπηρεσία -προϊόν που αυτή υπόσχεται, όπως προκύπτουν από τα μέσα κοινωνικής δικτύωσης. Παράλληλα ενδιαφέρον θα ήταν και η διαχείριση ενός συνόλου δεδομένων αρκετά μεγαλύτερου όγκου ώστε να χρησιμοποιηθούν και αξιολογηθούν και μέθοδοι βαθιάς μηχανικής μάθησης που θα έδιναν άλλη ευελιξία στην μοντελοποίηση του προβλήματος.

Σε ότι αφορά τη σύνδεση της επιτυχίας με την ωριμότητα του εκάστοτε επιστημονικού πεδίου, θα μπορούσαν να αξιοποιηθούν συνεντεύξεις, απόψεις και συμπεράσματα ειδημόνων του χώρου ώστε να κριθεί σε τι ποσοστό το εγχείρημα της εταιρείας είναι εφικτό ή παράτολμο κοκ. Ενδιαφέρον θα είχε κάτι τέτοιο να δοκιμαστεί σε διάφορα επιστημονικά πεδία.

Όσον αφορά τη χρονική σκοπιά της μελέτης, θεωρούμε πως η βέλτιστη πρακτική θα ήταν η 'ζωντανή' παρακολούθηση εταιρειών σε ένα χρονικό διάστημα στο τέλος του οποίου θα αναφέρεται και το αποτέλεσμα της πρόβλεψης. Σαφώς κάτι τέτοιο φαίνεται αρκετά χρονοβόρο και ασύμφορο όμως διασφαλίζεται έτσι η εγκυρότητα των δεδομένων. Σε αντίθετη περίπτωση πρέπει να γίνεται σαφής καθορισμός της ημερομηνίας που αφορούν τα συλλεχθέντα δεδομένα και αυτής που αφορά η πρόβλεψη του αλγόριθμου.

Μερικές επίσης κατευθύνσεις που θα μπορούσαμε να δώσουμε στο εν λόγω αντικείμενο είναι η δημιουργία ενός είδους ταξινόμησης των εταιρειών. Συχνή είναι η ίδρυση εταιρειών με κοινό ενδιαφέρον και στόχους την ίδια περίοδο. Σαφώς αν απευθύνονται στη ίδια αγορά την ίδια χρονική στιγμή είναι αρκετά παράδοξο να επιτύχουν όλες. Μια ταξινόμηση των εταιρειών με κριτήριο το ποσοστό επιτυχίας θα ήταν ιδιαίτερα ενδιαφέρουσα. Παράλληλα, εφόσον γνωρίζουμε ότι η επιτυχία είναι μια έννοια που δεν ορίζεται με σαφήνεια, θα μπορούσε επίσης να γίνει προσπάθεια πρόβλεψης άλλων μεγεθών όπως γύρων χρηματοδότησης, συμβάντων IPO, χρόνων επιβίωσης, μελλοντικού αριθμού εργαζομένων.

Τέλος, παρατηρώντας ότι πολλές από τις κλειστές εταιρείες εμφάνιζαν λόγω των χαρακτηριστικών τους αρκετές προοπτικές, θεωρούμε πως μια ακόμα κατεύθυνση της έρευνας θα μπορούσε να επικεντρώνεται στα βήματα που απαιτούνται από τη μεριά μιας αποτυχημένης εταιρείας για να οδηγηθεί στην επιτυχία. Με άλλα λόγια ποια θα ήταν η ελάχιστη αλλαγή στα δεδομένα εισόδου μας ώστε να ταξινομηθούν οι εταιρείες που ορθά έχουν ταξινομηθεί ως κλειστές, στην αντίθετη κλάση αυτή των επιτυχημένων. Η χρήση των δένδρων αποφάσεων θα βοηθούσε ιδιαίτερα στην απάντηση ενός τέτοιου ερωτήματος, στο οποίο η απάντηση θα ήταν ιδιαίτερα χρήσιμη αυτή τη φορά κυρίως στους εργαζομένους και ιδρυτές των εταιρειών και ύστερα στους επενδυτές.

Βιβλιογραφία

Will Badr. Why feature correlation matters...a lot! *https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4*, 2018.

Richard Bellman. *An Introduction to Artificial Intelligence : Can Computers Think?* Boyd amp; Fraser Pub. Co., 1978.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Stephen J. Brown, William Goetzmann, Roger G. Ibbotson, and Stephen A. Ross. Survivorship Bias in Performance Studies. *The Review of Financial Studies*, 5(4):553–580, 05 2015. ISSN 0893-9454. doi: 10.1093/rfs/5.4.553. URL https://doi.org/10.1093/rfs/5.4.553.

Camilla Calì and Maria Longobardi. Some mathematical properties of the roc curve and their applications. *Ricerche di Matematica*, 64(2):391–402, 2015. doi: 10.1007/s11587-015-0246-8. Dedicated to the Memory of Carlo Ciliberto.

F. R. da Silva Ribeiro Bento. Predicting start-up success with machine learning. 2018.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, Mar. 1996. doi: 10.1609/aimag.v17i3.1230. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230.

Jerome Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 11 2000. doi: 10.1214/aos/1013203451.

Tin Kam Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, page 278, USA, 1995. IEEE Computer Society. ISBN 0818671289.

Urvashi Jaitley. Why data normalization is necessary for machine learning models. *https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029*, 2018.

J.B.Rollins. Foundational methodology for data science. *IBM Analytics*, 2015.

Michael Shannon Kearns. Thoughts on hypothesis boosting. 1988.

David Masip Laura Calvet, Jésica de Armas and Angel A. Juan. Learnheuristics: hybridizing metaheuristics with machine learning for optimization with dynamic inputs. *Open Mathematics*, (15):261–280, 2017.

Daniel Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 11 1999. doi: 10.1038/44565.

Bai Lo. Useful properties of roc curves and auc scoring. *https://www.kaggle.com/learn-forum/53782*, 2018.

E.G.Alexandrakis M.G.Lioudakis. Emotional analysis in greek text using machine learning algorithms. 2017.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2.

P.L.Juell Nagesh Kadaba, K.E.Nygard. Integration of adaptive machine learning and knowledge-based systems for routing and scheduling applications. *Expert Systems with Applications*, 2(1): 15–27, 1991.

N.Christofides, A.Mingozzi, and P.Toth. State-space relaxation procedures for the computation of bounds to routing problems. *Networks*, 11(2):145–164, 1981.

Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994. doi: 10.1002/env.3170050203. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.3170050203.

C. Pan, Y. Gao, and Y. Luo. Machine learning prediction of companies' business success. 2018. URL http://cs229.stanford.edu/proj2018/report/88.pdf.

Joanne Peng, Kuk Lee, and Gary Ingersoll. An introduction to logistic regression analysis and reporting. *Journal of Educational Research - J EDUC RES*, 96:3–14, 09 2002. doi: 10.1080/00220670209598786.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Eric. Ries. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* New York: Crown Business, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. Cambridge University Press, 2014. URL http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

Boris Sharchilev, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten Rijke. Web-based startup success prediction. pages 2283–2291, 10 2018. doi: 10.1145/3269206.3272011.

Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 05 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr300. URL https://doi.org/10.1093/bioinformatics/btr300.

I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, 3rd.* Morgan Kaufmann, USA, 2011.

Guang Xiang, Zeyu Zheng, Miaomiao Wen, Jason I. Hong, Carolyn Penstein Rosé, and Chao Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In John G. Breslin, Nicole B. Ellison, James G. Shanahan, and Zeynep Tufekci, editors, *ICWSM*. The AAAI Press, 2012. URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2012.html#XiangZWHRL12.