Πανεπιστημιο Μακεδονιας

Τμημα εφαρμοσμενης πληροφορικης

Εργασία πάνω στην επιβλεπόμενη μάθηση και την ταξινόμηση

Κωνσταντίνος Πασβάντης

10 Νοεμβρίου 2022

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

Δώθεντός ένα αρχείο με δεδομένα από διάφορες εταιρείες ανά τα έτη, καλούμαι να αντιμετοπίσω ένα απλό, αλλά ταυτόχρονα απαιτητικό πρόβλημα: την δημιουργία μοντέλων ταξινόμησης με την βοήθεια των διαφόρων τεχνικών της μηχανικής μάθησης.
Τα δεδομένα περιγράφουν τις διαφορετικές καταστάσεις στις οποίες βρισκόταν η κάθε εταιρεία και στο τέλος κάθε παρατήρησης, δίνεται και η τελική κατάσταση της, δηλαδή, αν τελικά είναι υγιής ή αν έχει κυρήξει χρεωκοπία.
Με αυτά τα δεδομένα λοιπόν, πρέπει να συνταχθεί κώδικας ο οποίος θα διαβάζει τα δεδομένα από το αρχείο,με τα οποία θα εκπαιδεύονται τα διαφορετικά μοντέλα μηχανικής μάθησης, και θα προβλέπει τις εταιρείες εκείνες οι οποίες θα κυρήξουν πτώχευση.
Οι απαιττήσεις του προβλήαμτος είναι αυτές που το καθιστόυν απαιττητικο: τα μοντέλα ταξινόμησης πρέπει να προβλέπουν με ποσοστό επιτυχίας 62%τις εταιρείες οι οποίες θα πτωχεύσουν και με 70% τις εταιρείες οι οποίες είναι υγιείς.
Στην εργασία αυτή θα παρουσιαστούν τα μοντέλα επιβλεπόμενης μάθησης που χρησιμοποίηθηκαν με σκοπό την λύση του παραπάνω προβλήματος και θα συγκριθούν μεταξύ τους ως προς το ποιό μοντέλο έδωσε τα καλύτερα αποτελέσματα.
Η εργασία είναι χωρισμένη σε δύο μέρη. Στο πρώτο μέρος τα μοντέλα ταξινόμησης εκπαιδεύονται χωρίς να γίνει κάποια τροποποίηση ως προς τον αριθμό των δειγμάτων, ενώ στο δεύτερο μέρος, το train set δημιουργείται με τέτοιο τρόπο ώστε η αναλογια υγειών προς μη υγειών εταιρειών να είναι 3:1.
Πριν να ξεκινήσει το κύριο μέρος της εργασίας στο οποίο παραθέτονται οι αλγόριθμοι, οι μόνες τροποποιήσεις που κάνουμε στα δεδομένα για τώρα θα είναι να αφαιρεθεί η στήλη με την χρονία της κάθε παρατήρησης, αφού δεν μας παρέχει κάποια συγκεκριμένη πληροφορία με την οποία θα μπορούσαν να εκπαιδευτούν τα μοντέλα, και να γίνει *scale* ώστε οι τιμές των μεταβλητών να κυμαίνονται από 0 έως 1.
Επίσης θεωρείται δεδομένο ότι ο σκοπός του κάθε μοντέλου είναι να κατατάξει τις εξαρτημένες μεταβλητές του προβλήματος, στην συγκεκριμένη περίπτωση το αν μία εταιρεία χρεωκοπήσει ή όχι, με την βοήθεια των ανεξάρτητων μεταβλητών, οι οποίες θα είναι οι μεταβλητές των $Xtrain$. Με τον όρο μεταβλητές θα εννοούμε κάθε στήλη που υπάρχουν στα δεδομένα μας, ενώ κάθε εταιρεία θα αποτελέι και μια εγγραφη.
Επίσης το σύνολο των δεδομένων που έχουμε χωρίσει το αρχικό *dataset* για να κάνουμε τις προβλέψεις αποτελείται από 56 μη υγιείς και 2088 υγιείς επιχειρήσεις. Οπότε αφού θέλουμε ποσοστό επιτυχίας 62% για τις μη υγιείς, πρέπει να προβλεφθούν σωστά τουλάχιστον 35 από τις 56, και με παρόμοιο σκεπτικό, τουλάχιστον 1462 από τις 2088 υγιείς. Οπότε για να βρεθεί το ποσοστό πρόβλεψης υγιεών και μη υγειών επιχειρήσεων σε κάθε μοντέλο, αρκεί να διαιρέσουμε τον αριθμό ΤΠ με το 56 και τον αριθμό TN με το 2088 αντίστοιχα.

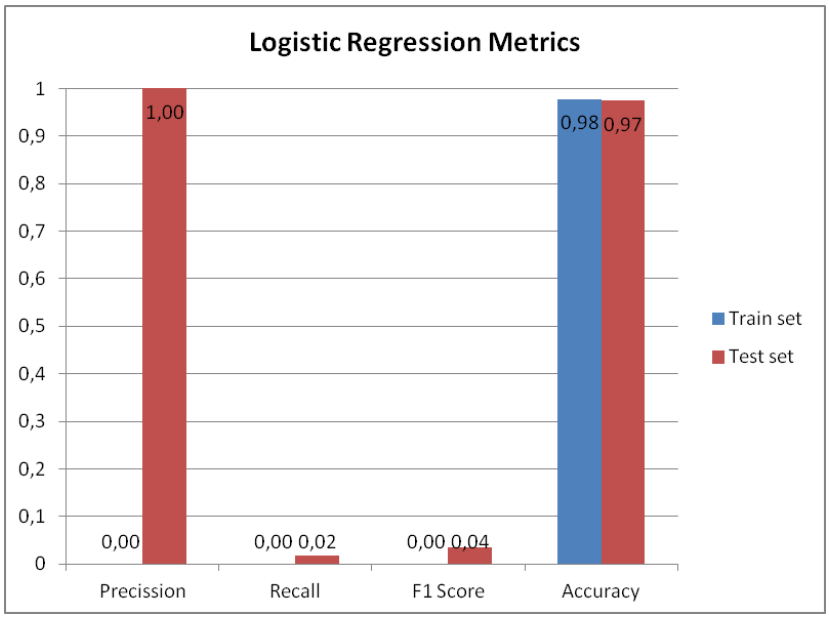
2 Μερος Α΄

2.1 Logistic Regression

Το μοντέλο αυτό ψάχνει τις καλύτερες δυνατές παραμέτρους ώστε να δημιουργήσει μια σιγμοειδή συνάρτηση της οποίας το σύνολο τιμών ανήκει στο $(0, 1)$. Εφόσον τα πιθανά αποτελέσματα που ζητάμε είναι δυο, είτε να πτωχεύσει είτε να μην πτωχεύσει μαι εταιρεία, αν συμβολίσουμε με 1 την κατάσταση στην οποία μια εταιρεία πτωχεύει και με 0 την κατάσταση στην οποία η εταιρεία είναι υγιείς, η σιγμοειδής συνάρτηση που αναφέρθηκε νωρίτερα μπορεί να φανεί ιδιαίτερα αποτελεσματική.

Στο training phase, το μοντέλο, έχοντας σαν δεδομένα όλες τις ζητούμενες μεταβλητές μαζί με τα αποτελέσματά τους,αρχικά προσαρμόζει μία τυχαία συνάρτηση της προηγούμενης μορφής σε αυτά, και στην συνέχεια ελέγχει αν υπάρχει κάποια άλλη συνάρτηση η οποία εξυπηρετεί τον σκοπό αυτό καλύτερα. Τελικά, αφού δημιουργήσει την συνάρτηση η οποία ανταποκρίνεται πιο ικανοποιητικά στα δεδομένα, εξετάζει ποια είναι η πιθανότητα μια παρατήρηση με κάποιες τιμές μεταβλητών να είναι κλάσης 1 ή 0. Αν η πιθανότητα αυτή είναι πάνω από 50%, τότε η προβλεπόμενη κλάση για αυτή την παρατήρηση είναι η 1, ενώ αν είναι κάτω από 50%, είναι η 0.

Τα αποτελέσματα από αυτό το μοντέλο δεν κατάφεραν να ικανοποιήσουν κανέναν από τους δύο περιορισμούς που μας ζητήθηκαν, κυρίως λόγω της αδυναμίας του μοντέλου να προσδιορίσει ποιές εταιρείες θα χρεοκωπήσουν.



Σχήμα 1: Μετρικές για το Logistic Regression

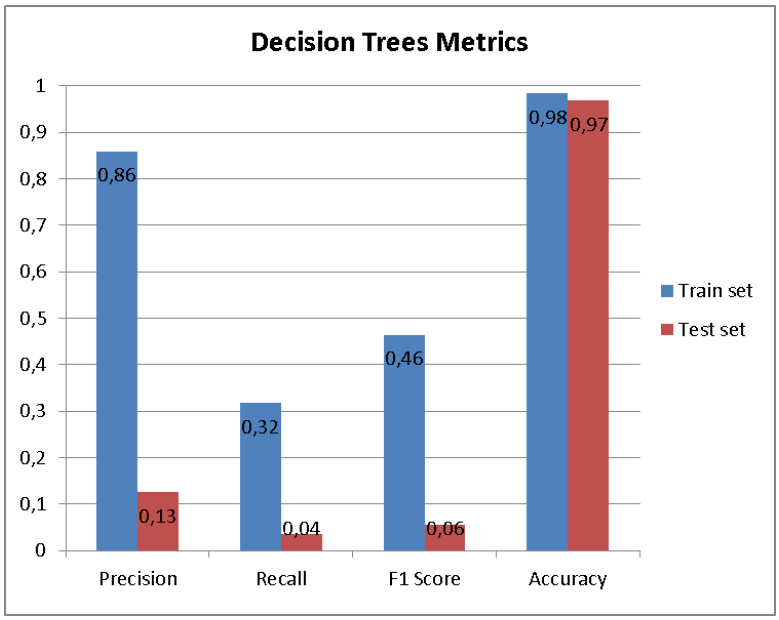
Πίνακας 1: Προβλέψεις για Linear Regression

	TP	TN	FP	FN
Train	0	8380	0	192
Test	1	2088	0	55

Όπως μπορούμε να δούμε η τιμή για το accuracy είναι πολύ ψηλά, κυρίως λόγω του μεγάλου αριθμού υγειών επειηρήσεων που προβλέπει το μοντέλο. Εφόσον επίσης βγήκε σαν αποτέλεσμα πρόβλεψης να χρεοκοπήσει μόνο μία εταιρεία, η οποία τελικά ήταν όντως χρεοκωπημένη, εξηγεί και το γεγονός ότι το precission στο test set είναι ίσο με την μονάδα.

2.2 Decision Tree Classifier

Σκοπός των δέντρων αποφάσεων είναι να κατηγοροποιήσουν τα δεδομένα κάνοντας προβλέψεις, οι οποίες είναι βασισμένες πάνω σε απαντήσεις ερωτήσεων που αφορούσαν τα δεδομένα τα οποία εκπαιδεύτηκαν. Πιο συγκεκριμένα, τα δέντρα αποφάσεων δημιουργουργούν μία ρίζα από την οποία προσπαθούν να χωρίσουν τα δεδομένα με τον καλύτερο δυνατό τρόπο. Ο τρόπος που το κάνουν αυτό είναι να δημιουργήσουν λογικές ερωτήσεις ωστέ να καταλήξουν σε μία απόφαση. Από την ρίζα δημιουργούνται κόμβοι, οι οποίοι με την σειρά τους δημιουργούν επιπλέον κόμβους που απεικονίζουν το τί απόφαση πρέπει να παρθεί για να συνεχίσουν σε κατώτερο επίπεδο. Στο κατώτατο επίπεδο βρίσκονται τα φύλλα κάθε δέντρου, τα οποία αντιπροσωπεύοουν τις πιθανές κατηγορίες αποτελεσμάτων (στην συγκερκιμένη περίπτωση χρεοκωπημένη και μη χρεωκοπημένη).



Σχήμα 2: Μετρικές για τα δέντρα αποφάεων

Όπως μπορούμε να δούμε, ο ταξινομητής αυτός δίνει καλύτερα αποτελέσματα σε σχέση με το Logistic Regression , αλλά πάλι δεν ικανοποιει τις απαιτήσεις του προβλήματός μας. Επίσης ξανασχολιάζουμε, το γεγονός ότι το accuracy στο test set είναι πάλι ψηλά δεν σημαίνει ότι το μοντέλο βγάζει σωστά αποτελέσματα. Για να το καταλάβουμε καλυτερα αυτό, ο τύπος που υπολογίζει το accuracy είναι :

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

. Όμως εδώ εφόσον το μοντέλο δεν προβλέπει αρκετά TP, TN και FP σε σχέση με τον αριθμό των TP (όπως φαίνεταται και από τον παρακάτω πίνακα), λογικό είναι η τιμή του κλάσματος να είναι υψηλή.

Πίνακας 2: Προβλέψεις για Decision Trees

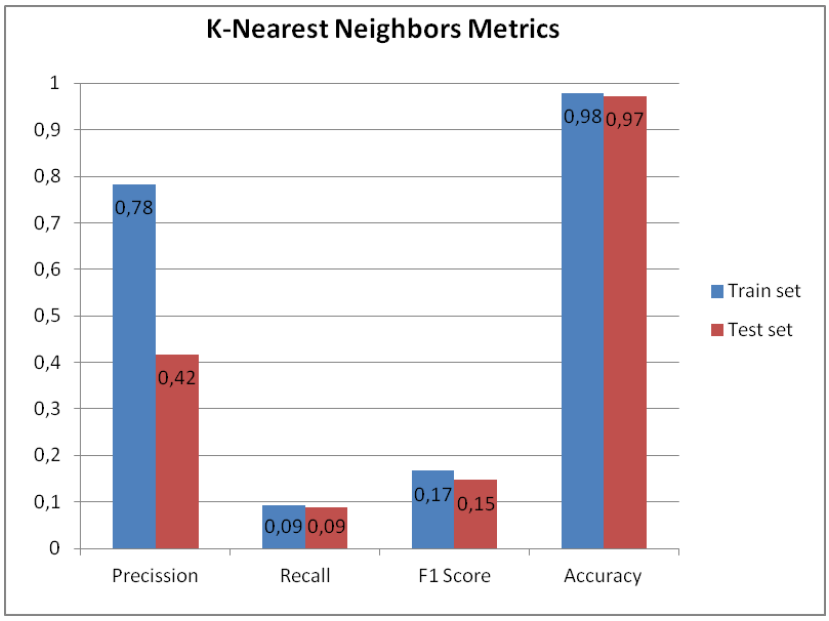
	TP	TN	FP	FN
Train	61	8370	10	131
Test	2	2074	14	54

2.3 k-Nearest Neighbors

Η βασική αρχή και το σκεπτικό αυτού του μοντέλου είναι ότι μία εγγραφή η οποία βρίκεται κοντά σε κάποια άλλη με συγκεκριμένη κλάση, θα έχει την ίδια κλάση μαζί της.

Ο αλγόριθμος των κ πλησιέστερων γειτόνων (στην περίπτωση μας 5 πλησίεστερων γειτόνων εφόσον δεν αλλάξαμε την παράμετρο κ), για να προβλέψει την κλάση μίας εγγραφής λειτουργεί ως εξής:
βρίσκει όλες τις αποστάσεις αυτής της εγγραφής από τις υπόλοιπες με τις οποίες έχουμε εκπαιδεύσει το μοντέλο. Συνήθως επιλέγεται η ευκλείδια. Αφού υπολογίσει όλες τις αποστάσεις, βρίσει τις 5 εγγραφές οι οποίες έχουν την μικρότερη απόσταση από την ζητούμενη. Η κλάση που εμφανίζεται τις περισσότερες φορές

από αυτές τις 5 εγγραφές, είναι ουσιαστικά και η κλάση την οποία προβλέπει το μοντέλο. Τέλος, τοποθετεί την εγγραφή την οποία προέβλεψε σε αυτή την κλάση και με παρόμοιο σκεπτικό συνεχίζει ο αλγόριθμος.



Σχήμα 3: Μετρικές για τους κ πλησιέστερους γείτονες

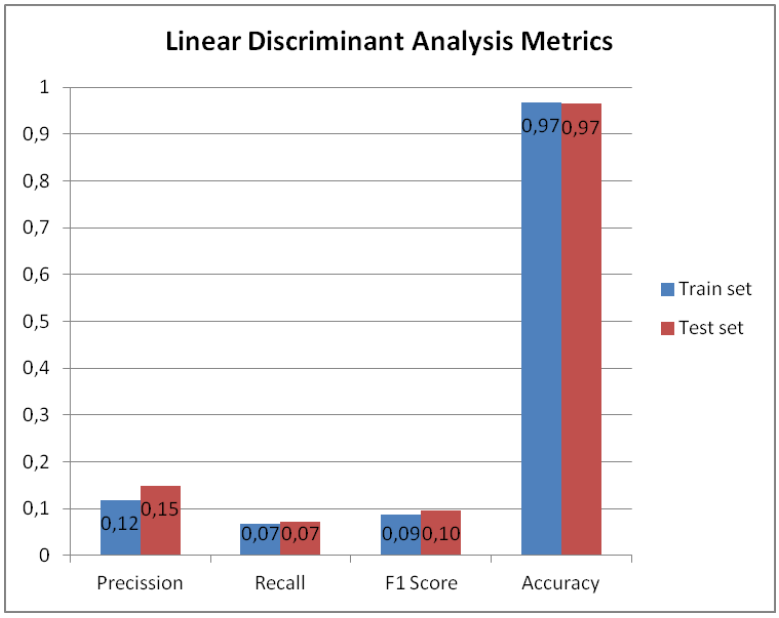
Πίνακας 3: Προβλέψεις για κ πλησιέστερους γείτονες

	TP	TN	FP	FN
Train	18	8375	5	174
Test	5	2081	7	51

Εδώ το μοντέλο παρόλο που μπόρεσε να βρει 5 από τις 56 μη υγιείς επιχειρήσεις στο *testset*, ενώ ο προηγούμενος αλγόριθμος μπόρεσε να βρει 7, βλέπουμε ότι τα στατιστικά των μετρικών είναι κατα πλειοψηφία καλύτερα όσον αφορά των αλγόριθμο των Decision Trees . Σε αυτό βοήθησε ότι ο αριθμός των TN αυξήθηκε, ενώ αυτός των FP μειώθηκε σημαντικά.

2.4 Linear Descriminant Analysis

Η συνοπτική περιγραφή αυτής της μεθόδου μπορεί να αποδοθεί στον αγγλικό όρο "dimensionality reduction".
Ψάχνει μια διαχωριστική γραμμή που χωρίζει τα δεδομένα με τον καλύτερο δυνατό τρόπο, αφού τα προβάλλει πανω σε αυτή τη γραμμή. Αυτό γίνεται με την ταυτόχρονη ικανοποίηση δύο συνθηκών: την μεγιστοποίηση απόστασης μεταξύ των μέσων κάθε κλάσης και την ελαχιστοποίηση της διασποράς που έχει η κάθε κλάση. Σε πολυδιάστατα προβλήματα όπως το δικό μας, ο όρος μέσος και διασπορά προκύπτουν από τις παραμέτρους της πολυδιάστατης κανονικής κατανομής. Εφόσον βρεθεί η διαχωριστική γραμμή, οι προβλέψεις γίνονται εκτιμώντας την πιθανότητα ένα καινούργιο σύνολο παρατηρήσεων να ανήκει σε κάποια κλάση. Αφού η διαχωριστική γραμμή δημιουργήθηκε με τέτοιο τρόπο ώστε να ικανοποιήσει τις δύο συνθήκες,το μοντέλο όταν βλέπει μια καινούργια εγγραφή, μπορεί να την τοποθετήσει στο σημείο της γραμμής στο οποίο έχει την μεγαλύτερη πιθανότητα να ανήκει, λόγω των τιμών των μεταβλητών της.



Σχήμα 4: Μετρικές για Linear Descriminant Analysis

Μπορούμε να δούμε ότι σε σύγκριση με τα προηγούμενα μοντέλα, δεν αποδίδει και τόσο καλα, εκτός βέβαια από το μοντέλο Logistic Regression.Επίσης παρατηρούμε και τον μεγάλο αριθμό FP και FN στο train set,πράγμα που δικαιολογεί γιατί το precission του train set είναι τόσο χαμηλό.

Πίνακας 4: Προβλέψεις για Linear Descriminant Analysis

	TP	TN	FP	FN
Train	13	8283	97	179
Test	4	2065	23	52

2.5 Gaussian Naive Bayes Classifier

Το μοντέλο στηρίζεται στο Θεώρημα του Bayes , έχοντας σαν βασικές αρχές ότι κάθε μεταβλητή του συνόλου δεδομένων είναι ανεξάρτητη και ισόνομη με κάθε άλλη μεταβλητή, και ότι όλες ακολουθούν κανονική κατανομή.
Για κάθε εγγραφή που θέλει να προβλέψει, υπολογίζει ποιά είναι η πιθανότητα να ανήκει στην κλάση 1 (μη υγιής) και ποιά να ανήκει στην κλάση 0 (υγιής) χρησιμοποιώντας τον τύπο του Bayes :

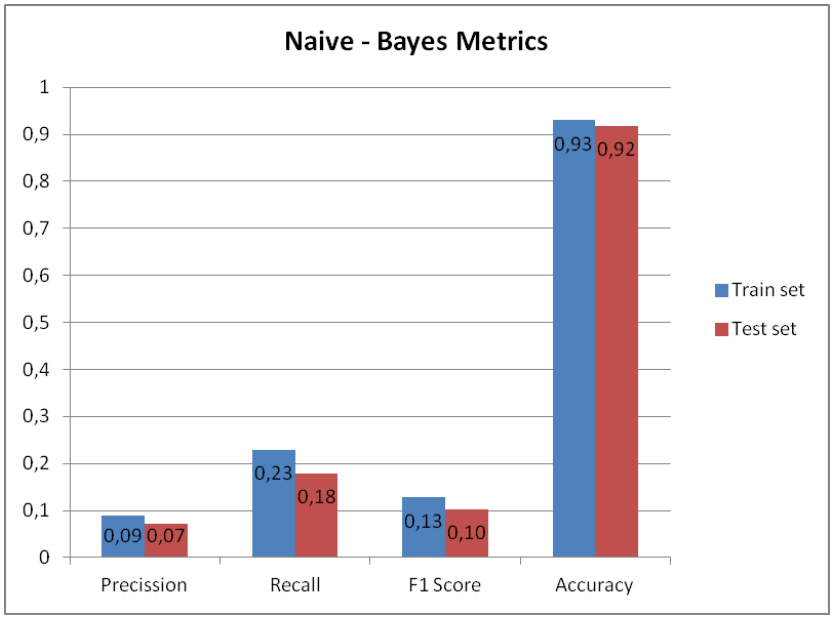
$$p(y \mid \mathbf{X}) = \frac{p(y)p(\mathbf{X} \mid y)}{p(\mathbf{X})}$$

όπου εδώ $y = 0$ ή 1, οι κλάσεις του προβλήματος και $\mathbf{X} = (X_1, X_2, ..., X_{11})$, οι μεταλητές της κάθε παρατήρησης.
Εφόσον παίρνουμε σαν δεδομένο ότι οι μεταλητές είναι ανεξάρτητες και ισόνομες, μπορούμε να πούμε ότι $p(\mathbf{X}) = p(X_1)p(X_2)...p(X_{11})$ και $p(\mathbf{X} \mid y) = p(X_1 \mid y)p(X_2 \mid y)...p(X_{11} \mid y)$.Επίσης αφού κάθε μεταβλητή ακολουθέι κανονική κατανομή, οι πιθανότητες $p(X_1)...$ μπορούν να υπολογιστούν από την συνάρτηση πυκνότητας πιθανότητας της κατανομής.

Πίνακας 5: Προβλέψεις για Gaussian Naive Bayes

	TP	TN	FP	FN
Train	44	7928	452	148
Test	10	1958	130	46

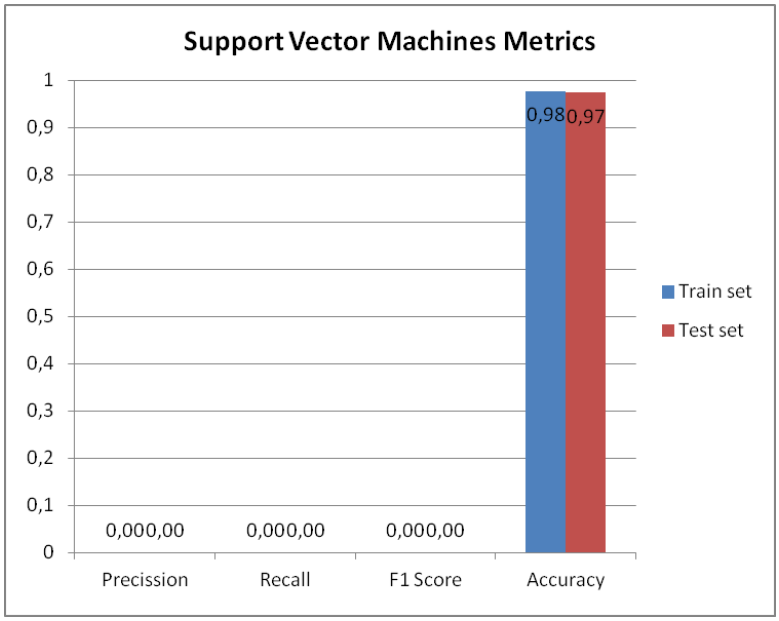
Το μοντέλο βλέπουμε ότι μπόρεσε να προβλέψει 10 από τις 56 μη υγιείς επιχειρήσεις, δηλαδή περισσότερες από κάθε άλλο μοντέλο, αλλά για ακόμα μία φορά το ποσοστό είναι πολυ πιο κάτων από το 62%.



Σχήμα 5: Μετρικές για Gaussian Naive Bayes

2.6 Support Vector Machines

Ο αλγόριθμος που χρησιμοποιούν τα support vector machines είναι παρόμοιος με αυτόν που χρησιμοποιεί το linear descriminant analysis . Δηλαδή, ψάχνει την καλύτερη γραμμή που χωρίζει τα δεδομένα, αυτή την φορά όμως δεν προβάλλει τις παρατηρήσεις πάνω στην ευθεία, αλλά δημιουργεί ένα υπερεπίπεδο.
Για να βρει την καλύτερη δυνατή γραμμή, αναζητάει ποια από όλες τις πιθανές γραμμές που χωρίζουν τα δεδομένα δημιουργεί το μέγιστο υπερεπίπεδο.Δηλαδή, βρίσκει από κάθε κλάση το σημείο το οποίο είναι πιο κοντά στην γραμμή, το οποίο ονομάζεται διάνυσμα υποστήριξης.Σκοπός είναι να μεγιστοποιηθεί η απόσταση μεταξύ των διανυσμάτων υποστήριξης και της γραμμής που δημιουργείται, ώστε να δημιουργηθεί και το μέγιστο υπερεπίπεδο.



Σχήμα 6: Μετρικές για Support Vector Machines

Το μοντέλο αυτό βλέπουμε ότι δεν κατάφερε να προβλέψει καμία επειχήρηση που θα χρεοκοπήσει, είτε αυτό είναι αλήθεια είτε ψέμα.

Πίνακας 6: Προβλέψεις για Support Vector Machines

	TP	TN	FP	FN
Train	0	8380	0	192
Test	0	2088	0	56

2.7 Neural Networks

Ένα νευρωνικό δίκτυο αποτελείται από το 1ο επίπεδο (input layer),το οποίο αποτελείται από νευρώνες οι οποίοι συμβολίζουν την κάθε μεταβλητή που υπάρχει στα δεδομένα του προβλήματος, τα κρυφά επίπεδα (hidden layers) τα οποία αποτελούνται απο βοηθητικούς νευρώνες ώστε να εκπαιδευτεί το μοντέλο, και το τελευταίο επίπεδο (output layer)το οποίο αποτελείται από τόσους νευρώνες όσες είναι και οι κλάσεις των δεδομένων.
Οι νευρώνες του κάθε επιπέδου συνδέονται με όλους τους νευρώνες του επόμενου με κάποιο βάρος.
Στο 1ο επίπεδο υπάρχουν οι τιμές της κάθε μεταβλητής μιας εγγραφής, οι οποίες τροφοδοτούν το επόμενο επίπεδο. Η διαδικασία ξεκινάει χρησιμοποιώντας την 1η εγγραφή. Συμβολίζω με A_i την ποσότητα που παράγεται για το 2ο επίπεδο :

$$A_i = \sum_{j=1}^{N} b_j A_j$$

, όπου

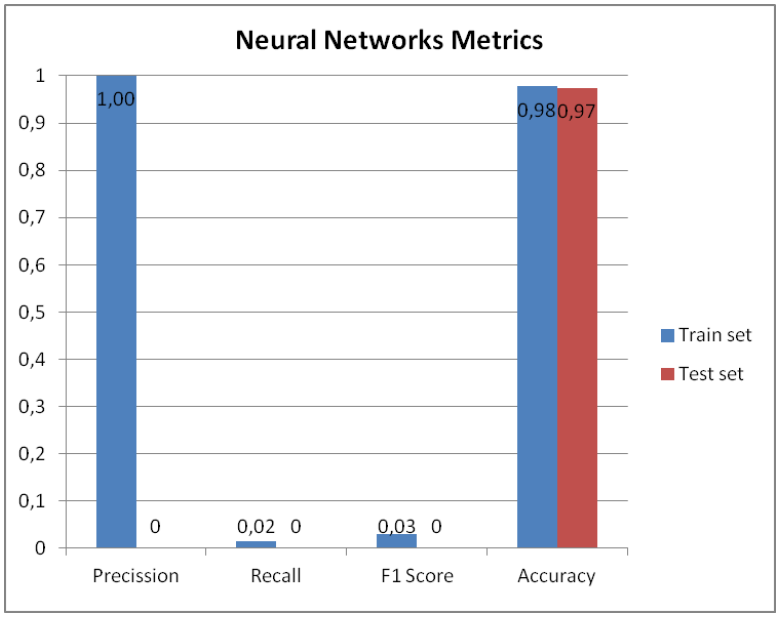
- N ο αριθμός των μεταβλητών,
- b_j η τιμή βάρους που συνδέεται ο νερώνας j του προήγουμενου επιπέδου με τον i νευρώνα,
- A_j η τιμή που υπάρχει στον j νευρώνα του προηγούμενου επιπέδου.

Εδώ να σημειωθεί ότι παίρνουμε την απλή περίπτωση όπου δεν υπάρχει $bias$.
Εφόσον υπολογισθεί για κάθε νευρώνα η τιμή A_i, τότε η τιμή κάθε νευρώνα του 2ου επιπέδου θα είναι ίση με $f(A_I)$, όπου f είναι η συνάρτηση ενεργοποίησης του συγκεκριμένου επιπέδου.
Με αυτό το σκεπτικό το νευρωνικό δίκτυο υπολογίζει τις τιμές των νευρώνων κάθε επιπέδου συναρτήσει των τιμών που έχουν οι νευρώνες στο προηγούμενο επίπεδο.
Στο τέλος η τιμή που θα έχουν οι νευρώνες του τελικού επιπέδου,θα καθορίσει και την κλάση που προβλέπει το νευρωνικό για την συγκεκριμένη εγγραφή. Αν η κλάση που προέβλεψε το μοντέλο για αυτή την εγγραφή είναι διαφορετική από την πραγματική, τότε με μια διαδικασία που ονομάζεται back propagation το νευρωνικό δίκτυο,αφού υπολογίσει το σφάλμα υπολογισμού, αλλάζει μερικά από τα βάρη με τέτοιο τρόπο ώστε να προκύψει καλύτερο αποτέλεσμα για τις επόμενες εγγραφές.
Αυτή η διαδικασία επαναλαμβάνεται για όλες τις εγγραφές στο train set ώστε να καταλήξει με το καλύτερο δυνατό μοντέλο πρόβλεψης νέων εγγραφών για κάποιο δείγμα που δεν έχει ξαναδεί.

Αξίζει να σημειωθεί ότι το μοντέλο δεν κατάφερε να προβλέψει καμία από τις 56 χρεοκωπημένες εταιρείες και ότι τα αποτελέσματα πρόβλεψης για το test set είναι ακριβώς τα ίδια με το support vector machines.

Πίνακας 7: Προβλέψεις για Neural Networks

	TP	TN	FP	FN
Train	3	8380	0	189
Test	0	2088	0	56



Σχήμα 7: Μετρικές για Neural Networks

2.8 Συμπεράσματα

Από ότι καταλάβαμε μέσα από τα διαγράμαμτα των μετρικών και τους πίνακες με τις προβλέψεις που έκανε το κάθε μοντέλο, κανένα δεν ήταν σε θέση ούτε να προβλέψει τις εταιρείες που θα χρεοκωπήσουν με ποσοστό πάνω απο 18%.
Το καλύτερο μοντέλο όσον αφορά τουλάχιστον αυτά τα δεδομένα που είχαμε, αποδείχθηκε να είναι το Gaussian Naive Bayes αφού προέβλεψε με ποσοστό 17% τις χρεοκωπημένες εταιρείες.

3 Μερος Β'

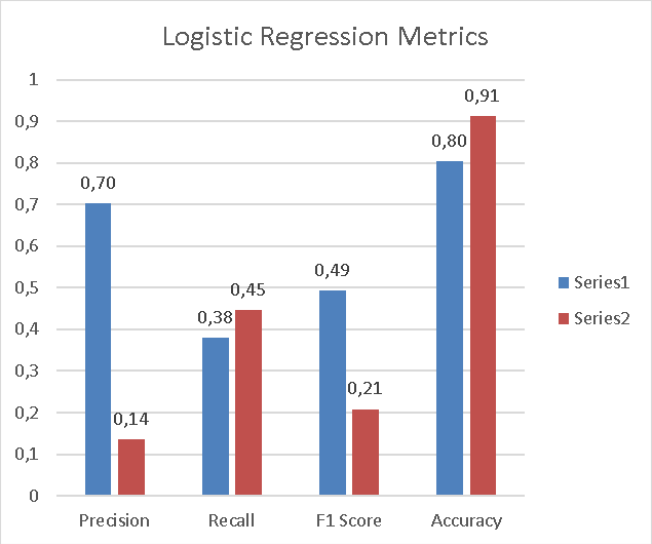
Η αδυναμία των μοντέλων στο 1ο μέρος της εργασίας οφειλόταν στο δείγμα, και συγκεκριμένα στην μεγάλη διαφορά υγειών και μη υγειών εταιρειών.Ο μεγάλος αριθμός υγειών εταιρειών έπαιξε καθοριστικό ρόλο στα αποτελέσματα κάθε ταξινομητή.
Όπως προαναφέρθηκε και στην εισαγωγή, στο Β΄ μερος θα παρουσιαστούν ακριβώς τα ίδια μοντέλα, με την μοναδική διαφορά ότι τα δεδομένα που δίνουμε για εκπαίδευση αυτή τη φορά θα έχουμε αναλογία 3:1 μεταξύ υγειών και μη υγειών εταιρειών.Σε αυτό το κεφάλαιο δεν θα περιγραφούν ξανά το τι κάνει το κάθε μοντέλο, αλλά θα σχολιαστούν τα αποτελέσματα κάθε μοντέλου μετά από αυτή την μικρή τροποποίηση.

3.1 Logistic Regression

Βλέπουμε από τον πρώτο κιόλας αλγόριθμο την διαφορά στα αποτελέσματα. Από εκεί που το ποσοστό πρόβλεψης των μη υγιεων εταρειών ήταν κάτω από 2%, τώρα ανέβηκε στο 44%. Μπορεί και πάλι να μην ικανοποιήθηκε η απαίτηση του προβλήματος, αλλά τουλάχιστον είμαστε θετικά προδιαθετιμένοι για τα αποτελέσματα των υπολοίπων αλγόριθμων.

Πίνακας 8: Προβλέψεις για Logistic Regression pt.B

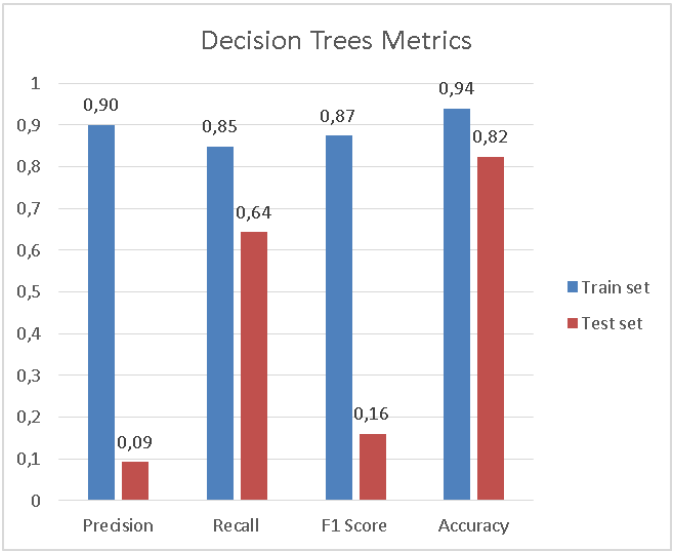
	TP	TN	FP	FN
Train	73	545	31	119
Test	25	1929	159	31



Σχήμα 8: Μετρικές για Logistic Regression pt.B

3.2 Decision Tree Classifier

Εδώ έχουμε την πρώτη φορά που ικανοποιούνται οι απαιττησεις του ροβλήματος. Βρέθηκαν 36 από τις 56 μη υγιείς επιχειρήσεις και 1731 από τις 2088 υγιείς. Βέβαια το *precission* εδώ πέρα είναι αρκετά μικρό, διότι το μοντέλο προέβλεψε συνολικά ότι 357 εταιρείες θα χρεοκωπήσουν, αυξάνοντας έτσι προφανώς τον αριθμό FP.



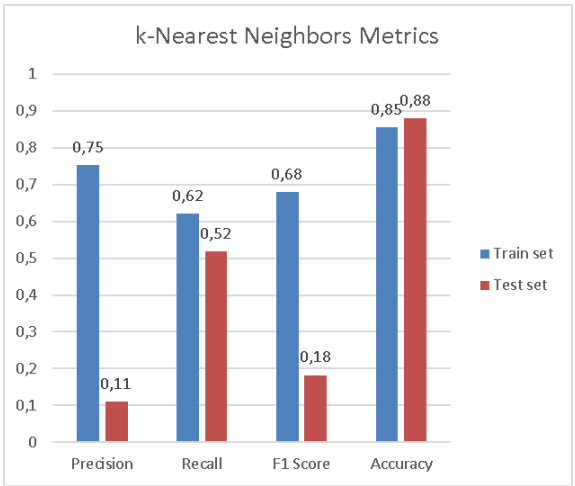
Σχήμα 9: Μετρικές για τα δέντρα αποφάεων pt.B

Πίνακας 9: Προβλέψεις για Decision Trees pt.B

	TP	TN	FP	FN
Train	163	558	18	29
Test	36	1731	357	20

3.3 k-Nearest Neighbors

Με ποσοστό επιτυχίας 52% πάλι δεν ικανοποιούνται οι απαιτήσεις του προβλήματος.Ίσως με διαφορετικό αριθμο κ, δηλαδη άν αλλάζαμε τον αριθμό που δέχεται το μοντέλο ώστε να υπολογίσει τις κ πλησιέστερες αποστάσεις να παίρναμε καλύτερα αποτελέσματα, αλλά με τις default επιλογές (κ=5), ο αλγόριθμος φαίνεται να αποτυγχάνει.

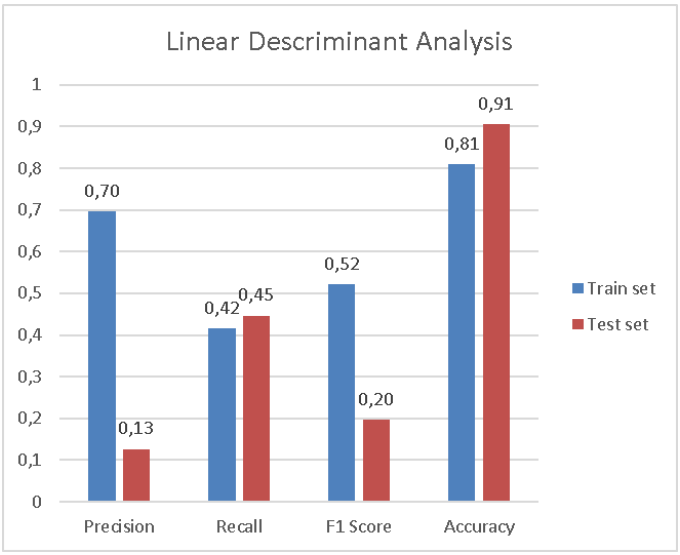


Σχήμα 10: Μετρικές για τους κ πλησιέστερους γείτονες pt.B

Πίνακας 10: Προβλέψεις για κ πλησιέστερους γείτονες pt.B

	TP	TN	FP	FN
Train	119	537	39	73
Test	29	1855	233	27

3.4 Linear Descriminant Analysis



Σχήμα 11: Μετρικές για Linear Descriminant Analysis pt.B

Το συγεκριμένο μοντέλο βλέπω ότι προβλέπει λιγότερα TP σε σχέση με τα 3 προηγούμενα, αλλά βλέπω επίσης ότι προβλέπει πολύ περισσότερες εταιρείες οι οποίες είναι υγιείς. Δεν παίζει όμως μεγάλο ρόλο καθώς η κύρια απαίττηση του προβλήματος δεν ικανοποιείται.

Πίνακας 11: Προβλέψεις για Linear Descriminant Analysis pt.B

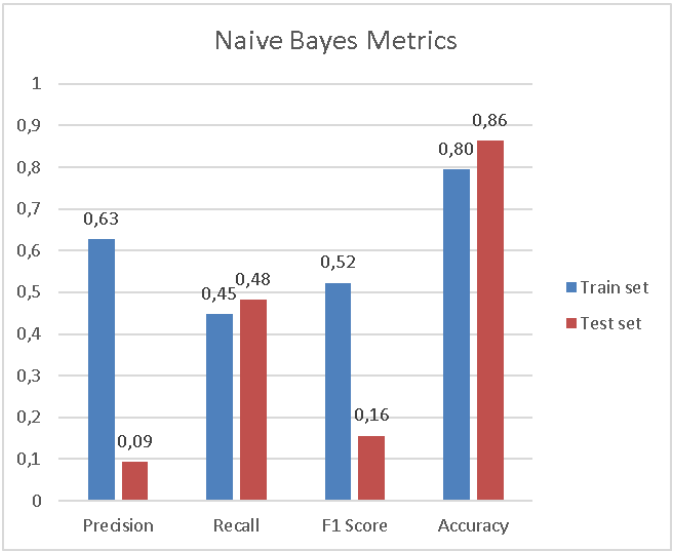
	TP	TN	FP	FN
Train	80	541	35	112
Test	25	1916	172	31

3.5 Gaussian Naive Bayes

Μπορεί η μετρική *recall* να βγάζει καλύτερο αποτέλεσμα συγκριτικά με το Linear Descriminant Analysis αλλά όλες οι υπόλοιπες μετριές βγάζουν χειρότερα αποτελεσμάτα.

Πίνακας 12: Προβλέψεις για Gaussian Naive Bayes pt.B

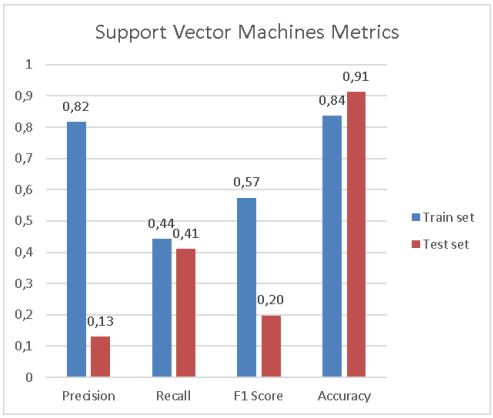
	TP	TN	FP	FN
Train	86	525	51	106
Test	27	1825	263	29



Σχήμα 12: Μετρικές για Gaussian Naive Bayes pt.B

3.6 Support Vector Machines

Εν,ω κανονικά θα περίμενε κανείς αυτό το μοντέλο να μας δώσει τα καλύτερα δυνατά αποτελέσματα, αφού δουλεύει καλύτερα με μικρότερο αριθμό δεδομένων, παρατηρούμε ότι κάτι τέτοιο δεν ισχύει. Μάλιστα, το ποσοστό επιτυχημένων προβλέψεων για τις μη υγιείς εταιρείες είναι ίσο με 41%, λιγότερο από κάθε μοντέλο στο Β΄ Μέρος της άσκησης έως τώρα.



Σχήμα 13: Μετρικές για Support Vector Machines pt.B

Πίνακας 13: Προβλέψεις για Support Vector Machines pt.B

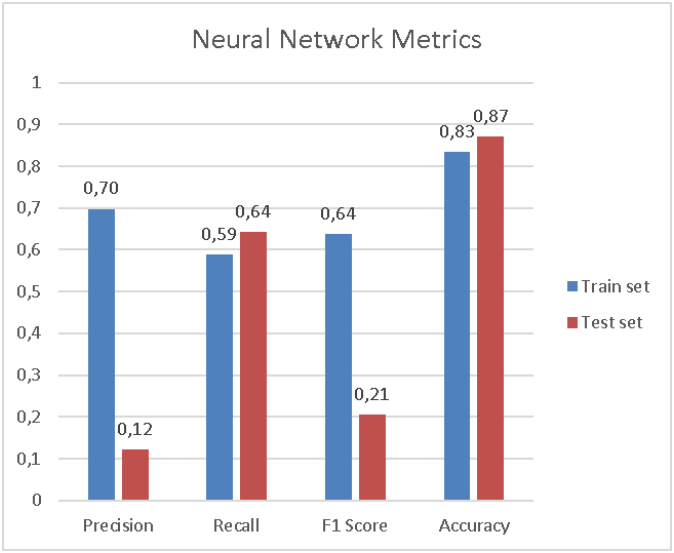
	TP	TN	FP	FN
Train	85	557	19	107
Test	23	1935	153	33

3.7 Neural Networks

Εδω παρατηρούμε ότι έχουμε και δεύτερο ταξινομητή ο οποίος προβλέπει με ικανοποιητικό ποσοστό τις χρεωκοπημένες εταιρείες, και μάλιστα είναι καλύτερος από τον πρώτο αφού βλέπουμε ότι το ποσοστό ανέρχεται στο 64 %. Να σημειωθέι ότι δεν πειράξαμε σε μεγάλο βαθμό το νευρωνικό, δηλαδή δεν βάλαμε πολλά κρυφά επίπεδα, αντιθέτως με μόνο ένα κρυφό επίπεδο με 4 νευρώνες βλέπουμε ότι έχουμε επιθυμητό αποτέλεσμα.%.

Πίνακας 14: Προβλέψεις για Neural Networks pt.B

	TP	TN	FP	FN
Train	113	527	49	79
Test	36	1830	258	20



Σχήμα 14: Μετρικές για Neural Networks pt.B

4 Τελικά Συμπεράσματα

Παρά τις προσδοκίες που είχαμε για τα μοντέλα ταξινόμησης μετά την αποτυχία που είχαν όταν τροφοδοτήθηκαν με όλα τα δεδομένα εκπαίδευσης, ότι θα βρισκόταν σε θέση να προβλεπουν σε ικανοποιητικό ποσοστό την κλάση κάθε εταιρείας, βλέπουμε ότι κάτι τέτοιο δεν συμβάινει. Μπορούμε να συνειδητοποιήσουμε λοιπόν ότι η αναλογία δύο κλάσεων στις οποίες ένας ταξινομητής εκπαιδεύεται και κάνει προβλέψεις παίζει σημαντικό ρόλο στα αποτελέσματά που παράγει.

Τα μόνα μοντέλα που κατάφεραν να βγάλουν ικανοποιητικά αποτελέσματα ήταν αυτά των δέντρων αποφάσεων και των νευρωνικών δικτύων. Όλα τα άλλα μοντέλα δεν ήταν σε θέση να προβλέψουν με ποσοστό πάνω από 62% τις χρεοκωπημένες εταιρείες, καθιστόντας τα άχρηστα για μελλονική χρήση. Το ποιό μοντέλο είναι καλύτερο από κάποιο άλλο μπορεί να βρεθέι κοιτώντας τα γραφήματα τα οποία παρουσιάστηκαν στο Β΄ Μερος της εργασίας.

Εδώ αξίζει να σημειωθέι ότι ενώ βρέθηκαν μόνο δύο μοντέλα που παρήγαγαν επαρκή αποτελέσματα, αυτό δεν σημαίνει ότι με διαφορετική επιλογή παραμέτρων θα εξακολουθουν τα υπόλοιπα να μην δίνουν σωστά αποτελέσματα. Για παράδειγμα τα support vector machines μπορεί να μας εξυπηρετούσαν αν διαλέγαμε διαφορετικό kernel . Γενικά εφόσον χρησιμοποιήσαμε τις *default* παραμέτρους κάθε μοντέλου δεν ήταν καθόλου σίγουρο ότι με αυτά τα δεδομένα θα έχουμε κάποιο σωστό αποτέλεσμα.

Συνοψίζοντας, τα καλύτερο μοντέλα για αυτό το πρόβλημα από ότι φαίνεται είναι αυτό των νευρωνικών δικτύων και των δεντρών αποφάσεων, τα οποία προβλέπουν με ποσοστό 64% τις εταιρείες.