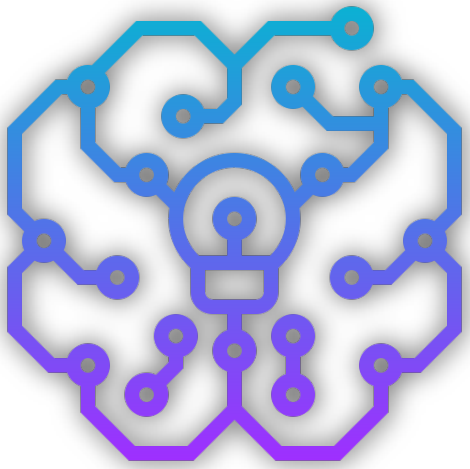


Machine Learning Project

Αργύριος Κορωναίος
23/10/2022

Περιεχόμενα

Πίνακας Εικόνων

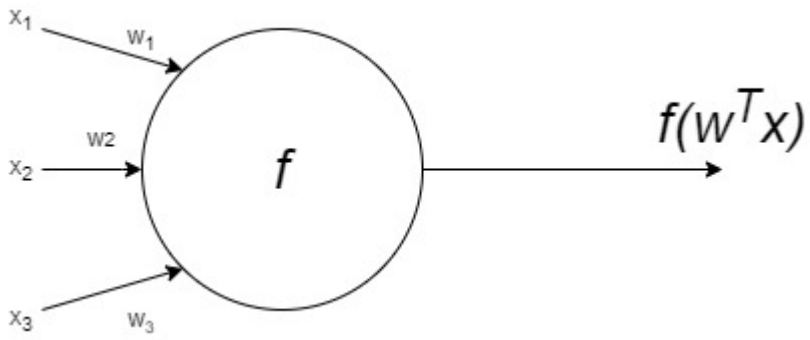
Περιγραφή Προβλήματος

Το πρόβλημα που δίνεται προς λύση είναι ένα πρόβλημα διαχωρισμού δυο διαφορετικών κλάσεων. Πιο συγκεκριμένα, τα δεδομένα που έχουμε στην διάθεσή μας είναι κάποιοι χρηματοπιστωτικοί δείκτες και πληροφορίες για την δραστηριότητα από 10.716 ελληνικές εταιρείες. Δίνεται, επίσης, και επιπλέον πληροφορία για την κατάσταση της εταιρίας, ταξινομώντας το σύνολο των εταιριών σε δύο κλάσεις ξένες μεταξύ τους τις υγιής και τις χρεωκοπημένες εταιρίες. Η δοκιμασία, λοιπόν, είναι η εξής να δημιουργηθεί ένα κατάλληλο μοντέλο ταξινόμησης που θα εντοπίζει τις εταιρείες που θα χρεωκοπήσουν, έχοντας ως είσοδο τους διάφορους χρηματοπιστωτικοί δείκτες και πληροφορίες για την δραστηριότητα των εταιριών. Το μοντέλο πρέπει να πληρεί κάποιες συγκεκριμένες προδιαγραφές, οι οποίες είναι: με ποσοστό επιτυχίας τουλάχιστον 62% να εντοπίζει τις εταιρείες που θα πτωχεύσουν και ταυτόχρονα, με ποσοστό μεγαλύτερο του 70% να εντοπίζονται οι υγιής εταιρίες που δεν θα πτωχεύσουν. Αρχικά, έγινε μια προεργασία πάνω στα δεδομένα έτσι ώστε να χρησιμοποιηθούν για την εκπαίδευση των διαφορών μοντέλων. Τα δεδομένα χωρίστηκαν αρχικά στα δεδομένα εισόδου που δεν είναι άλλο από τις τιμές στους 11 χρηματοπιστωτικοί δείκτες και δεδομένα εξόδου που είναι η κατάσταση της εταιρείας, χρεωκοπημένη ή υγιής. Στην συνέχεια τα δεδομένα χωρίστηκαν με τυχαίο τρόπο σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου με αναλογία 0,75/0,25. Επειδή υπάρχει διάνυσμα στόχων όλα τα παρακάτω μοντέλα που θα δημιουργηθούν είναι μοντέλα επιβλεπόμενης μάθησης. Θα μετρηθούν και αξιολογηθούν οι επιδόσεις τους πάνω και στα δύο σύνολα δεδομένων, εκπαίδευσης και ελέγχου. Σκοπός δεν είναι παρά άλλος από το να βρεθεί ένα μοντέλο που να ικανοποίει τις προδιαγραφές που αναφέρθηκαν πιο πάνω στα δεδομένα ελέγχου.

Συνοψίζοντας, το πρόβλημα που καλούμαστε να επιλύσουμε είναι ένα πρόβλημα κατηγοριοποίησης δύο κλάσεων. Οι μέθοδοι μάθησης που θα εφαρμοστούν είναι όλες μάθηση με επίβλεψη καθώς έχουμε πλήρη εικόνα την έξοδο. Ενδεικτικά, θα εκπαιδευτούν συνολικά επτά διαφορετικά μοντέλα. Τα οποία είναι: Γραμμική Διακριτική Ανάλυση, Λογιστική Παλινδρόμηση, Δέντρα Απόφασης, κ Πλησιέστεροι Γείτονες, Naive Bayes, Μηχανές Διανυσμάτων Υποστήριξης και Νευρωνικά Δίκτυα. Λόγω του θεωρήματος καθολικής προσέγγισης για τα νευρωνικά δίκτυα, θα δοθεί σε αυτά μεγαλύτερη έμφαση.

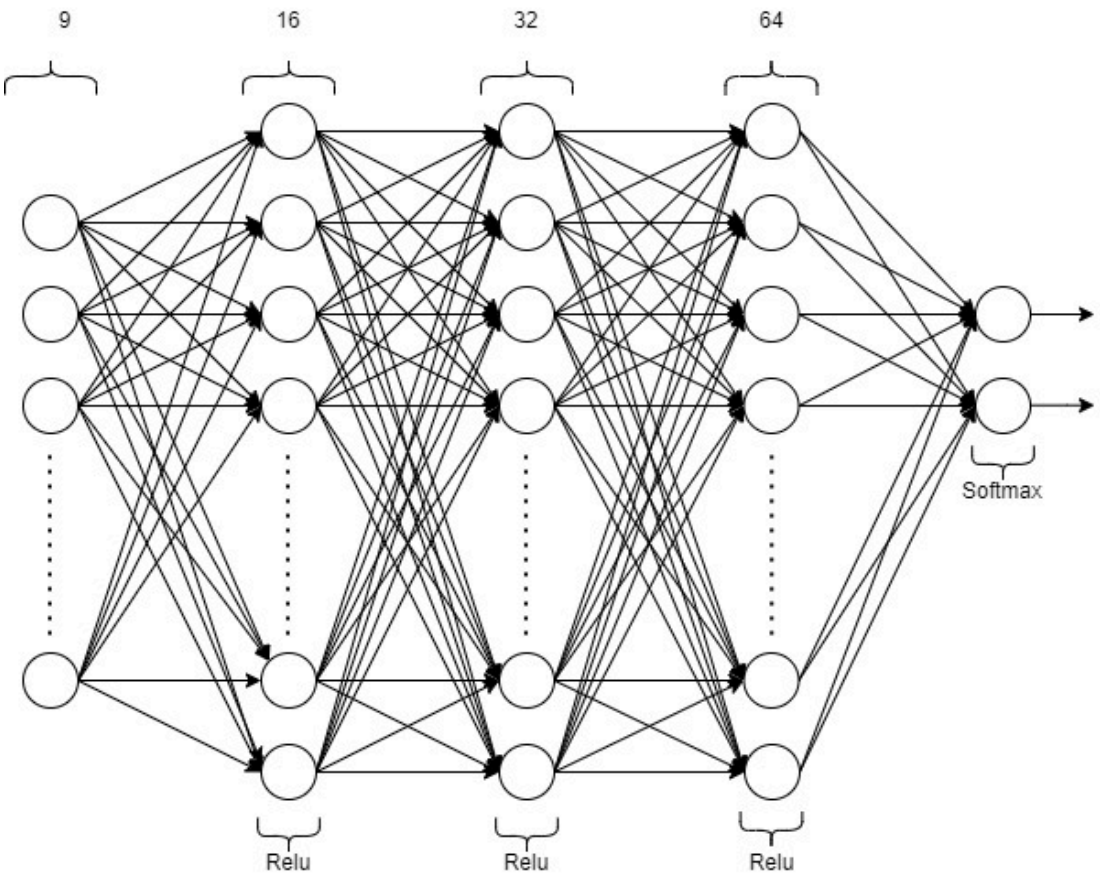
1η Μέθοδος: Νευρωνικό Δίκτυο

Αποτελούν είδος τεχνητής νοημοσύνης, βασισμένο στην προσπάθεια ενσωμάτωσης και προσομοίωσης σε ένα υπολογιστικό σύστημα των βασικών χαρακτηριστικών της ανθρώπινης σκέψης για να επιλυθούν πρακτικά προβλήματα. Οι είσοδος στον νευρώνα είναι ένα διάνυσμα το οποίο πολλαπλασιάζεται με τα συνοπτικά βάρη. Ύστερα το αποτέλεσμα αυτό περνά μέσα από την συνάρτηση ενεργοποίησης του νευρώνα και η τελική έξοδος έχει την μορφή $f(w^Tx)$. Συνδυάζοντας πολλούς νευρώνες σε πολλά στρώματα έχουμε ένα νευρωνικό δίκτυο. Οι παράμετροι που εκπαιδεύονται είναι τα συνοπτικά βάρη w_i. Για την εκπαίδευση χρησιμοποιείτε ο αλγόριθμος back propagation. Στον οποίον τα σφάλματα γυρίζουν «πίσω» με την βοήθεια των μερικών παραγώγων.



Εικόνα 1: Απλός τεχνητός νευρώνας

Υλοποιήθηκε νευρωνικό δίκτυο με την παρακάτω αρχιτεκτονική,



Εικόνα 2:Αρχιτεκτονική 1ου NN

Η αποτελεσματικότητα του παραπάνω νευρωνικού φαίνεται από τον παρακάτω πίνακα.

Train Confusion Matrix	
TP = 42	FP = 4
FN = 131	TN = 7860

Test Confusion Matrix	
2	73
2	2599

Table 1: Confusion Matrix of 1st NN

2η Μέθοδος: Λογιστική Παλινδρόμηση

Η λογιστική Παλινδρόμηση προσπαθεί να προβλέψει τις τιμές της ποιοτικής μεταβλητής Y μέσα από το παρακάτω λογιστικό μοντέλο

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Όπου $p(x) = P(Y = y|X = x)$ με $x = (x_1, x_2, \dots, x_n)$ οι τιμές των τυχαίων μεταβλητών X_i.

Το επόμενο μοντέλο που δοκιμάστηκε ήταν με χρήση της λογιστικής παλινδρόμησης. Τα αποτελέσματα του όπως θα δούμε παρακάτω ήταν απογοητευτικά. Καθώς, το μοντέλο έκανε overfit τα δεδομένα με αποτέλεσμα να μη έχει καθόλου καλά score στα δεδομένα ελέγχου.

Train Confusion Matrix	
TP = 1	FP = 0
FN = 172	TN = 7864

Test Confusion Matrix	
0	0
75	2604

Table 2: Confusion Matrix of Logistic Regression

Όπως παρατηρούμε από το Test Confusion Matrix, πρακτικά το μοντέλο κατηγοριοποίησε όλες τις εταιρίες ως υγιείς.

3η Μέθοδος: Δέντρο Απόφασης

Τα δέντρα απόφασης ή δέντρα λήψης αποφάσεων προβλέπουν την τιμή μιας ποιοτικής μεταβλητής. Το δέντρο φτιάχνεται από αλγορίθμους όπως ID3 ή Gini Index. Το πλεονέκτημα της μεθόδους είναι η εύκολη ερμηνεία των αποτελεσμάτων μέσα από ένα δενδρόγραμμα αποτελούμενο από κόμβους και φύλλα.

Δημιουργήθηκε ακόμα ένα μοντέλο με την χρήση δέντρου απόφασης. Χωρίς να γίνει κάποιο κλάδεμα (pruning). Αλλά ορίστηκε μέγιστο βάθος 6 και είχαμε τα παρακάτω αποτελέσματα.

Train Confusion Matrix	
TP = 30	FP = 4
FN = 143	TN = 7860

Test Confusion Matrix	
0	31
75	2593

Table 3:Confusion Matrix of Decision Tree

Ο λόγος που στην εκπαίδευση του δέντρου υπήρχε μέγιστο βάθος ήταν ότι υπό διαφορετικές συνθήκες το μοντέλο θα «παπαγάλιζε» τα δεδομένα και δεν θα μάθαινε τίποτα.

4η Μέθοδος: Γραμμική Διακριτική Ανάλυση

Τέταρτο μοντέλο είναι αυτό της γραμμικής διακριτικής ανάλυσης (Linear Disciminat Analysis). Ο σκοπός της γραμμικής Διακριτής Ανάλυσης είναι να προβλέψει τις τιμές μιας ποιοτικής μεταβλητής Y με την χρήση ποσοτικών ή ποιοτικών μεταβλητών X_i. Θεωρούμε ότι $X_i|Y=y \sim N(\mu_1,\sigma_1)$, όπου y οι διάφορες τιμές της ποιοτικής μεταβλητής Y. Έτσι έχουμε τόσες κανονικές κατανομές όσες και οι διαφορές τιμές της Y.

Οδηγούμαστε στον υπολογισμό της κανονικής ή διαχωριστικής μεταβλητής $Z=w_1X_1+w_2X_2+\ldots+w_nX_n$. Έτσι αν η Z ξεπεράσει της τιμή c είναι στην κατηγορία 1 αλλιώς είναι στην κατηγορία 0. Στο δοθέν πρόβλημα η παραπάνω τεχνική είχε με τα ακόλουθα αποτελέσματα στα δεδομένα εκπαίδευσης και ελέγχου.

Train Confusion Matrix	
TP = 9	FP = 72
FN = 164	TN = 7792

Test Confusion Matrix	
5	31
70	2573

Table 4:Confusion Matrix of Linear Disciminat Analysis

5η Μέθοδος: Naïve Bayes

Ο Naïve Bayes classifier χρησιμοποίει τον κανόνα του Bayes για να προβλέψει την κλάση C_k του σημείου x.

$$p(C_k|\boldsymbol{x}) = \frac{p(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})} \text{ άρα } \tilde{y} = arg \max_{k\in\{1,2,\ldots K\}} p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

Επόμενο μοντέλο είναι Naïve Bayes. Με τα παρακάτω αποτελέσματα,

Train Confusion Matrix	
TP = 40	FP = 424
FN = 133	TN = 7440

Test Confusion Matrix	
20	139
55	2465

Table 5:Confusion Matrix of Naive Bayes

6η Μέθοδος: κ- Κοντινότεροι Γείτονες

Ο αλγόριθμος ταξινόμησης των κ-Κοντινότερων Γειτόνων προβλέπει την τιμή της κατηγορικής μεταβλητής Y βάση την τιμή της μεταβλητής που έχουν οι γείτονες του σημείου. Πιο συγκεκριμένα κάθε παρατήρηση στα δεδομένα είναι ένα σημείο στον ενδιαστάτο χώρο. Έστω $X=(X_1, X_2,\ldots, X_n)$ το σημείο προς κατηγοριοποίηση. Υπολογίζουμε βάση μιας μετρικής την απόσταση από τα υπόλοιπα σημεία του συνόλου. Επιλέγουμε τα κ κοντινότερα για να αποφασίσουμε την κλάση του σημείου μέσα από ψηφοφορία, για αυτό το κ είναι περιττός αριθμός για να μην υπάρχει ισοψηφία.

Για την επόμενη μέθοδο δημιουργήθηκαν 3 διαφορετικά μοντέλα, για τις διάφορες τιμές του κ=3,5,7. Τα τρία αυτά μοντέλα ήταν πολύ κοντά όσον αφορά την αποτελεσματικότητα. Με τις μέσες επιδόσεις να είναι οι εξής:

Train Confusion Matrix	
TP = 24	FP = 7
FN = 147	TN = 7857

Test Confusion Matrix	
1	5
74	2599

Table 6: Confusion Matrix of k-NN

7η Μέθοδος: Μηχανές Διανυσμάτων Υποστήριξης

Η κεντρική ιδέα πίσω από τα SVM συνοψίζεται στο ότι για ένα δοθέν δείγμα εκπαίδευσης, η μηχανή διανυσμάτων υποστήριξης δημιουργεί ένα βέλτιστο υπερεπίπεδο απόφασης έτσι ώστε τα περιθώρια διαχωρισμού μεταξύ των δύο κλάσεων να είναι μέγιστα. Ανάγει δηλαδή το πρόβλημα κατηγοριοποίησης σε ένα πρόβλημα βελτιστοποίησης στο οποίο η βέλτιστη λύση είναι υπαρκτή και μπορεί να βρεθεί. Με την χρήση kernel μπορεί να λύση περίπλοκα προβλήματα.

Τελευταίο μοντέλο είναι ένα SVM. Στο οποίο δεν έγινε κάποιο tuning στις παραμέτρους. Χρησιμοποιήθηκαν οι default ρυθμίσεις. Έχοντας τα παρακάτω αποτελέσματα.

Train Confusion Matrix	
TP = 0	FP = 0
FN = 173	TN = 7440

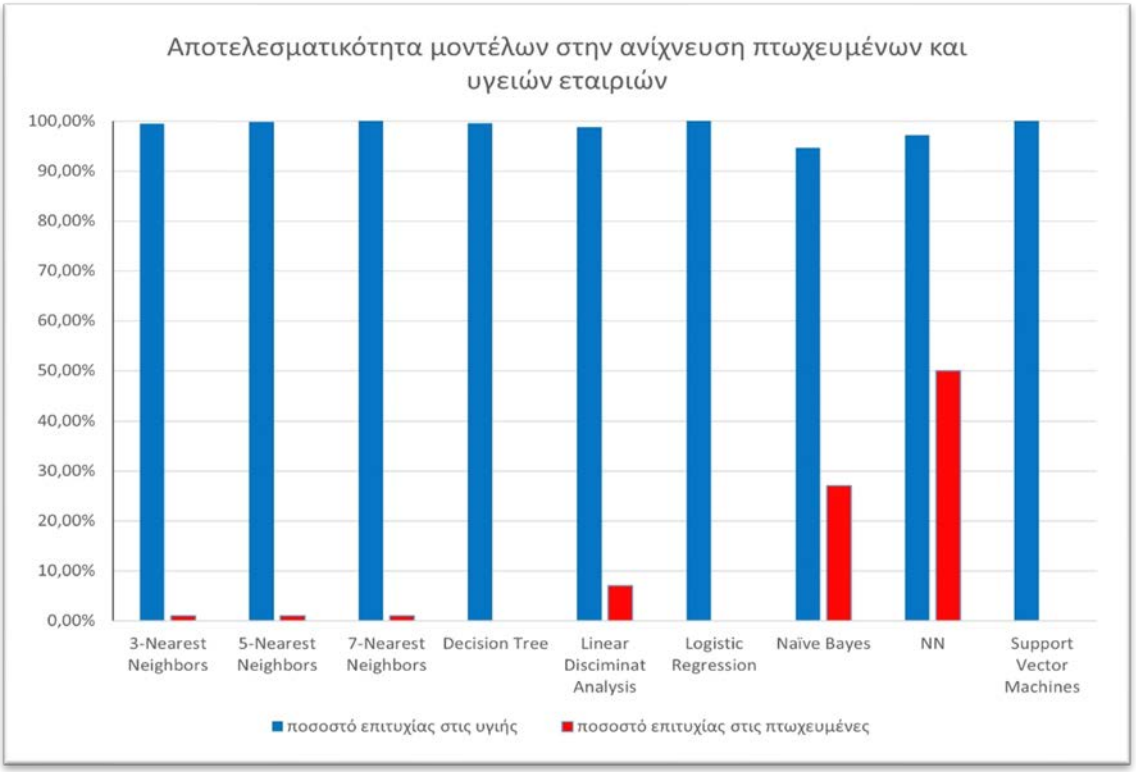
Test Confusion Matrix	
0	0
75	2465

Table 7: Confusion Matrix of SVM

Σε αυτήν την περίπτωση το μοντέλο δεν μπόρεσε να ξεχωρίσει τις υγιής από τις πτωχευμένες εταιρίες και απλά ταξινόμησε όλες τις εταιρίες ς οικονομικά υγιείς.

Σχολιασμός πρώτων αποτελεσμάτων

Όπως βλέπουμε και στο παρακάτω ραβδόγραμμα, όλα τα μοντέλα έχουν πολύ καλή συμπεριφορά όσον αφορά τον εντοπισμό των οικονομικά υγιή εταιριών. Ξεπερνώντας κατά πολύ το κατώφλι του 70%, μάλιστα όλα τα μοντέλα έχουν ποσοστό επιτυχίας πάνω από 90%. Τα ποσοστά, όμως, δεν είναι το ίδιο καλά και στην ανίχνευση πτωχευμένων εταιριών. Με κάποιες μεθόδους να ξεχωρίζουν, πιο συγκεκριμένα αυτές των NN και των Naïve Bayes, παρόλα αυτά ακόμα και αυτές έχουν απόδοση πιο κάτω από το επιθυμητό 62%.



Εικόνα 3: Γράφημα αποτελεσματικότητας των μοντέλων

2η Φάση

Παρατηρούμε, πως η αναλογία πτωχευμένων και εύρωστων εταιριών στα δεδομένα εκπαίδευσης είναι περίπου 1 προς 43. Δηλαδή για κάθε μια πτωχευμένη εταιρία υπάρχουν 43 οικονομικά υγιείς. Αυτή η αναλογία δημιουργεί μια «προκατάληψη» (bias) σε όλα τα μοντέλα, σε διαφορετικό βαθμό στο κάθε ένα, υπέρ της κλάσης των εύρωστων εταιριών. Τα αποτελέσματα, αυτής της προκατάληψης παρατηρούνται μέσα από τα πολύ χαμηλά scores των μοντέλων στην εύρεση χρεωκοπημένων εταιριών. Για αυτόν ακριβώς το λόγω αποφασίστηκε να γίνει μια τροποποίηση στα δεδομένα εκπαίδευσης. Αφαιρούνται με τυχαίο τρόπο υγιής εταιρίες με σκοπό η αναλογία πτωχευμένων και εύρωστων εταιριών να είναι 1 προς 3. Φυσικά δεν κάνουμε κάποια αλλαγή στα δεδομένα ελέγχου, τα οποία τα αφήνουμε ως έχουν.

2ο SVM

Δημιουργήθηκε νέο SVM στο οποίο έγινε tunning να έχει τα καλύτερα αποτελέσματα. Η επιλογή τους SVM classifier δεν είναι τυχαία, αφού τα SVM δουλεύουν πολύ καλά με μικρό όγκο δεδομένων και το training dataset έχει μικρύνει κατά πολύ μετά την

αλλαγή που έγινε στην 2η φάση. Έτσι, μετά το tunning έχουμε ένα SVM με rbf kernel, C=0.1 και γ=1. Το οποίο έχει τα παρακάτω αποτελέσματα

Train Confusion Matrix	
TP = 72	FP = 18
FN = 101	TN = 501

Test Confusion Matrix	
29	143
46	2461

Όπως μπορούμε να συγκρίνουμε με την βοήθεια του Confusion πίνακα *(Table 7: Confusion Matrix of SVM)* υπάρχει βελτίωση άλλα όχι τόσο μεγάλη ώστε να ισχυριστούμε ότι λύσαμε το πρόβλημα, καθώς το ποσοστό εύρεσης πτωχευμένων εταιριών είναι 42% χαμηλότερο από επιθυμητό όριο.

2ο Naïve Bayes

Επόμενη προσπάθεια είναι η εκπαίδευση ενός Naïve Bayes classifier με τα νέα δεδομένα. Με τα εξής αποτελέσματα:

Train Confusion Matrix	
TP = 95	FP = 61
FN = 78	TN = 458

Test Confusion Matrix	
43	342
32	2262

Παρατηρώντας τα αποτελέσματα και συγκρίνοντας με αυτά του παλαιού μοντέλου *(Table 5:Confusion Matrix of Naive Bayes)* βλέπουμε πως υπάρχει βελτίωση όσον αφορά την ανίχνευση πτωχευμένων εταιριών. Με το ποσοστό να φτάνει το 57%, χαμηλότερο από το επιθυμητό 62%

2ο 3-NN με απόσταση cosine

Δημιουργήθηκε ένα μοντέλο 3-NN και μετά από πειράματα επιλέχθηκε η απόσταση cosine έναντι της ευκλείδειας. Η απόσταση cosine ορίζεται ως εξής:

$$D_C(A,B) = 1 - \frac{\sum_{i=1}^{n} A_i B_i}{||A||\,||B||}$$

Με τα παρακάτω αποτελέσματα:

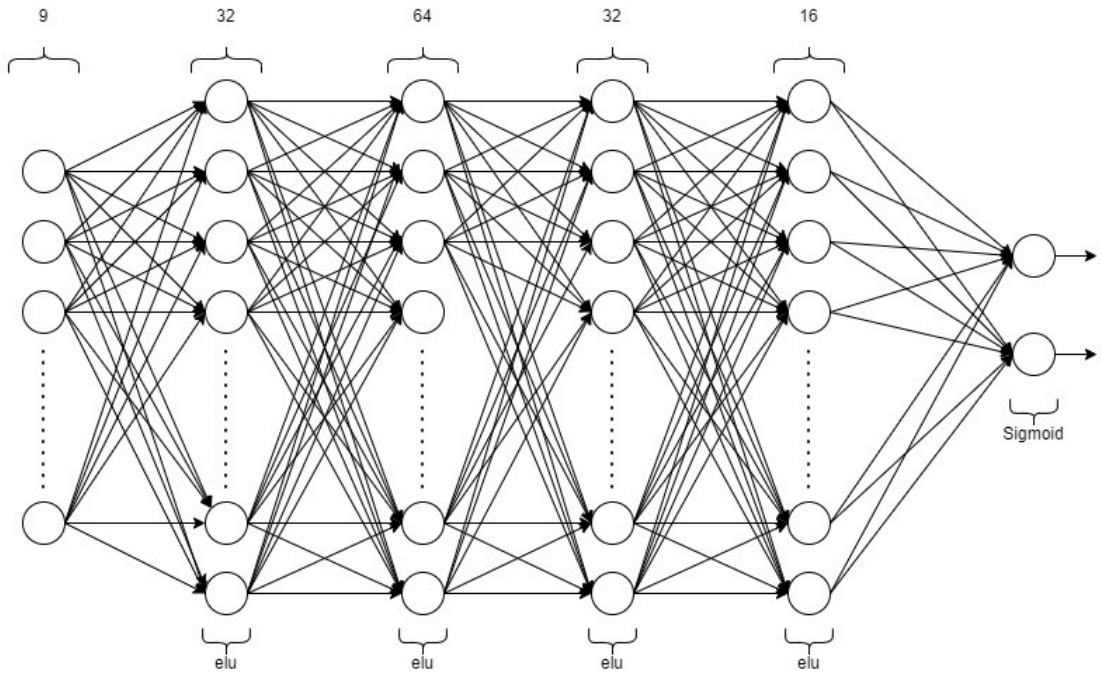
Train Confusion Matrix	
TP = 113	FP = 34
FN = 60	TN = 485

Test Confusion Matrix	
34	278
41	2326

Βλέπουμε ότι ο 3-NN classifier έχει καλύτερη απόδοση στο training set από τον Naïve Bayes όμως δεν έχει καλύτερα αποτελέσματα στο test set αφού έχει ποσοστό ανίχνευση πτωχευμένων εταιριών κάτω από το 50%

2ο NN

Επόμενη υλοποίηση είναι αυτή ενός νευρωνικού δικτύου το οποίο, όμως, θα διαφέρει από το αρχικό ως προς την αρχιτεκτονική. Επειδή το πρόβλημα είναι αρκετά σύνθετο δημιουργήθηκε ένα νευρωνικό με περισσότερα layers και περισσότερους νευρώνες, η αρχιτεκτονική του οποίου φαίνεται στο παρακάτω σχήμα. Σαν συνάρτηση ενεργοποιήσεις των νευρώνων μετά από πειραματικές δοκιμές επιλέχθηκε η elu αντί της relu. Επίσης στο τελευταίο layer χρησιμοποιήθηκε η σιγμοειδής συνάρτηση και όχι η softmax γιατί η πρώτη έχει καλύτερη συμπεριφορά σε προβλήματα δύο κλάσεων.

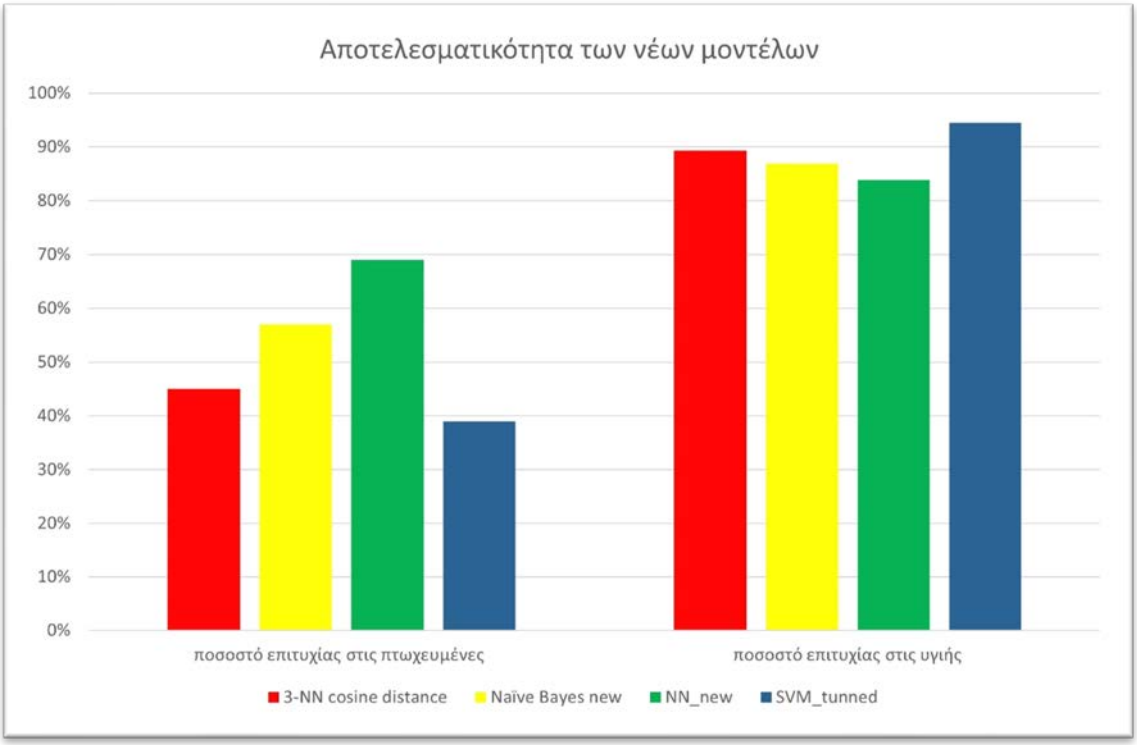


Εικόνα 4: Αρχιτεκτονική 2ου NN

Train Confusion Matrix	
TP = 122	FP = 78
FN = 51	TN = 441

Test Confusion Matrix	
52	419
23	2185

Όπως βλέπουμε από το παραπάνω πίνακα τα αποτελέσματα είναι εξαιρετικά. Αφού το NN μοντέλο όχι μονό είναι αρκετά καλύτερο από τα προηγούμενα αλλά και ικανοποιεί τις προδιαγραφές, αφού έχει ποσοστό ανιχνεύσεις των πτωχευμένων εταιριών 69% και ποσοστό ανιχνεύσεις των υγειών εταιριών με ποσοστό 84%.



Εικόνα 5: Γράφημα αποτελεσματικότητας των νέων μοντέλων

Όπως βλέπουμε από το παραπάνω γράφημα το NN έχει με διαφορά την καλύτερη συμπεριφορά ως προς την εύρεση πτωχευμένων εταιριών. Όμως, δεν είναι το καλύτερο στην εύρεση εύρωστων εταιριών, αυτή η διαφορά, βέβαια δεν είναι πάρα πολύ μεγάλη για να πούμε ότι το μοντέλο πρέπει να απορριφθεί αφού είναι πολύ πάνω από το όριο που έχει τεθεί.

<u>Συμπεράσματα</u>

Το πρόβλημα που ήταν προς επίλυση ο διαχωρισμός των πτωχευμένων και τον υγειών εταιριών. Τα μοντέλα που δημιουργήθηκαν στην $1^{η}$ φάση της επίλυσης δεν είχαν τα στάνταρ απόδοση. Αυτό γιατί στο σύνολο δεδομένων εκπαιδεύσεις η αναλογία των δυο κλάσεων ήταν τέτοια που προκαλούσε μια «προκατάληψη» υπέρ της μιας κλάσης. Για να λυθεί αυτό το ζήτημα, αφαιρέθηκαν τυχαίες υγιής εταιρίες μονό από το σύνολο των δεδομένων εκπαιδεύσεις ώστε η τελική αναλογία να είναι 1 πτωχευμένη για κάθε 3 εύρωστες εταιρίες. Η αλλαγή αυτή βελτίωσε δραματικά τις επιδόσεις των μοντέλων. Καθώς όπως είδαμε όλα τα μοντέλα είχαν καλύτερη επίδοση μετά την αλλαγή στο σύνολο των δεδομένων εκπαιδεύσεις. Για να ικανοποιηθούν οι αρχικές απαιτήσεις απόδοσης δοκιμαστικά 4 διαφορετικά μοντέλα. Τα καλύτερα σκορ επιτευχθήκαν από ένα σύνθετο νευρωνικό δίκτυο με 4 κρυφά επίπεδα και συνολικά 155 νευρώνες.