

Classification Problems

Παναγιώτης Μίσιος (**ics22042**)

Μηχανική Μάθηση

12/12/2025

Περιεχόμενα

1	Εισαγωγή	3
2	Μεθοδολογία	4
	Logistic Regression	4
	k-Nearest Neighbors	6
	Gaussian Naive Bayes	7
	Support Vector	9
3	Συμπεράσματα	10

1 Εισαγωγή

Η συγκεκριμένη εργασία έχει στόχο την εύρεση του καλύτερου δυνατού μοντέλου για τον εντοπισμό εταιρειών που θα κηρύξουν χρεωκοπία. Δηλαδή το πρόβλημα που αντιμετωπίζεται είναι αυτό της ταξινόμησης (classification).

Τα ζητούμενα είναι:

1. να διαβάζονται τα δεδομένα από το παρεχόμενο αρχείο excel.
2. να τυπώνει στην οθόνη ανά έτος τα ακόλουθα στοιχεία:
 - a. Αριθμό υγιών και χρεοκοπημένων επιχειρήσεων
 - b. Την min, max, average τιμή για κάθε δείκτη
3. να κανονικοποιηθούν τα δεδομένα στο διάστημα [0,1], να υλοποιηθούν στη συνέχεια τέσσερα (4) classification μοντέλα της επιλογής μας, των οποίων τα αποτελέσματα στα train και test sets θα συγκριθούν βάσει των τιμών:
 - a. False Positive (FP)
 - b. False Negative (FN)
 - c. True Positive (TP)
 - d. True Negative (TN)
 - e. Precision (TP / TP + FP)
 - f. Accuracy (TP + TN / TP + TN + FP + FN)
 - g. Recall (TP / TP + FN)
 - h. F1 score (2TP / 2TP + FP + FN)
4. να αποθηκεύονται τα αποτελέσματα σε αρχείο comma separated value (.csv)

2 Μεθοδολογία

Οι ταξινομητές (classifiers) που επιλέχθηκαν είναι οι:

- Logistic Regression
- k-Nearest Neighbors
- Gaussian Naive Bayes
- Support Vector

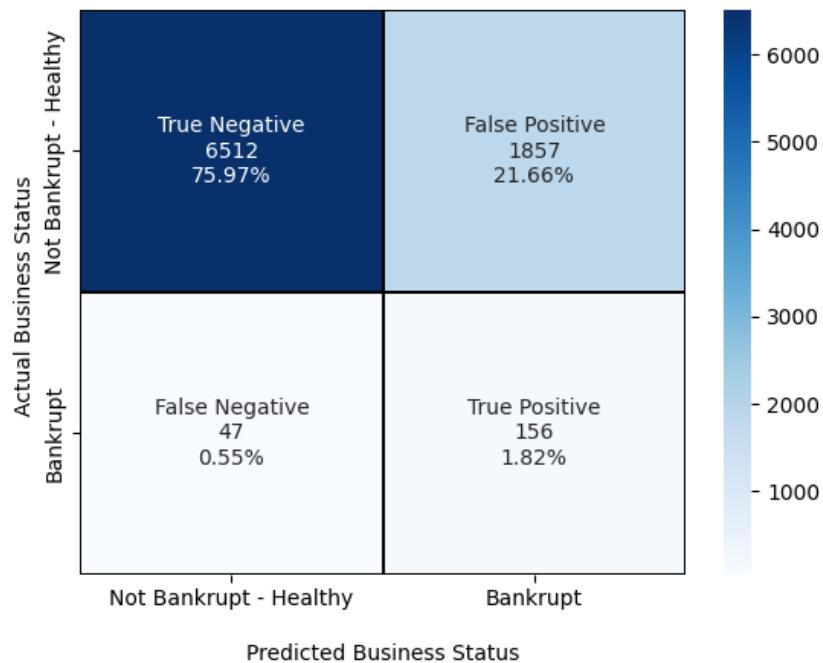
Ξεκινώντας, αφού εισαχθούν οι ταξινομητές από την βιβλιοθήκη sklearn, καθώς και οι υπόλοιπες απαραίτητες βιβλιοθήκες (numpy, matplotlib, seaborn, pandas), τυπώνονται τα αποτελέσματα τιμών υγείας, καθώς και το αν είναι υγιείς ή χρεοκοπημένες οι επιχειρήσεις του δείγματος μας από το έτος 2006 έως το έτος 2009, ανα έτος (τα πραγματικά δεδομένα δηλαδή).

Για την εκκίνηση εκπαίδευσης των μοντέλων, επιλέγεται αυθαίρετα αναλογία training - test set 80% - 20% των δεδομένων, αντίστοιχα. Στη συνέχεια, πραγματοποιείται κανονικοποίηση των δεδομένων στο διάστημα [0, 1].

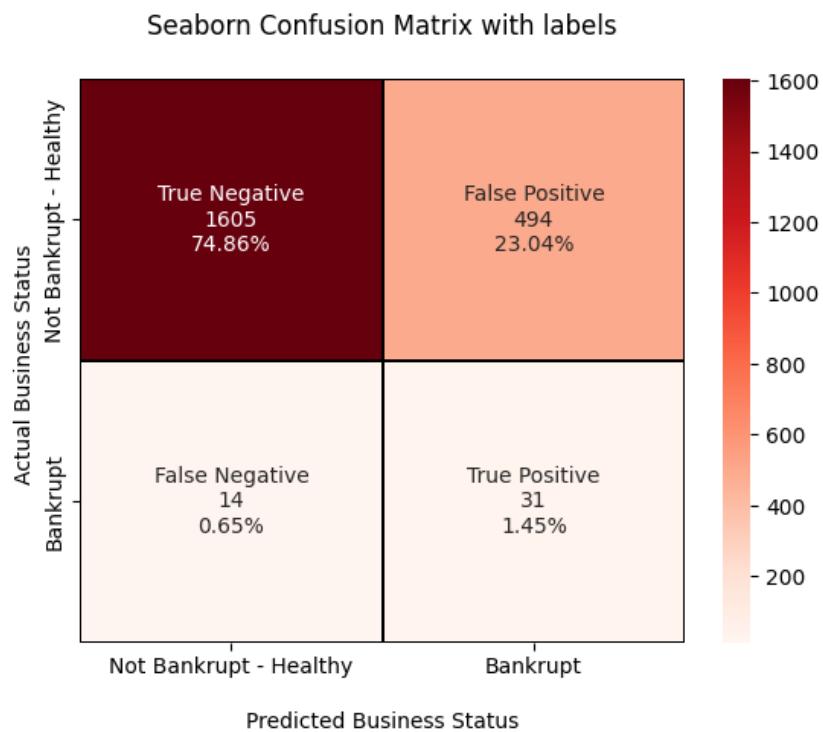
Εκπαιδεύεται πρώτος ο **logistic regression classifier** με class weight = balanced, διότι οι κλάσεις είναι άνισα κατανεμημένες. Τυπώνονται τα αποτελέσματα των τιμών Accuracy, Precision, Recall και F1, καθώς και οι μήτρες σφαλμάτων (confusion ή error matrices) των training και test sets, αντίστοιχα (χρησιμοποιείται η βιβλιοθήκη seaborn για τα γραφήματα):

- LogR Accuracy Score (Training Set): 0.7780
LogR Accuracy Score (Test Set): 0.7631
- LogR Precision Score (Training Set): 0.0775
LogR Precision Score (Test Set): 0.0590
- LogR Recall Score (Training Set): 0.7685
LogR Recall Score (Test Set): 0.6889
- LogR F1 Score (Training Set): 0.1409
LogR F1 Score (Test Set): 0.1088

Seaborn Confusion Matrix with labels



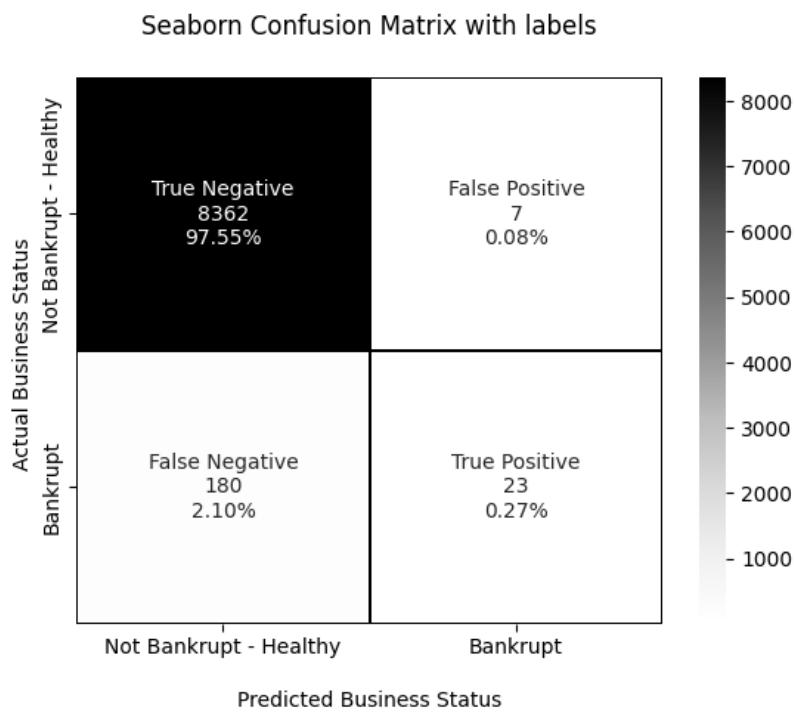
LogR Confusion Matrix (Training Set)



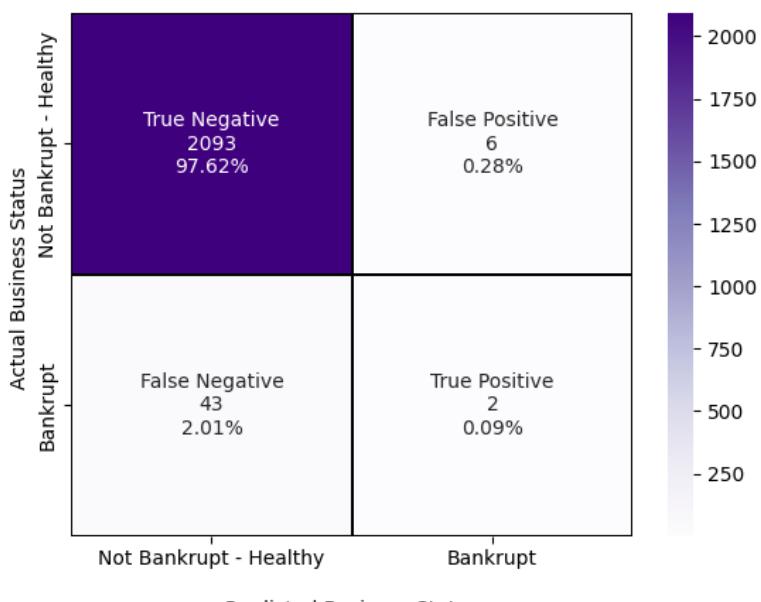
LogR Confusion Matrix (Test Set)

Ακολουθεί ο **kNN classifier** με αριθμό $k = 5$. Όπως και πριν, τυπώνονται τα αποτελέσματα των τιμών Accuracy, Precision, Recall και F1, καθώς και οι μήτρες σφαλμάτων (confusion ή error matrices) των training και test sets, αντίστοιχα:

- kNN Accuracy Score (Training Set): 0.9782
kNN Accuracy Score (Test Set): 0.9771
- kNN Precision Score (Training Set): 0.7667
kNN Precision Score (Test Set): 0.2500
- kNN Recall Score (Training Set): 0.1133
kNN Recall Score (Test Set): 0.0444
- kNN F1 Score (Training Set): 0.1974
kNN F1 Score (Test Set): 0.0755



Seaborn Confusion Matrix with labels

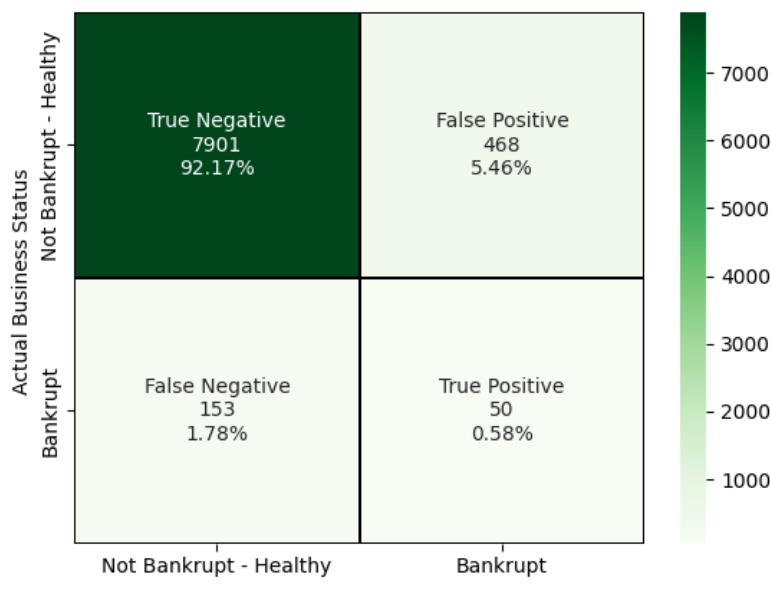


kNN Confusion Matrix (Test Set)

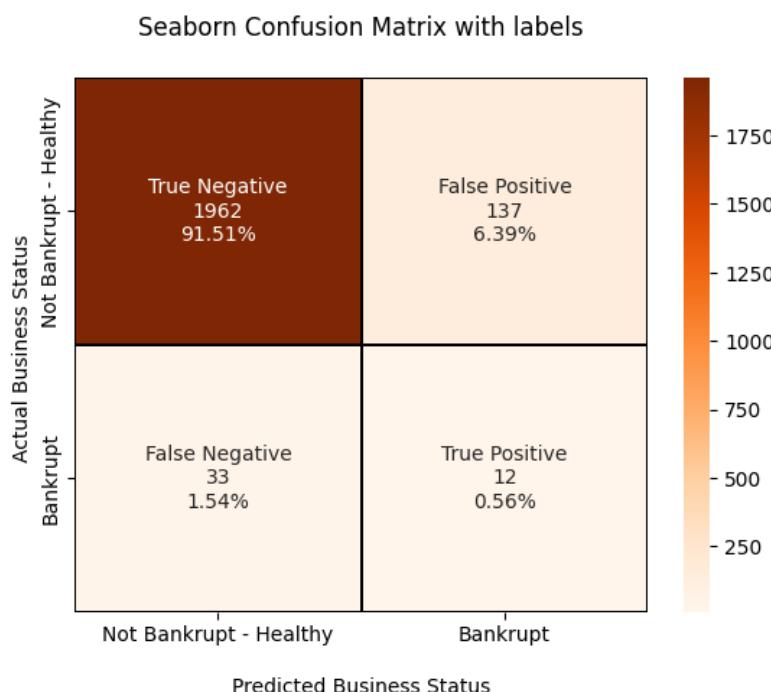
Στη συνέχεια, εκπαιδεύεται ο **Gaussian naive Bayes classifier**. Παρομοίως, τυπώνονται τα αποτελέσματα των τιμών Accuracy, Precision, Recall και F1, καθώς και οι μήτρες σφαλμάτων (confusion ή error matrices) των training και test sets, αντίστοιχα:

- Bayes Accuracy Score (Training Set): 0.9276
Bayes Accuracy Score (Test Set): 0.9207
- Bayes Precision Score (Training Set): 0.0965
Bayes Precision Score (Test Set): 0.0805
- Bayes Recall Score (Training Set): 0.2463
Bayes Recall Score (Test Set): 0.2667
- Bayes F1 Score (Training Set): 0.1387
Bayes F1 Score (Test Set): 0.1237

Seaborn Confusion Matrix with labels



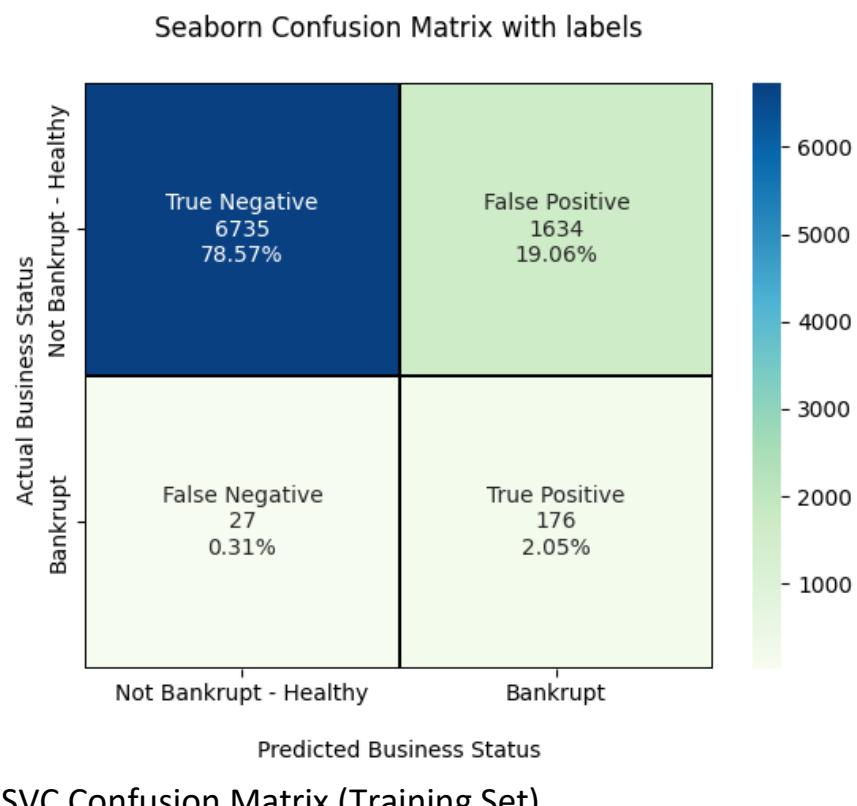
Bayes Confusion Matrix (Training Set)



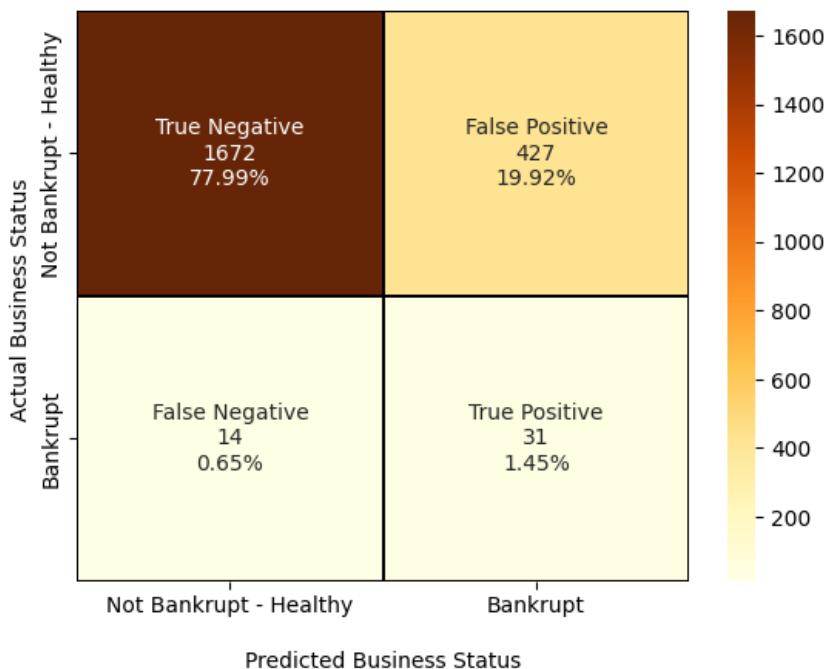
Bayes Confusion Matrix (Test Set)

Τέλος, ο **support vector classifier**. Για ακόμη μια φορά, τυπώνονται τα αποτελέσματα των τιμών Accuracy, Precision, Recall και F1, καθώς και οι μήτρες σφαλμάτων (confusion ή error matrices) των training και test sets, αντίστοιχα:

- SVC Accuracy Score (Training Set): 0.8062
SVC Accuracy Score (Test Set): 0.7943
- SVC Precision Score (Training Set): 0.0972
SVC Precision Score (Test Set): 0.0677
- SVC Recall Score (Training Set): 0.8670
SVC Recall Score (Test Set): 0.6889
- SVC F1 Score (Training Set): 0.1749
SVC F1 Score (Test Set): 0.1233



Seaborn Confusion Matrix with labels



SVC Confusion Matrix (Test Set)

Όλες οι τιμές που απέδωσαν οι διάφοροι classifiers τελικά καταγράφονται σε ένα .csv αρχείο.

Συμπεράσματα

Από τα αποτελέσματα των τιμών που παρήγαγαν οι classifiers, μπορούμε να εξάγουμε τα ακόλουθα:

- Ο **logistic regression classifier** αποδίδει τις καλύτερες τιμές **precision** (απόστασης μεταξύ παρατηρήσεων) στα training και test sets αλλά πολύ κακές τιμές **recall** (επιτυχούς ανάκλησης), κάτι που οδηγεί σε σχετικά μέτριες τιμές F1. Οι τιμές **accuracy** (ακρίβειας) είναι οι χειρότερες εκ των τεσσάρων classifiers (~77.7% στο training, ~76.3% στο test set).
- Ο **kNN classifier** αποδίδει τις καλύτερες τιμές **accuracy** στα training και test sets, με ελάχιστη διαφορά μεταξύ τους (97.8% και 97.7% αντίστοιχα). Οι τιμές **recall** είναι επίσης πολύ καλές, αλλά παρατηρείται μεγάλη χειροτέρευση στην τιμή του

test set σε σχέση με το training set. Πολύ μεγάλη διαφορά παρατηρείται επίσης στις τιμές **precision**, καθώς η τιμή στο training set είναι πολύ καλή αλλά χειροτερεύει πολύ στο test set. Συνεπώς, παρόμοια βελτίωση παρατηρείται στο F1 score, η τιμή του είναι η χειρότερη εκ των τεσσάρων classifiers στο training set αλλά η καλύτερη στο test set.

- Ο **Gaussian naive Bayes classifier** αποδίδει ισορροπημένα καλές τιμές σε όλες τις κατηγορίες, με τις μικρότερες πτώσεις αποτελεσμάτων στα test sets σε σχέση με τα αποτελέσματα των training sets, αλλά σε καμία κατηγορία δεν αποδίδει τις καλύτερες, με εξαίρεση την τιμή F1 στο test set, που είναι οριακά η καλύτερη.
- Ο **support vector classifier** αποδίδει πολύ καλές τιμές **precision** αλλά τις χειρότερες τιμές **recall**, κάτι που κάνει τις τιμές F1 μέτριες συνολικά. Παρατηρείται όμως αισθητή χειροτέρευση τους από τις τιμές του training set στις τιμές του test set. Οι τιμές **accuracy** είναι σχετικά κακές, ξεπερνάνε κατά λίγο μόνο τις τιμές του logistic regression classifier.

Οι μεγάλες διαφορές των αποτελεσμάτων μεταξύ training και test sets προκύπτουν κατά βάση από τον αυθαίρετο διαχωρισμό των συνόλων αυτών, καθώς επιλέγεται το πρώτο 80% των δεδομένων μας ως training set, και το υπόλοιπο 20% ως test set, χωρίς να εγγυόμαστε την αντιπροσωπευτικότητα των test δεδομένων με κάποιο τρόπο.

Με βάση τα αποτελέσματα F1 στο test set, οι ταξινομητές, από τον καλύτερο προς τον χειρότερο, είναι: **Gaussian naive Bayes > Support Vector > Logistic Regression > k-Nearest Neighbors**.