ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΜΣ «ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Μεθοδολογίες Μηχανικής Μάθησης για την Πρόγνωση Πιστωτικού Κινδύνου** **Machine Learning Methodologies for Credit Risk Prediction**
Ονοματεπώνυμο Φοιτητή	**Κωνσταντίνος Πληθάκης**
Πατρώνυμο	**Παναγιώτης**
Αριθμός Μητρώου	**ΜΠΠΛ17042**
Επιβλέπων	**Διονύσιος Σωτηρόπουλος, Καθηγητής**

ΠΕΙΡΑΙΑΣ

Σεπτέμβριος 2022

Τριμελής Εξεταστική Επιτροπή

Δ.Σωτηρόπουλος	Γ.Τσιχριντζής	Ε.Σακκόπουλος
Επίκουρος Καθηγητής	Καθηγητής	Αναπληρωτής Καθηγητής

Όνομα Επώνυμο (Name Surname)

Constantinos Plithakis

A.M.: MPPL17042

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΠΡΟΒΛΕΨΗ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ

1.1 Εισαγωγή

Από τον 20ο αιώνα και έπειτα στον τραπεζικό χώρο και τον χρηματοοικονομικό τομέα ξεκίνησαν να χρησιμοποιούνται μαθηματικές μέθοδοι για την επίλυση χρόνιων προβλημάτων που ταλάνιζαν την παγκόσμια αγορά, με αξιοσημείωτο παράδειγμα την αποφυγή πτώχευσης μιας εταιρείας ή/και ολόκληρου του χρηματιστηρίου. Με την ραγδαία ανάπτυξη της τεχνολογίας, της Πληροφορικής και των υπολογιστικών μεθόδων θεωρήθηκε απαραίτητο τα μέσα αυτά να ενταχθούν και να χρησιμοποιηθούν με βέλτιστο τρόπο στον χρηματοπιστωτικό κόσμο [1].

Στην Διπλωματική μας εργασία θα ασχοληθούμε με τον κλάδο της Τεχνητής Νοημοσύνης (Artificial Intelligence) και συγκεκριμένα της Μηχανικής και της Βαθιάς Μάθησης (Machine Learning and Deep Learning) στον χώρο της Οικονομίας και ειδικότερα, της Ανάλυσης Ρίσκου και διενέργειας προβλέψεων πάνω στο ζήτημα της χρηματοοικονομικής πίστωσης.

Τα πιο συνήθη τεχνολογικά εργαλεία που, πλέον, κατέχουν περίοπτη θέση στην Οικονομία, και σε όλους τους σύγχρονους κλάδους γενικότερα, και τα οποία χρησιμοποιούνται για να την ανάλυση και την πρόβλεψη χρηματοπιστωτικού κινδύνου (Credit Risk Analysis and Predictions) είναι οι αλγόριθμοι Ταξινόμησης Μηχανικής Μάθησης (Classification Machine Learning Algorithms) και τα Νευρωνικά Δίκτυα (Neural Networks). Μάλιστα, τις δύο αυτές τεχνικές θα αναπτύξουμε εκτενώς στη συνέχεια, με υλοποίηση αλγορίθμων σε πραγματικά δεδομένα για τη διενέργεια προβλέψεων.

Τέλος, για την υλοποίηση της του πρακτικού μέρους της έρευνας μας, θα χρησιμοποιήσουμε το προγραμματιστικό περιβάλλον της Python, η οποία στο παρόν αποτελεί την «νούμερο ένα» γλώσσα προγραμματισμού παγκοσμίως, ιδιαίτερα για θέματα και πειράματα στον τομέα της Ανάλυσης Δεδομένων και της Τεχνητής Νοημοσύνης.

1.2 Μια ιστορική αναδρομή

Οι πρώτες αναλύσεις, οι οποίες πραγματοποιήθηκαν, ιστορικά, στο σύγχρονο χρηματοπιστωτικό σύστημα για την ανάλυση ρίσκου και την πρόβλεψη της πιθανότητας πτώχευσης μιας εταιρείας χρησιμοποιούσαν στατιστικές μεθόδους και στατιστικά μοντέλα [2]. Συγκεκριμένα, αυτές οι μέθοδοι ήταν ποσοτικές και χρησιμοποιούσαν τυχαίες μεταβλητές για την πρόβλεψη πιθανής πτώχευσης ενός οργανισμού έως και λίγα χρόνια πριν να συμβεί αυτό [3].

Αργότερα, και συγκεκριμένα την τελευταία δεκαετία, ξεκίνησαν να εντάσσονται και να αναπτύσσονται μοντέλα Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης για την επεξεργασία και μοντελοποίηση τραπεζικών δεδομένων και τη διενέργεια προβλέψεων χρηματοπιστωτικού κινδύνου [3]. Τελευταία μάλιστα, παρατηρείται και μια άνοδος στην κατασκευή νευρωνικών δικτύων και μοντέλων βαθιάς μάθησης στην οικονομία και τις τράπεζες [4].

Η ειδοποιός διαφορά μεταξύ των πρώτων και των δεύτερων μεθόδων που αναπτύχθηκαν και προσαρμόστηκαν στα τραπεζικά δεδομένα αποτελεί το γεγονός πως οι αλγόριθμοι Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης κατείχαν μεγαλύτεροι υπολογιστική ικανότητα στην προσομοίωση των δεδομένων, πληθώρα παραμέτρων και μαθηματικών αλγορίθμων με αποτέλεσμα την αύξηση της παραγωγικότητας των πειραμάτων – μοντέλων και την πιο έγκαιρη και ακριβή δημιουργία προβλέψεων [3].

Ένα τρανταχτό παράδειγμα αυτής της σημαντικής διαφοράς μεταξύ των παραδοσιακών στατιστικών μεθόδων και αυτών των σύγχρονων – εξελιγμένων, αποτελεί η περίπτωση της πτώχευσης του οργανισμού «Lehman Brothers Holdings Inc.». Από τους παγκόσμιους αναλυτές και επιστήμονες έχει εξαχθεί το συμπέρασμα πως αν υπήρχε ένας αξιόπιστος και αποτελεσματικός αλγόριθμος ανάλυσης του χρηματοπιστωτικού κινδύνου, όπως αυτοί που αναπτύσσονται από τις τράπεζες σήμερα, η εταιρεία θα είχε καταφέρει να αποφύγει την πτώχευση [5].

1.3 Χρηματοπιστωτικό Σκορ

Το χρηματοπιστωτικό σκορ και η βαθμολόγηση του βασίζεται στους αλγορίθμους ταξινόμησης Μηχανικής Μάθησης [6]. Το χρηματοπιστωτικό σκορ αποτελεί, σήμερα, τον σημαντικότερο δείκτη αξιολόγησης και σύγκρισης μοντέλων που σχετίζονται με την οικονομία, το χρηματιστήριο και τις τράπεζες όσον αφορά τις προβλέψεις ρίσκου, πτώχευσης τιμών κλπ. [7].

Ειδικότερα, για το πρόβλημα μας, την πρόβλεψη χρηματοπιστωτικού ρίσκου τα credit scores βασίζονται στα αποτελέσματα των μοντέλων ταξινόμησης (classification) μηχανικής μάθησης και τα νευρωνικά δίκτυα (πάλι) για ταξινόμηση [8]. Σε αυτές τις περιπτώσεις η μεταβλητή πρόβλεψης σε ένα σύνολο δεδομένων που μελετάμε είναι ένα δυαδικό ποιοτικό χαρακτηριστικό με τιμές 0 και 1 ή ναι και όχι (yes or no) και αφορά για παράδειγμα αν ένας υποψήφιος δανειολήπτης είναι αξιόπιστος ή όχι για να πάρει δάνειο, αν ένας οργανισμός θα πτωχεύσει ή όχι και άλλα [9].

Ενδεικτικά, κάποιοι από τους πιο γνωστούς και ευρέως διαδεδομένους αλγορίθμους ταξινόμησης αποτελούν τα μοντέλα: K-Nearest Neighbors (KNN) – Κ εγγύτεροι γείτονες, Support Vector Machines (SVM) – Μηχανές Διανυσμάτων Υποστήριξης, Λογιστική Παλινδρόμηση (Logistic Regression) και Δέντρα Απόφασης (Decision Trees) [9], [10].

Φυσικά, υπάρχουν και άλλοι χρηματιστηριακοί δείκτες αξιολόγησης μοντέλων ανάλυσης κινδύνου ή πτώχευσης, όπως το NPL, δηλαδή τα μη εξυπηρετούμενα δάνεια. Αυτό το μοντέλο σχετίζεται με την πρόβλεψη του δανείου εκείνου από έναν (υποψήφιο) πελάτη το οποίο τείνει να είναι μη εξυπηρετούμενο και έτσι να αποφευχθεί η χορήγηση του αντιμετωπίζοντας κατάλληλα την κατάσταση [11]. Ωστόσο, όπως είναι φανερό από την βιβλιογραφία και τα παγκόσμια τεκταινόμενα, οι αλγόριθμοι Τεχνητής Νοημοσύνης είναι αυτοί που έρχονται πρώτοι, τόσο στην χρήση, όσο και στην επιτυχία προβλέψεων [12].

1.4 Ανίχνευση απάτης

Ένα άλλο σημαντικό κεφάλαιο, πέραν της ανάλυσης χρηματοπιστωτικού ρίσκου σε ένα πρότζεκτ, αποτελεί η ανίχνευση απάτης (Fraud detection) σε τράπεζες και σε άλλα οικονομικά δεδομένα. Συγκεκριμένα, οι αλγόριθμοι Μηχανικής Μάθησης και τα Νευρωνικά Δίκτυα χρησιμοποιούνται τα τελευταία χρόνια σε πολλούς οργανισμούς για τον έλεγχο της αξιοπιστίας των πελατών και των δεδομένων τους με σκοπό την αποφυγή εξαπάτησης από αυτούς, αλλά και από άλλους κακόβουλους παράγοντες [13], [14].

2 ΜΕΘΟΔΟΛΟΓΙΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΠΡΟΒΛΕΨΗ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ

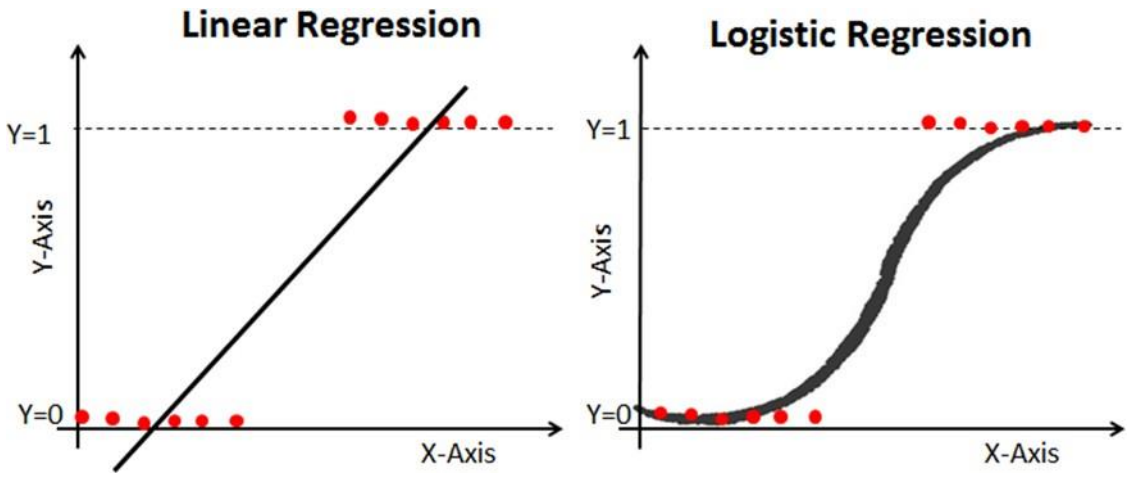
Παρακάτω παρουσιάζουμε μερικούς από τους βασικότερους αλγόριθμους μηχανικής μάθησης για προβλήματα ταξινόμησης.

2.1 Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση είναι ένας αλγόριθμος μηχανικής μάθησης ταξινόμησης, που χρησιμοποιείται όταν η τιμή της μεταβλητής στόχου είναι κατηγορηματικής φύσης, δηλαδή δεν εκφράζει ποσότητα αλλά κάποιο ποιοτικό χαρακτηριστικό [15]. Μάλιστα, χρησιμοποιείται σχεδόν πάντα, όταν τα δεδομένα μας, έχουν δυαδική - δίτιμη έξοδο, δηλαδή όταν η μεταβλητή πρόβλεψης – στόχου (target - class variable) ανήκει σε μια τάξη εκ των δύο τάξεων (0 ή 1 – yes or no) [16].

Το αποτέλεσμα προσδιορίζεται χάρη στη χρήση μιας λογαριθμικής συνάρτησης (όχι γραμμική), συνήθως τη σιγμοειδή (ή και κάποια διαφορετική), η οποία εκτιμά μια πιθανότητα και στη συνέχεια καθορίζει την πλησιέστερη τάξη (θετική ή αρνητική- 1 ή 0) στην τιμή πιθανότητας που έχει ληφθεί [16].

Μπορούμε να θεωρήσουμε τη λογιστική παλινδρόμηση ως μέθοδο ταξινόμησης της οικογένειας των εποπτευόμενων αλγόριθμων μάθησης (supervised learning algorithms). Χρησιμοποιώντας στατιστικές μεθόδους, η λογιστική παλινδρόμηση επιτρέπει τη δημιουργία ενός αποτελέσματος το οποίο, στην πραγματικότητα, αντιπροσωπεύει την πιθανότητα ότι μια δεδομένη τιμή εισόδου ανήκει σε μια δεδομένη κατηγορία από την έξοδο [17].



Εικόνα 1: Σύγκριση γραμμικής με λογιστική παλινδρόμηση [18]

2.2 KNN (K-Nearest Neighbors)

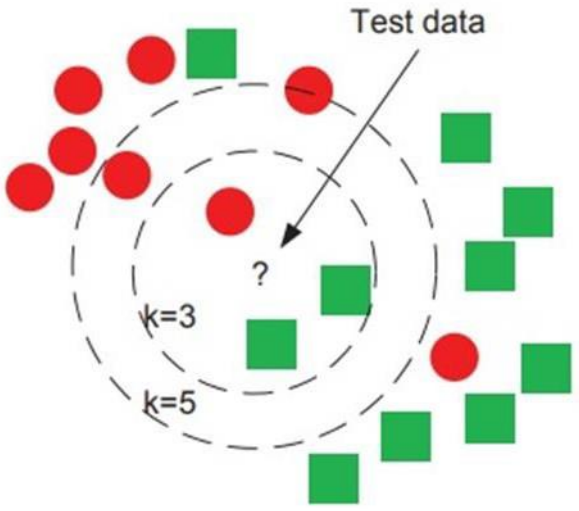
Οι Κ εγγύτεροι - πλησιέστεροι γείτονες (K Nearest Neighbors) είναι ένας αλγόριθμος που χρησιμοποιείται, ευρέως, για προβλήματα ταξινόμησης και παλινδρόμησης [19]. Ο αλγόριθμος λειτουργεί επαναληπτικά, καθώς αποθηκεύει βασικά όλες τις διαθέσιμες μεταβλητές για να ταξινομήσει τις νέες τιμές που έχουν κατά πλειοψηφία τις περισσότερες κοινές από τους γείτονες τους [20].

Η συνάρτηση απόστασης που χρησιμοποιεί ο αλγόριθμος αποτελεί την σημαντικότερη παράμετρο του και είναι μια από τις ακόλουθες μετρικές [21]:

1) Euclidean
2) Manhattan
3) Minkowski
4) Chebyshev
5) Hamming

Ενώ οι τρεις πρώτες συναρτήσεις απόστασης χρησιμοποιούνται για συνεχείς μεταβλητές, η συνάρτηση απόστασης Hamming χρησιμοποιείται για ποιοτικά χαρακτηριστικά [21].

Το σημαντικότερο βήμα υλοποίησης του αλγορίθμου KNN είναι η εύρεση του καταλληλότερου και αποδοτικότερου K που θα μεγιστοποιεί την ακρίβεια του μετέπειτα κατασκευαζόμενου μοντέλου (δηλαδή πόσοι εγγύτεροι γείτονες θα υπάρχουν). Γενικά, αυτός ο αλγόριθμος έχει μεγάλο υπολογιστικό κόστος. Πέρα, από την δημιουργία μοντέλων μηχανικής μάθησης με KNN, ο αλγόριθμος αυτός πολλές φορές χρησιμοποιείται και για άλλους σκοπούς, όπως για παράδειγμα η μείωση των διαστάσεων των δεδομένων, η εύρεση απομακρυσμένων και μη επιθυμητών τιμών (outliers) και άλλα [22].



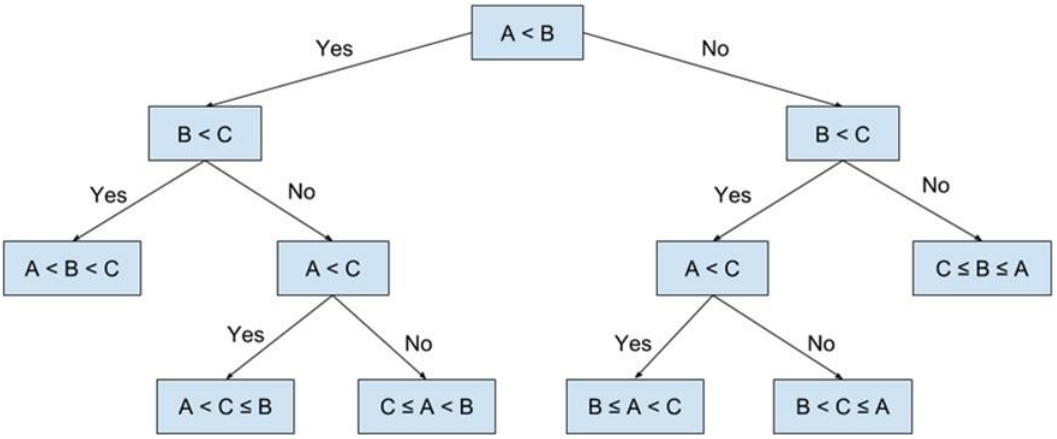
Εικόνα 2: Παράδειγμα KNN μοντελοποίησης [22]

2.3 Δέντρα Απόφασης (Decision Trees)

Ένας άλλος σημαντικός αλγόριθμος ταξινόμησης είναι τα δέντρα απόφασης. Τα δέντρο απόφασης είναι μια δομή κόμβων και ακμών - με βάση την οποία ο πληθυσμός ταξινομείται χρησιμοποιώντας μια επεξηγηματική μεταβλητή (xi) σε κάθε κόμβο και λαμβάνοντας μια απόφαση σχετικά με τις διάφορες επιλογές διαχωρισμού των υπόλοιπων μεταβλητών – κόμβων. Η κορυφή του δέντρου είναι ο κόμβος ρίζα του μοντέλου, ενώ τα επόμενα επίπεδα κόμβων είναι οι κόμβοι παιδιά. Στο τελευταίο επίπεδο του δέντρου βρίσκονται οι τερματικοί κόμβοι που περιγράφουν την τελική απόφαση – αποτέλεσμα, δηλαδή την τελική ταξινόμηση [23].

Ο τρόπος με τον οποίο μεταπηδούμε διαδοχικά από την ρίζα του δέντρου στους τερματικούς κόμβους του μοντέλου καθορίζεται από μια πληθώρα κριτηρίων που αποτελούν

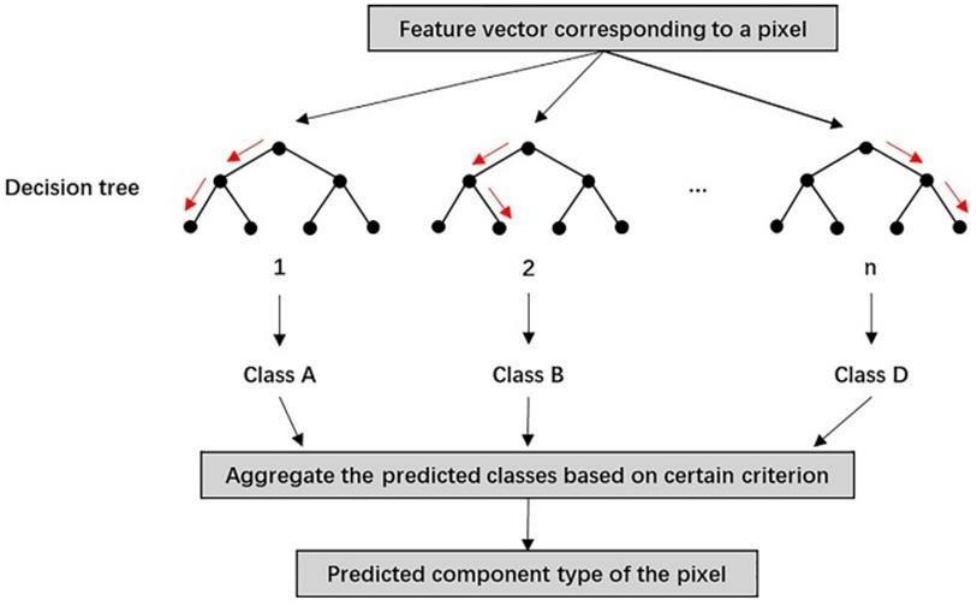
τις παραμέτρους του αλγορίθμου. Δηλαδή, η κάθε απόφαση που συντελείται εξαρτάται από την πολυπλοκότητα, τις παραμέτρους και της φύσης των δεδομένων που χρησιμοποιεί το μοντέλο και δεν ακολουθείται κάποια συγκεκριμένη ακολουθία αποφάσεων [24].



Εικόνα 3: Θεωρητικό παράδειγμα δομής δέντρου απόφασης [24]

2.4 Τυχαίο δάσος (Random Forest)

Ο αλγόριθμος Random Forest αποτελεί έναν αλγόριθμο εποπτευόμενης μάθησης που χρησιμοποιείται συχνά, τόσο για προβλήματα ταξινόμησης, όσο και σε προβλήματα παλινδρόμησης. Η γενική ιδέα αυτού του αλγορίθμου είναι πως η απόφαση που παίρνει σε κάθε επιλογή πραγματοποιείται βάση πολλαπλών αποφάσεων από πολλά δέντρα απόφασης συγχρόνως. Έτσι, συγχωνεύοντας πολλά δέντρα μαζί, το μοντέλο που κατασκευάζεται είναι ικανό να παράγει πιο ακριβείς και συνεπείς λύσεις – προβλέψεις, αφού αξιολογεί μεγάλη πληθώρα επιλογών και αποφάσεων για να εξάγει την καλύτερη κάθε φορά [25].



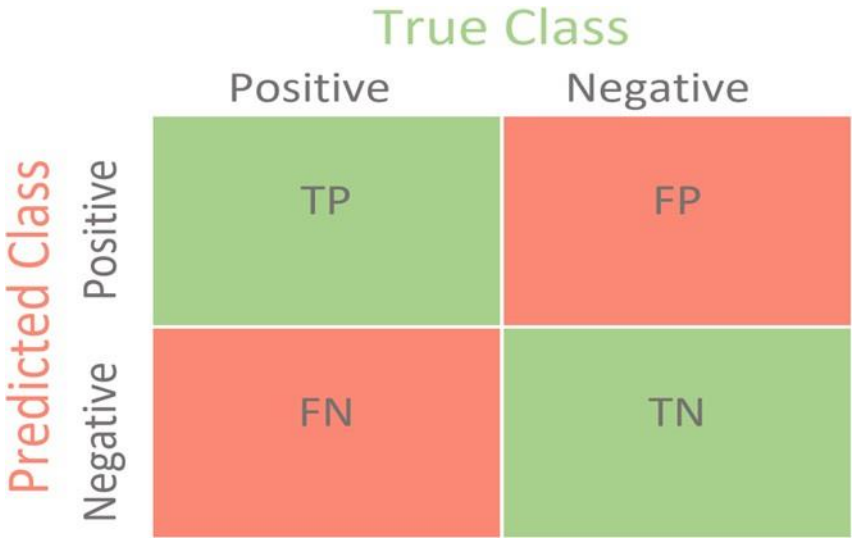
Εικόνα 4: Οπτικοποίηση γενικού μοντέλου «τυχαίου δάσους» [25]

2.5 Αξιολόγηση αποτελεσμάτων μοντέλων (Model Evaluation)

Για την αξιολόγηση και την σύγκριση των αποτελεσμάτων των προβλέψεων για κάθε ένα από τα μοντέλα ταξινόμησης που κατασκευάζονται χρησιμοποιούνται οι κάτωθι μετρικές αξιολόγησης:

1) Πίνακας Σύγχυσης (Confusion Matrix)

Ο πίνακας σύγχυσης αποτελεί το πιο διαδεδομένο μέτρο ελέγχου και ακρίβειας ενός μοντέλου μηχανικής μάθησης για ταξινόμηση. Σε ένα πρόβλημα δυαδικής ταξινόμησης ο πίνακας αυτός είναι ένας 2 x 2 πίνακας, όπου οι στήλες παρουσιάζουν τις παρατηρήσεις που βρίσκονται στην πραγματική τάξη τους (0 ή 1), ενώ οι γραμμές αντιπροσωπεύουν τις προβλεπόμενες κλάσεις των τιμών σύμφωνα με το εκάστοτε μοντέλο [26]. Η απεικόνιση αυτού του πίνακα είναι ως εξής:



Εικόνα 5: Απεικόνιση πίνακα σύγχυσής [26]

Οι συμβολισμοί και η ερμηνεία τους σε κάθε επιμέρους «κουτί» του πίνακα είναι:

- *TP (True Positive):* για κάθε παρατήρηση που προβλέφθηκε (ταξινομήθηκε) ότι ανήκει στην θετική ή 1η τάξη και πραγματικά ανήκε σε αυτή.
- *FP (False Positive):* για κάθε παρατήρηση που προβλέφθηκε ότι ανήκει στην 1η τάξη αλλά πραγματικά ανήκε στην αρνητική ή μηδενική
- *FN (False Negative):* για κάθε παρατήρηση που ταξινομήθηκε στην αρνητική κλάση (μηδενική), αλλά στην πραγματικότητα ανήκε στην θετική – 1η
- *TN (True Negative):* για κάθε παρατήρηση που ταξινομήθηκε στην μηδενική τάξη και όντως ανήκε σε αυτήν

2) Μετρικές μοντέλου ταξινόμησης (Classification Report)

Οι παρακάτω τέσσερις μετρικές, οι οποίες όλες μαζί δημιουργούν μια ολοκληρωμένη αναφορά για το μοντέλο ταξινόμησης (classification report) και υπολογίζονται από μαθηματικούς τύπους σύμφωνα και με τα στοιχεία του πίνακα σύγχυσης [26]:

- Precision (Ακρίβεια):

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}\ or\ \frac{True\ Positive}{Total\ Predicted\ Positive}$$

- Recall (Ανάκληση):

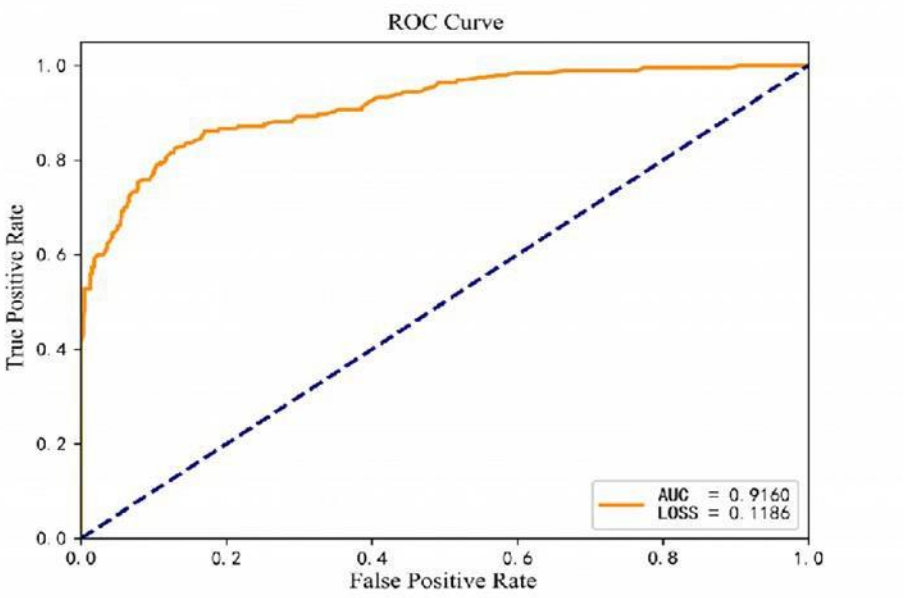
$$Recall = \frac{True\ Positive}{True\ Positive + False\ negative}\ or\ \frac{True\ Positive}{Total\ Actual\ Positive}$$

- F1 score (Ακρίβεια F1):

$$F1 = 2 * Precision * \frac{Precision * Recall}{Precision + Recall}$$

3) ROC – AUC score και ROC curve

Η ROC καμπύλη και το ROC – AUC score αποτελούν και αυτά με τη σειρά τους δύο πολύ χρήσιμα μέτρα αξιολόγησης και ακρίβειας των προβλέψεων σε προβλήματα ταξινόμησης μηχανικής μάθησης. Ωστόσο, εδώ η ακρίβεια έχει να κάνει με το πόσο δυνατό ένα μοντέλο είναι, και ικανό, στο να ξεχωρίζει καλύτερα τις δύο τάξεις της δυαδικής μεταβλητής πρόβλεψης. Όσο υψηλότερο είναι το σκορ της καμπύλης, τόσο καλύτερο και πιο αποδοτικό θα είναι το μοντέλο. Τέλος, στο γράφημα της καμπύλης οι τιμές στους δύο άξονες δίνονται από το False Positive Rate – Ποσοστό τιμών FP (x άξονας) και True Positive Rate – Ποσοστό τιμών TP (y άξονας) [27].



Εικόνα 6: Παράδειγμα ROC καμπύλης [27]

3 ΜΕΘΟΔΟΛΟΓΙΕΣ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ

3.1 Βαθιά μάθηση (Deep learning)

Η βαθιά μάθηση (deep learning) αποτελεί έναν σύγχρονο κλάδο της Τεχνητής Νοημοσύνης με πληθώρα εφαρμογών και εργασιών. Το deep learning έχει πρακτικές εφαρμογές και ερευνητικά πρότζεκτ, τόσο όσον αφορά την εποπτευόμενη μάθηση, όσο και την μη εποπτευόμενη [28].

Στα τέλη της δεκαετίας του 1980, τα νευρωνικά δίκτυα (Neural Networks)έγιναν ένα διαδεδομένο θέμα στον τομέα της Μηχανικής Μάθησης (ML) καθώς και της Τεχνητής νοημοσύνης (AI), λόγω της εφεύρεσης διαφόρων αποτελεσματικών μεθόδων μάθησης και δομών δικτύων [29].

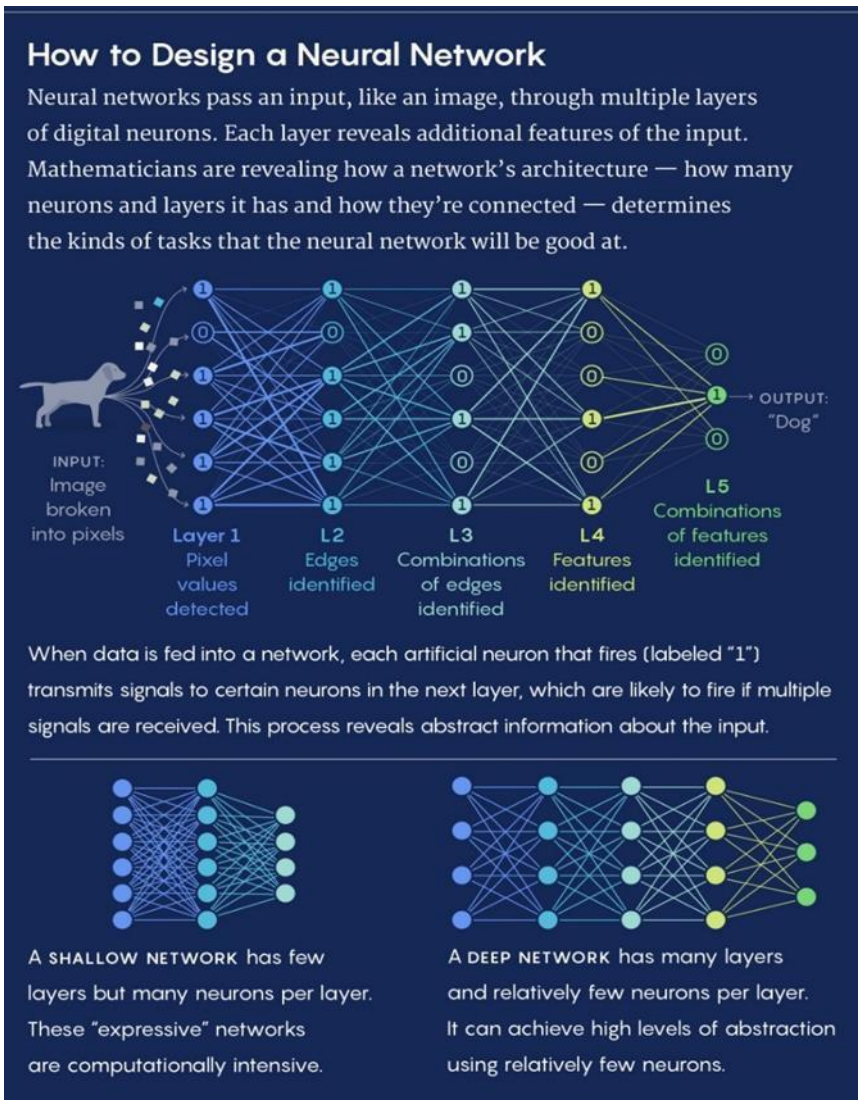
Χαρακτηριστικοί αλγόριθμοι που χρησιμοποιούνται κλασικά για την κατασκευή και την υλοποίηση νευρωνικών δικτύων αποτελούν οι Backpropagation αλγόριθμοι και αυτό-οργανωμένες αρχιτεκτονικές δικτύων που είναι πολυεπίπεδα (multilayer) [30]. Η βασική διαφορά των μοντέλων των νευρωνικών δικτύων με τις υπόλοιπες κλασικές μεθόδους μηχανικής μάθησης είναι πως λειτουργούν όπως ένας νευρώνας και κατ’ επέκταση ένας ανθρώπινος οργανισμός, δηλαδή το δίκτυο έχοντας κάποια ιδιαίτερα χαρακτηριστικά και συγκεκριμένη μορφολογία παίρνοντας στην είσοδο του τα εκάστοτε δεδομένα, μαθαίνει μόνο του πως να τα χειριστεί, ώστε να κάνει τις επιθυμητές κάθε φορά προβλέψεις [1], [31].

Η έννοια της βαθιάς μάθησης (deep) εισήχθη στην επιστημονική κοινότητα το 2006 και παραπέμπει στην έννοια του τεχνητού νευρωνικού δικτύου (Artificial Neural Network - ANN) [32], [33]. Η αντίστοιχη θεωρία έβαλε τα θεμέλια για την αναγέννηση της έρευνας πάνω στα νευρωνικά δίκτυα, με αποτέλεσμα από τότε και ύστερα να αναπτύσσονται τα νευρωνικά δίκτυα νέας γενιάς. Είναι αξιοσημείωτο δε, πως οι αρχιτεκτονικές και οι μεθοδολογίες που αναπτύσσονται για τα δίκτυα αυτά εξελίσσονται συνεχώς, με αποτέλεσμα συνέχεια να ξεπηδούν καινούριοι αλγόριθμοι και βελτιστοποιημένα μοντέλα [34].

Οι πρακτικές εφαρμογές των νευρωνικών δικτύων είναι πολλαπλές και εξίσου σημαντικές με χαρακτηριστικά παραδείγματα την δημιουργία μοντέλων ταξινόμησης και παλινδρόμησης σε πολύ μεγαλύτερο όγκο δεδομένων σε σύγκρισης με τις τεχνικές της απλής Μηχανικής Μάθησης, την επεξεργασία και μοντελοποίηση εικόνων και βίντεο, την ανίχνευση πλαστά κατασκευασμένων εικόνων ή και την δημιουργία τους και άλλα [1]. Χαρακτηριστικό παράδειγμα αποτελεί το σύστημα ανίχνευσης εικόνων και αυτόματου οδηγού των ηλεκτροκίνητων αυτοκινήτων της “Tesla”, το οποίο βασίζεται κατά κύριο λόγο σε εξειδικευμένα και πολύπλοκα νευρωνικά δίκτυα [35].

Μερικές από τις βασικότερες γνωστές αρχιτεκτονικές και μεθοδολογίες νευρωνικών δικτύων αποτελούν οι εξής [36]:

- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Artificial Neural Networks (ANN)
- Long - Short Term Memory Networks (LSTM) etc.



Εικόνα 7: Παράδειγμα σχεδίασης νευρωνικού δικτύου με πολλαπλά layer (επίπεδα) για ταξινόμηση – πρόβλεψη εικόνας [37]

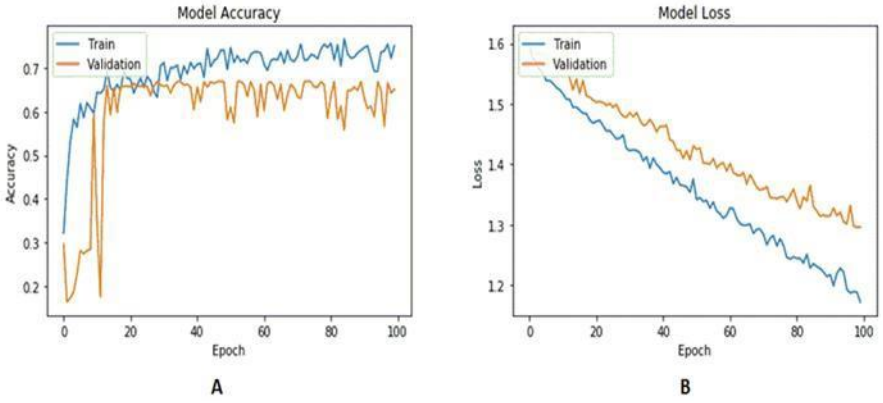
3.2 Αξιολόγηση μοντέλων Νευρωνικών Δικτύων για προβλήματα ταξινόμησης

Στα προβλήματα Τεχνητής Νοημοσύνης, στα οποία ο σκοπός είναι η πρόβλεψη της τάξης μια δυαδικής μεταβλητής μέσω διαδοχικής ταξινόμησης (binary classification), και όπου τα μοντέλα που αναπτύσσονται και εξυπηρετούν τον στόχο είναι νευρωνικά δίκτυα, η αξιολόγηση πραγματοποιείται σε δύο στάδια [38].

Το 1ο στάδιο είναι ήδη γνωστό και αποτελεί την αξιολόγηση με τις ίδιες μετρικές που χρησιμοποιούμε για τα αποτελέσματα στα μοντέλα ταξινόμησης στην Μηχανική Μάθηση. Δηλαδή, και εδώ, μπορούν να υπολογιστούν οι μετρικές Precision, Recall, F1-Score [38] και να κατασκευαστεί ο αντίστοιχος πίνακας σύγχυσης (Confusion Matrix) για περαιτέρω ανάλυση (που ουσιαστικά περιλαμβάνει και υπολογίζει τα υπόλοιπα σκορ) [39].

Το 2ο στάδιο, το οποίο συναντάται μόνο στα νευρωνικά δίκτυα είναι η αξιολόγηση μέσω των καμπύλων μάθησης. Υπάρχουν δύο καμπύλες μάθησης (learning curves): η καμπύλη μάθησης για την ακρίβεια του εκάστοτε μοντέλο (learning curve of accuracy of a model) και η καμπύλη μάθησης για την απώλεια (δεδομένων - στοιχείων) (learning curve of loss). Και οι δύο αυτές καμπύλες έχουν στους δύο άξονες τα σύνολα εκπαίδευσης δεδομένων και ελέγχου (training and test sets) και αναπαριστούν την συσχέτιση μεταξύ τους κατά την

διενέργεια μιας πρόβλεψης που κάνουμε σε ένα μοντέλο Τεχνητής Νοημοσύνης. Επιπλέον, μέσα από αυτά τα γραφήματα μπορεί να εξαχθούν συμπεράσματα για το αν ένα μοντέλο κάνει overfitting (ύπερ - προσαρμογή) ή underfitting (ύπο - προσαρμογή) - και τα δύο αυτά φαινόμενα καλό είναι να αποφεύγονται [40].



Εικόνα 8: Παράδειγμα καμπυλών μάθησης και σύγκριση μεταξύ τους [41]

Να σημειώσουμε εδώ ότι εποχές (Epochs) προσδιορίζουν τον αριθμό των επαναλήψεων που θα συντελείται ένας αλγόριθμος νευρωνικού δικτύου έως ότου «σταματήσει» και εξάγει κάποιο επιθυμητό ή μη αποτέλεσμα (πρόβλεψη) [42].

4 ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ

4.1 Το σύνολο δεδομένων

Στην συνέχεια θα αναλύσουμε και θα κατασκευάσουμε μοντέλα μηχανικής και βαθιάς μάθησης, ώστε να ερευνήσουμε στην πράξη πως λειτουργούν οι αλγόριθμοι τεχνητής νοημοσύνης στην πράξη. Επίσης, θα συγκρίνουμε τα αποτελέσματα μας, ώστε να εξάγουμε το καλύτερο μοντέλο και μερικά χρήσιμα συμπεράσματα.

Το σύνολο δεδομένων που χρησιμοποιούμε είναι το «Credit Risk Dataset» και είναι δημόσια διαθέσιμο για κάθε ερευνητή και επιστήμονα να το επεξεργαστεί και να το αναλύσει με οποιονδήποτε τρόπο αυτός κρίνει, στην διαδικτυακή βάση συνόλων δεδομένων Kaggle (https://www.kaggle.com/laotse/credit-risk-dataset).

Αποτελεί ένα σύνολο δεδομένων με 32581 υποψήφιους δανειολήπτες, δηλαδή 32581 άτομα τα οποία ερευνάται αν θα έχουν ή όχι χορηγηθεί δάνειο από μια τράπεζα. Με άλλα λόγια πραγματοποιείται ανάλυση ρίσκου ώστε να προβλεφθεί αν ένας υποψήφιος πελάτης είναι αξιόπιστος στο να πάρει δάνειο ή όχι. Τα δεδομένα μας έχουν 12 μεταβλητές, οι οποίες αποτελούν τα χαρακτηριστικά των ατόμων και εκ των οποίων οι 11 αφορούν διάφορα προσωπικά χαρακτηριστικά του κάθε (υποψήφιου) δανειολήπτη, ενώ η 12η μεταβλητή είναι η μεταβλητή πρόβλεψης ή μεταβλητή – στόχος και αφορά την κατάσταση δανείου (loan_status) με τιμές 0 και 1 (0: non default – 1: default). Ακριβέστερα, η μεταβλητή πρόβλεψης του συνόλου δεδομένων μας σχετίζεται με την αξιοπιστία των δανειοληπτών – πελατών και είναι ο κίνδυνος της πιστωτικής αθέτησης εκ μέρους τους, δηλαδή ο κίνδυνος που αναλαμβάνει ο δανειστής (τράπεζα) στην περίπτωση που ο δανειολήπτης δεν δύναται να πραγματοποιήσει τις απαιτούμενες πληρωμές του δανείου που κατέχει. Η τιμή 0: non-default μεταφράζεται στο γεγονός πως ο δανειολήπτης είναι αξιόπιστος και θα πραγματοποιήσει τις απαιτούμενες πληρωμές του δανείου και η τιμή 1: default αναφέρεται στους μη-αξιόπιστους πελάτες.

Στους παρακάτω πίνακες παρουσιάζουμε όλες τις μεταβλητές του συνόλου δεδομένων μας, την ερμηνεία τους και τον τύπο δεδομένων τους:

Πίνακας Ι

Όνομα μεταβλητής	**Περιγραφή**
person_age	η ηλικία (σε χρόνια)
person_income	το ετήσιο εισόδημα
person_home_ownership	ο τρόπος κατοίκισης (π.χ. ενοικίαση)
person_emp_length	έτη επαγγελματικής κατάρτισης
loan_intent	ο τύπος – είδος δανείου
loan_grade	ο βαθμός δανείου
loan_amnt	το χρηματικό ποσό του δανείου
loan_int_rate	οι τόκοι του δανείου
loan_status	η κατάσταση – πρόθεση του δανειολήπτη όσον αφορά το δάνειο

Όνομα μεταβλητής	Περιγραφή
loan_percent_income	ποσοστιαίο ετήσιο εισόδημα
cb_person_default_on_file	ιστορικό δανειολήπτη όσον αφορά την μέχρι τώρα αξιοπιστία του(ναι ή όχι)
cb_preson_cred_hist_length	χρονικό πιστωτικής διάρκειας στο παρελθόν (χρόνια)

Πίνακας ΙΙ

Όνομα μεταβλητής	Περιγραφή
person_age	numeric - integer
person_income	numeric - integer
person_home_ownership	categorical
person_emp_length	numeric - float
loan_intent	categorical
loan_grade	categorical
loan_amnt	numeric - integer
loan_int_rate	numeric - float
loan_status	categorical
loan_percent_income	numeric - float
cb_person_default_on_file	categorical
cb_preson_cred_hist_length	numeric - integer

Όπως παρατηρούμε το σύνολο δεδομένων περιέχει 5 κατηγορικές – ποιοτικές μεταβλητές, ενώ τα υπόλοιπα 7 χαρακτηριστικά είναι ποσοτικά – αριθμητικά (4 ακέραια και 3 δεκαδικά).

4.2 Εργαλεία Υλοποίησης

Για να υλοποιήσουμε την ανάλυση και μοντελοποίηση του συνόλου δεδομένων που έχουμε με την δημιουργία προβλέψεων με διάφορες μεθόδους τεχνητής νοημοσύνης χρησιμοποιήσαμε την γλώσσα προγραμματισμού Python και συγκεκριμένα το γραφικό περιβάλλον της Google, Google Collaboratory.

Η Python είναι η καλύτερη και πιο αποτελεσματική γλώσσα προγραμματισμού στον κόσμο αυτή την στιγμή. Είναι ιδιαίτερα γνωστή κυρίως για πρότζεκτ δημιουργίας εφαρμογών, ιστοσελίδων, παιχνιδιών και ιδιαίτερα για ανάλυση και επεξεργασία δεδομένων και την έρευνα και ανάπτυξη ποικίλων μεθόδων στον κλάδο της τεχνητής νοημοσύνης [43].

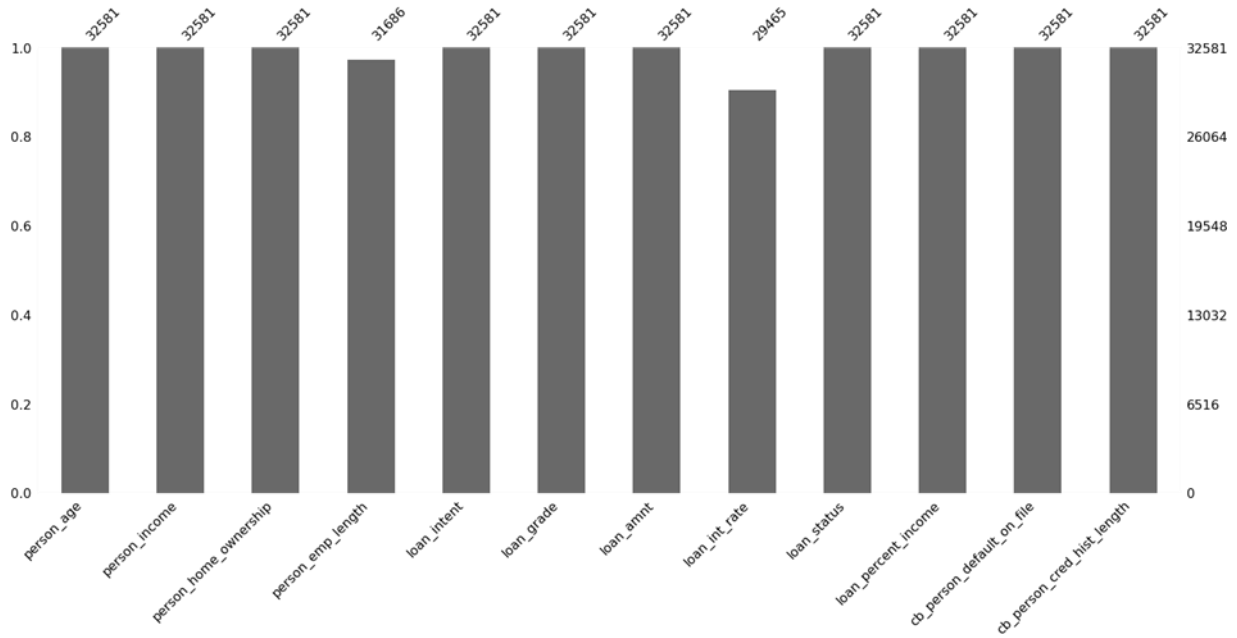
Το Google Collaboratory, εν συντομία Google Collab, αποτελεί μια πλατφόρμα που αναπτύχθηκε από την Google και λειτουργεί δωρεάν στον διαδικτυακό περιηγητή του καθενός προσφέροντας αρκετές ευκολίες στον χρήστη στον προγραμματισμό του με Python [44].

4.3 Προετοιμασία και Καθαρισμός δεδομένων

4.3.1 NA values

Το 1ο βήμα πριν προχωρήσουμε στην δημιουργία μοντέλων μηχανικής μάθησης και σε νευρωνικά δίκτυα είναι να εξερευνήσουμε τα δεδομένα μας, ώστε να βρούμε αν έχουμε NA τιμές, δηλαδή τιμές που είτε λείπουν είτε είναι λάθος (π.χ. το άπειρο), όπως επίσης και να διορθώσουμε τους τύπους των δεδομένων σε περίπτωση που αυτοί είναι λάθος. Γενικά, όταν ξεκινάμε ένα πρότζεκτ ανάλυσης και μοντελοποίησης δεδομένων πολλές φορές ερχόμαστε αντιμέτωποι με ακατάστατα δεδομένα και επομένως πρέπει να τα μετασχηματίσουμε.

Στο παρακάτω γράφημα οπτικοποιούμε τις εκλειπόμενες τιμές, αν υπάρχουν.



Στις στήλες – μεταβλητές “person_emo_length” και “loan_int_rate” έχουν πολλές τιμές που λείπουν, συγκεκριμένα, 895 και 3116 αντίστοιχα. Επειδή οι μεταβλητές αυτές εκφράζουν ποσότητα επιλέγουμε να μην διαγράψουμε τις αντίστοιχες σειρές – παρατηρήσεις του συνόλου δεδομένων, αλλά να αντικαταστήσουμε τις τιμές αυτές με τον αντίστοιχο μέσο όρο της στήλης.

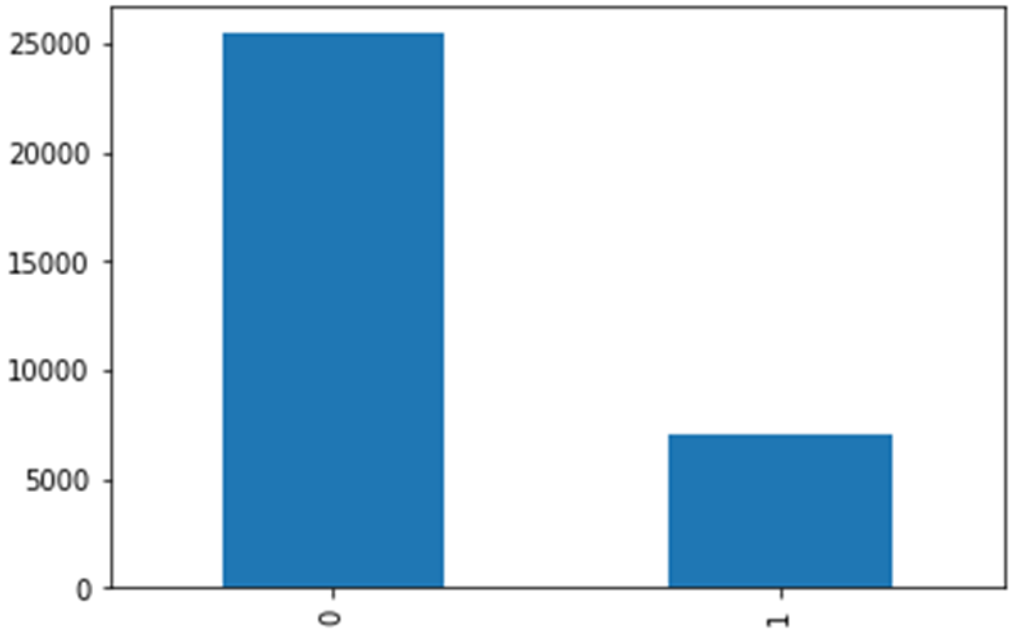
4.3.2 Label encoding

Από την θεωρία γνωρίζουμε πως στα προβλήματα Μηχανικής Μάθησης και Νευρωνικών Δικτύων που πραγματοποιούμε προβλέψεις για μοντέλα ταξινόμησης και παλινδρόμησης, τα δεδομένα που γίνονται δεκτά ως είσοδο στα σύνολα εκπαίδευσης και ελέγχου δεν πρέπει – απαγορεύεται να είναι ποιοτικά, δηλαδή μη – αριθμητικά [45]. Όπως είναι φανερό, όμως, σχεδόν πάντα, έχουμε και κατηγορικά δεδομένα σε ένα πλαίσιο δεδομένων. Έτσι, προκειμένου να μην χαθούν αυτά τα δεδομένα και η πληροφορία που προσφέρουν δουλεύουμε ως εξής. Πραγματοποιούμε κωδικοποίηση «ετικέτας» των κατηγορικών

μεταβλητών με κάποια μεθοδολογία κωδικοποίησης με πιο γνωστές (στην Python) Label Encoder και One hot Encoder [45].

4.3.3 Smote Technique – from Imbalanced Data to Balanced

Το τελευταίο βήμα είναι η μετατροπή των «ακατάστατων» μη – ισορροπημένων δεδομένων (imbalanced data) σε ισορροπημένο (balanced). Αυτό συμβαίνει σε ένα πρόβλημα ταξινόμησης υπάρχει μεγάλη διαφορά στο πλήθος των παρατηρήσεων που ανήκουν σε μια τάξη. Επιπλέον, αυτό προκαλεί μείωση της απόδοσης των μετέπειτα αναπτυσσόμενων μοντέλων μηχανικής μάθησης και τις περισσότερες φορές, ερευνητικά, διορθώνεται [46].

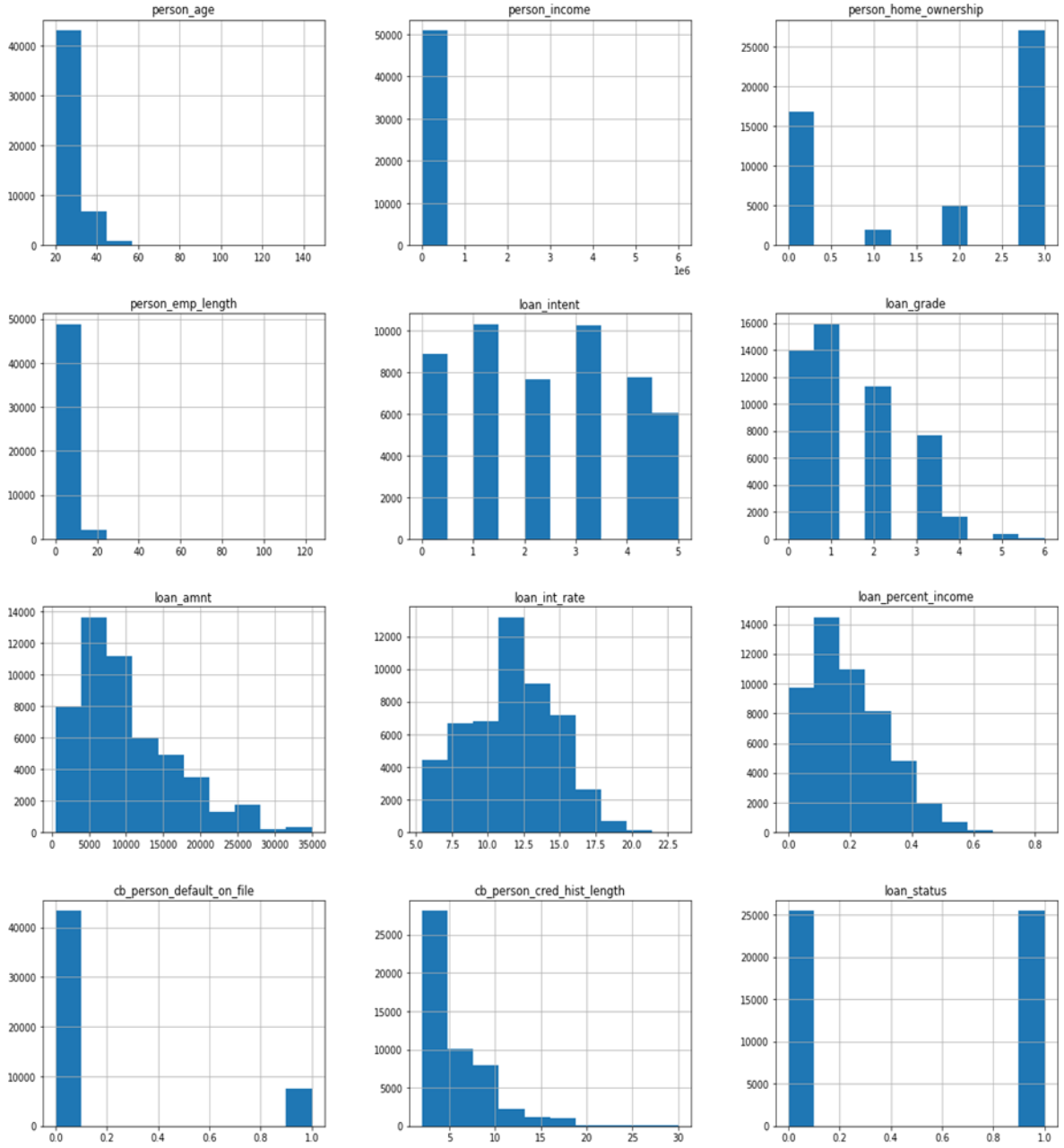


Στα δεδομένα μας παρατηρούμε ότι από τους 32581 δανειολήπτες, οι 25473 ανήκουν στην τάξη 0 της μεταβλητής πρόβλεψης, ενώ μόλις οι 7108 ανήκουν στην τάξη 1. Αυτό σημαίνει πως χρησιμοποιούμε την διαδικασία SMOTE και κάνουμε υπερ-δειγματοληψία με αποτέλεσμα το τελικό σύνολο δεδομένων προς χρήση να αποτελείται από 50946 εγγραφές (και 12 μεταβλητές).

4.3.4 Data Analysis

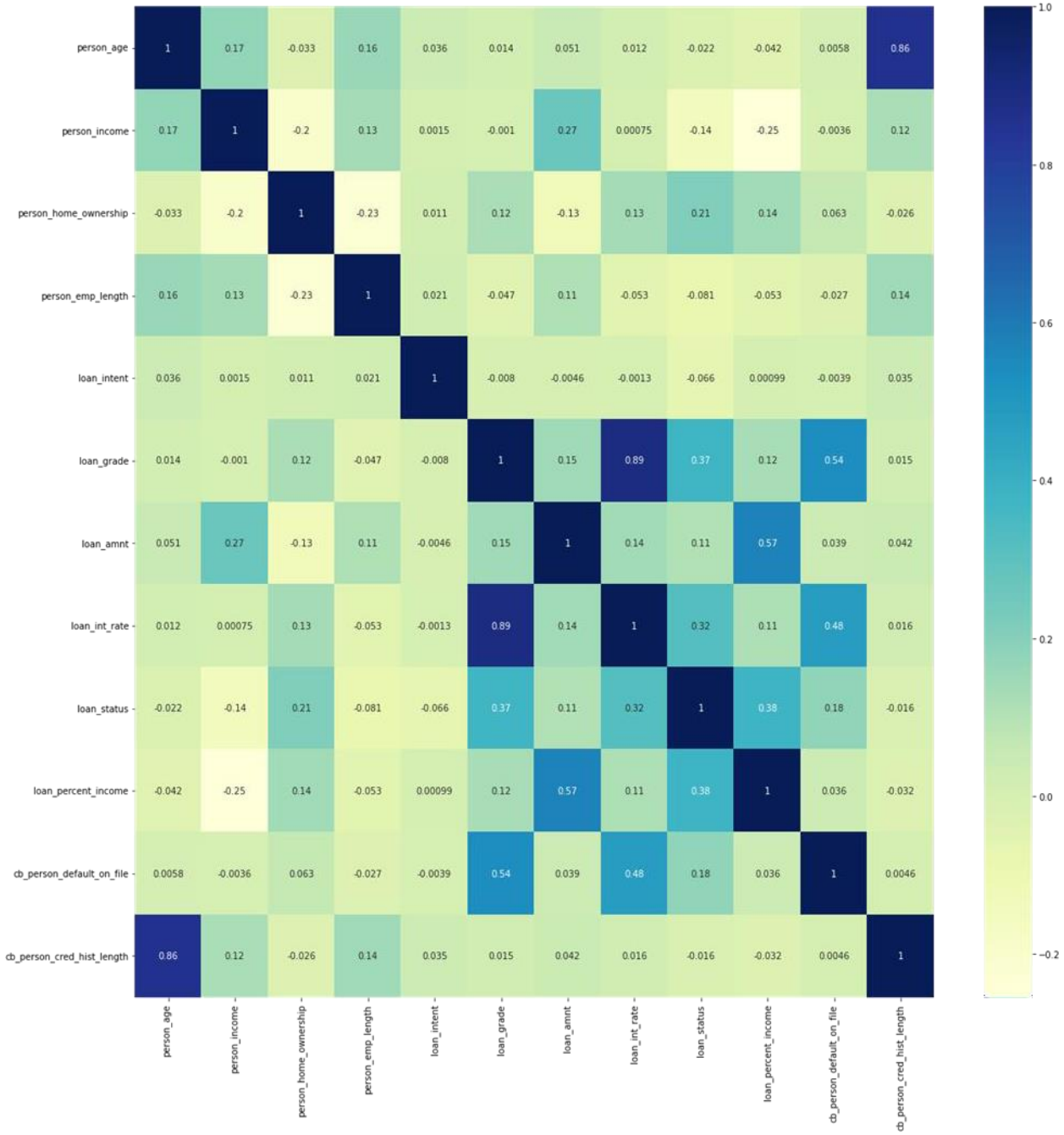
1) Ιστογράμματα

Τα ιστογράμματα αποτελούν ιδιαίτερα σημαντικά στατιστικά γραφήματα, τα οποία χρησιμοποιούνται για την κατανόηση των κατανομών των μεταβλητών ενός πλαισίου δεδομένων, καθώς και για την παρατήρηση και τον υπολογισμό διάφορων άλλων στατιστικών μέτρων (π.χ. εύρος τιμών) [47].



2) Heatmap – Χάρτης πληροφορίας και συσχετίσεων

Ο Heatmap ή πίνακας «θερμότητας» αποτελεί ένα από τα σημαντικότερα γραφήματα στον τομέα της ανάλυσης δεδομένων, καθώς μας παρέχει πληροφορία για το ποιες μεταβλητές σχετίζονται περισσότερο μεταξύ τους και ποιες όχι. Ο συντελεστής στατιστικής συσχέτισης που χρησιμοποιεί αυτός ο πίνακας είναι ο συντελεστής Pearson. Οι μεταβλητές συγκρίνονται και αναλύονται ανά ζεύγη των δύο και εξάγεται η συσχέτιση που είναι ένα αριθμητικό ποσοστό. Όσο πιο κοντά στη μονάδα είναι το ποσοστό τόσο μεγαλύτερη η συσχέτιση και περισσότερη πληροφορία παρέχεται [47].



3) Pairplot – διάγραμμα ζευγών

Το Pairplot αποτελεί ένα εξίσου σημαντικό γράφημα στο οποίο παρουσιάζονται τόσο οι συσχετίσεις μεταξύ δύο διαφορετικών μεταβλητών, όσο και η κατανομή των τιμών τους σε ένα διάγραμμα σημείων. Το αξιοσημείωτο, εδώ, είναι πως υπάρχει και χρωματικός διαχωρισμός των τιμών με βάση την κλάση στην οποία ανήκουν, όσον αφορά την μεταβλητή πρόβλεψης. Τέλος, στην κύρια διαγώνιο παρατηρούμε ιστογράμματα κατανομής των μεταβλητών [47].

5 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Παρακάτω παρουσιάζουμε τα αποτελέσματα των προβλέψεων των μοντέλων που αναπτύχθηκαν στην Python, σύμφωνα με τις μετρικές αξιολόγησης που περιγράψαμε παραπάνω.

5.1 K-Nearest Neighbors

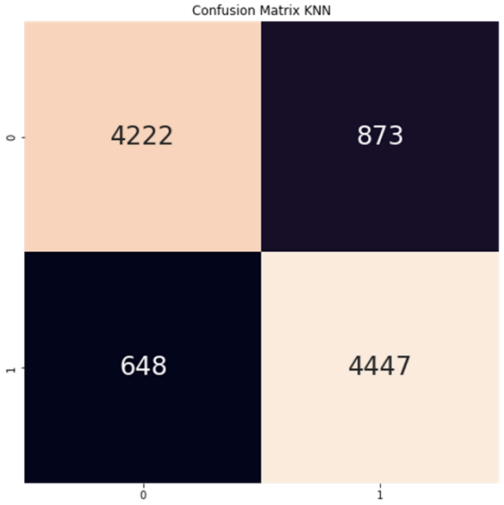
Classification Report:

	Precision	Recall	f1 - score	Support
0	0.87	0.83	0.85	5095
1	0.84	0.87	0.85	5095
accuracy			**0.85**	10190
macro avg	0.85	0.85	0.85	10190
weighted avg	0.85	0.85	0.85	10190

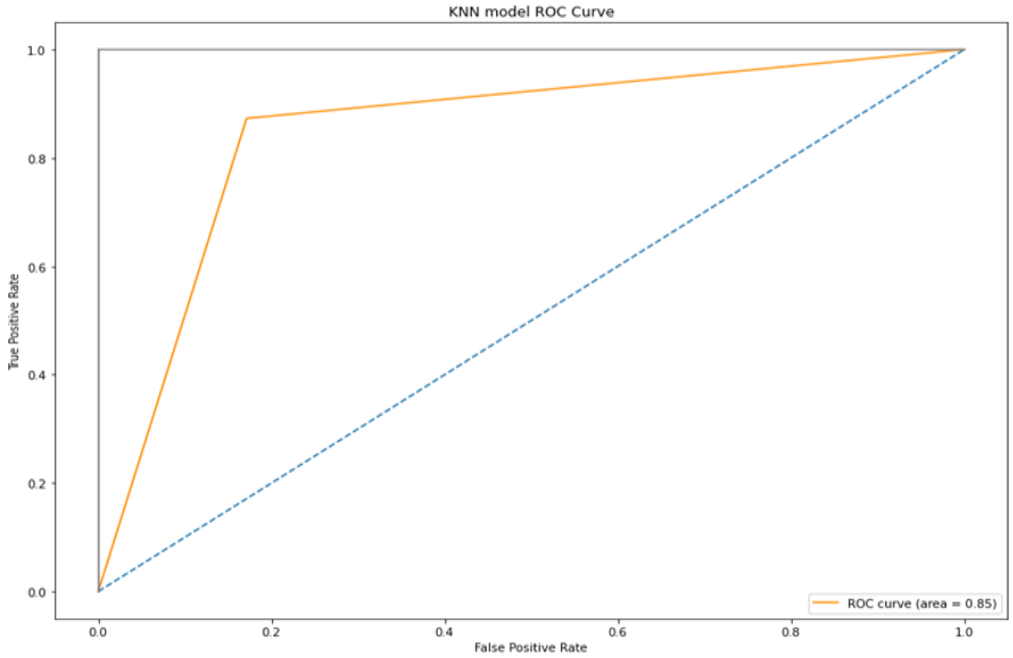
Accuracy score & ROC accuracy score:

Accuracy score	**ROC accuracy score**
85.07 %	85.07 %

Confusion Matrix:



ROC Curve:



5.2 Logistic Regression

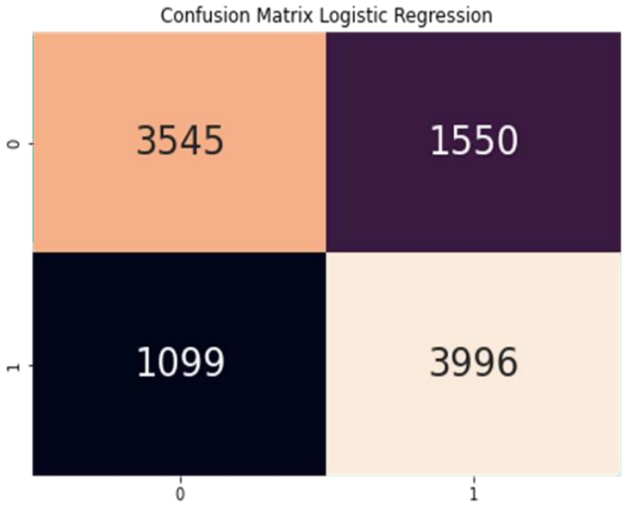
Classification Report:

	Precision	Recall	f1 - score	Support
0	0.76	0.70	0.73	5095
1	0.72	0.78	0.75	5095
accuracy			0.74	10190
macro avg	0.74	0.74	0.74	10190
weighted avg	0.74	0.74	0.74	10190

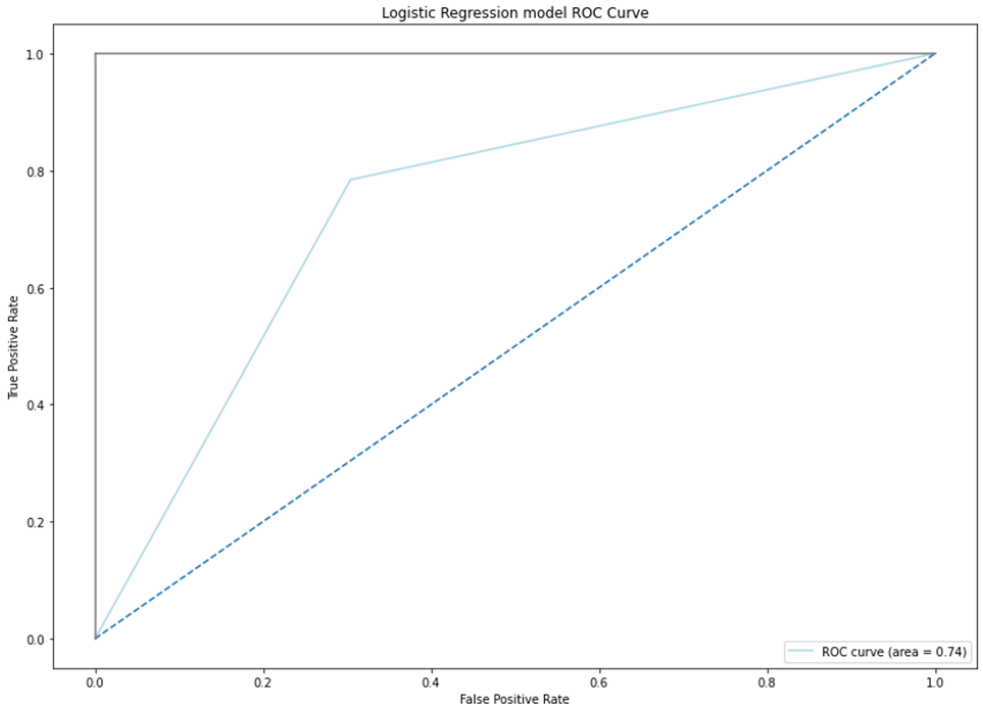
Accuracy score & ROC accuracy score:

Accuracy score	ROC accuracy score
74.00 %	74.00 %

Confusion Matrix:



ROC Curve:



5.3 Decision Tree

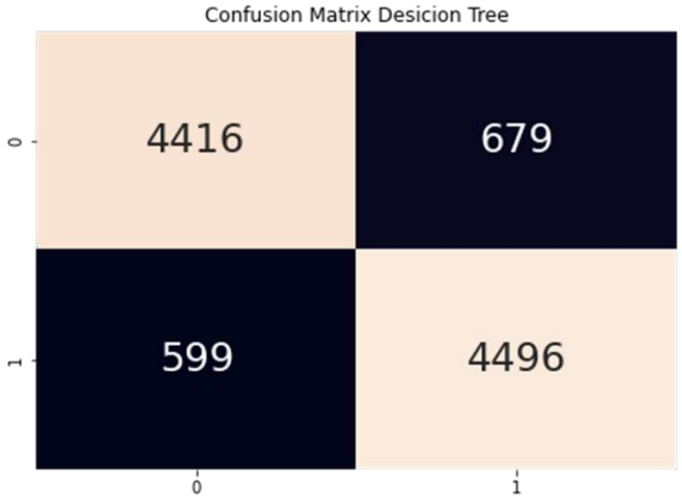
Classification Report:

	Precision	Recall	f1 - score	Support
0	0.88	0.87	0.87	5095
1	0.87	0.88	0.88	5095
accuracy			0.87	10190
macro avg	0.87	0.87	0.87	10190
weighted avg	0.87	0.87	0.87	10190

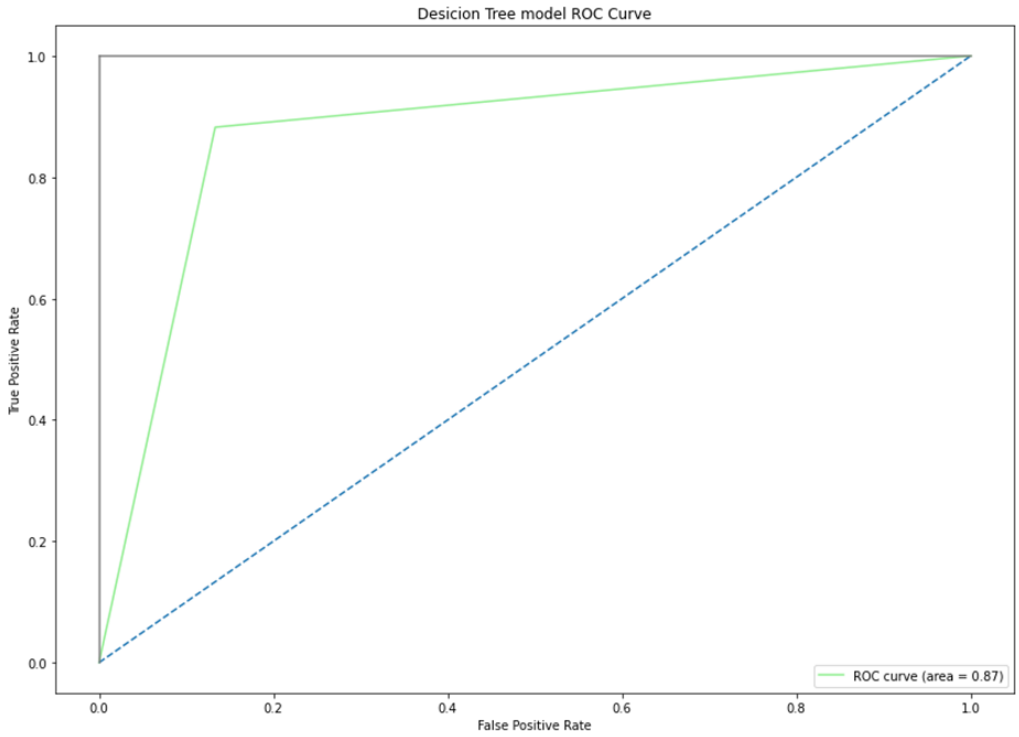
Accuracy score & ROC accuracy score:

Accuracy score	ROC accuracy score
87.46 %	87.49 %

Confusion Matrix:



ROC Curve:



5.4 Random Forest:

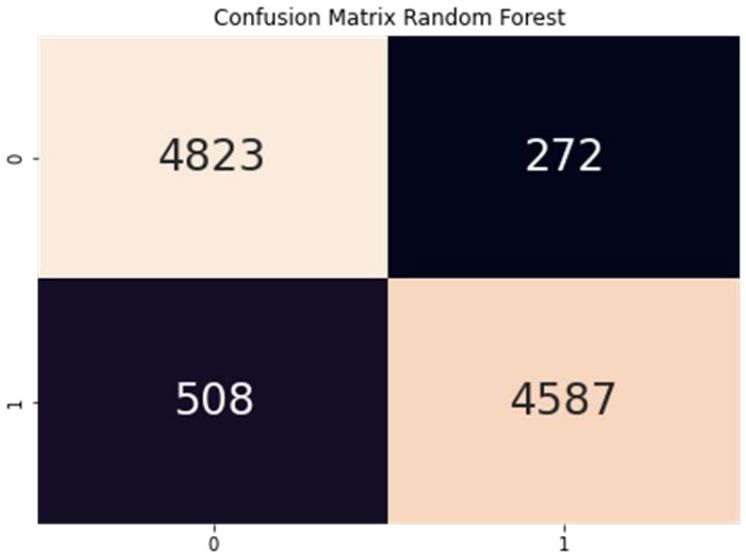
Classification Report:

	Precision	Recall	f1 - score	Support
0	0.90	0.95	0.93	5095
1	0.94	0.90	0.93	5095
accuracy			0.92	10190
macro avg	0.92	0.92	0.92	10190
weighted avg	0.92	0.92	0.92	10190

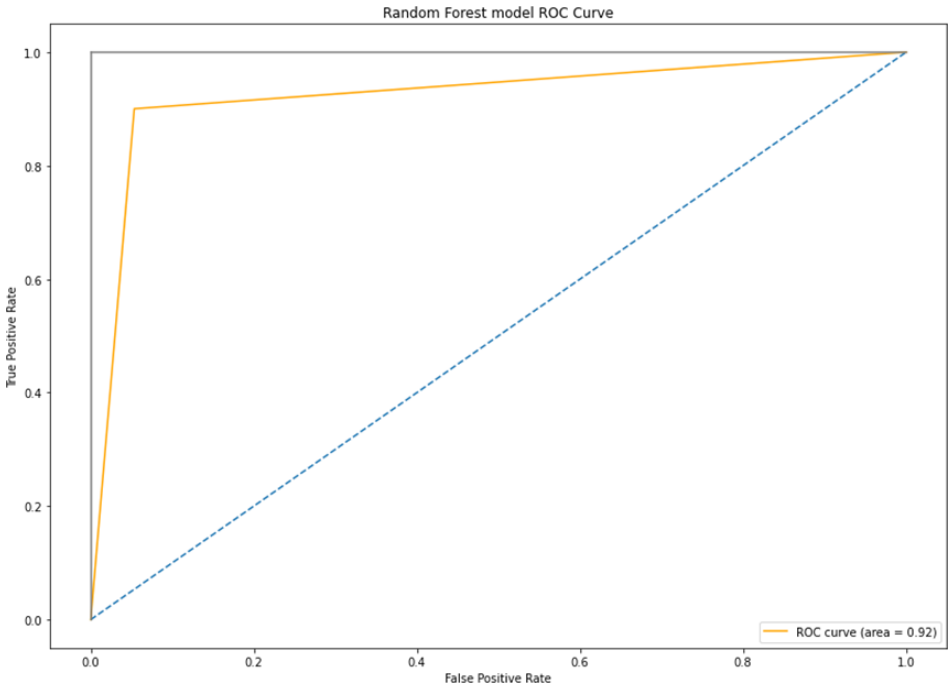
Accuracy score & ROC accuracy score:

Accuracy score	ROC accuracy score
92.35 %	92.35 %

Confusion Matrix:



ROC Curve:



5.5 1st Neural Network:

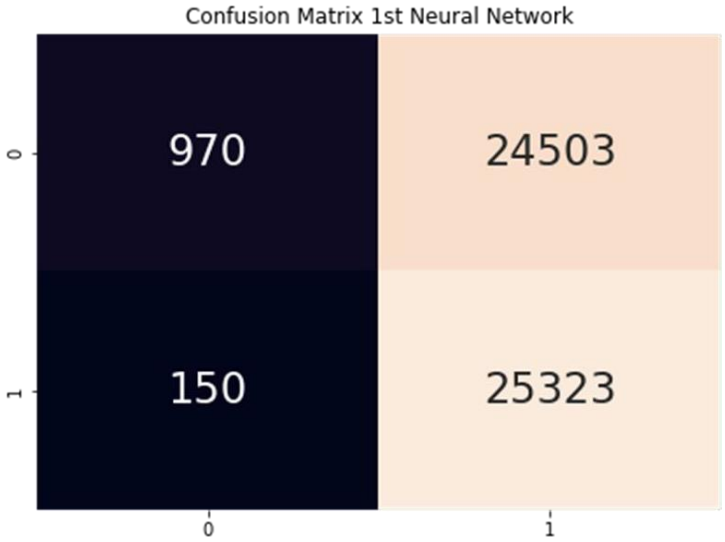
Classification Report:

	Precision	Recall	f1 - score	Support
0	0.87	0.04	0.07	25473
1	0.51	0.99	0.67	25473
accuracy			0.52	50946
macro avg	0.69	0.52	0.37	50946
weighted avg	0.69	0.52	0.37	50946

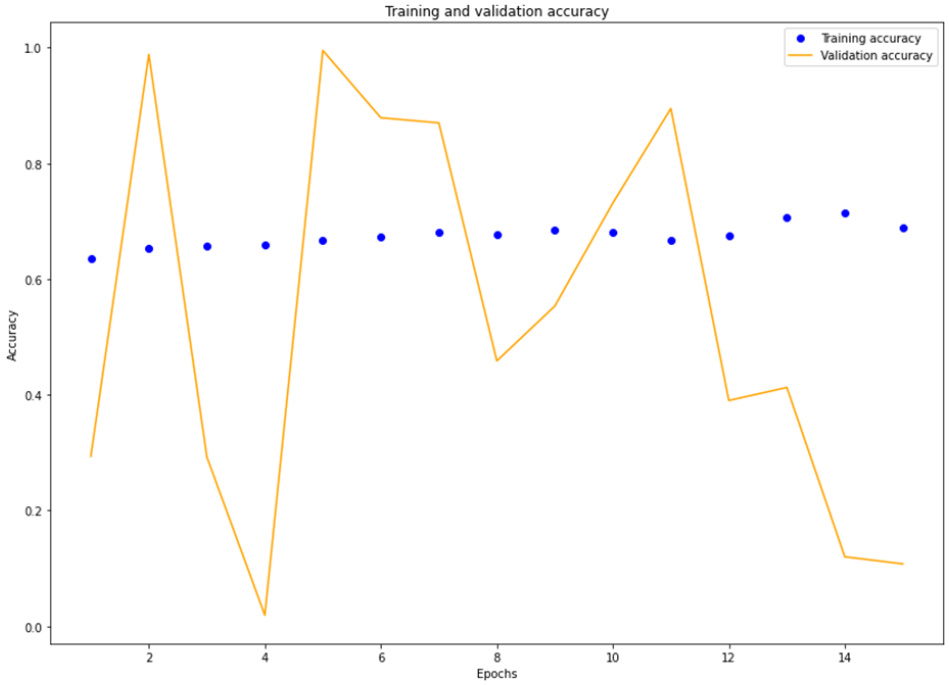
Accuracy score & ROC accuracy score:

Accuracy score	ROC accuracy score
51.60 %	51.60 %

Confusion Matrix:



Accuracy Learning Rate Curve:



5.6 2nd Neural Network:

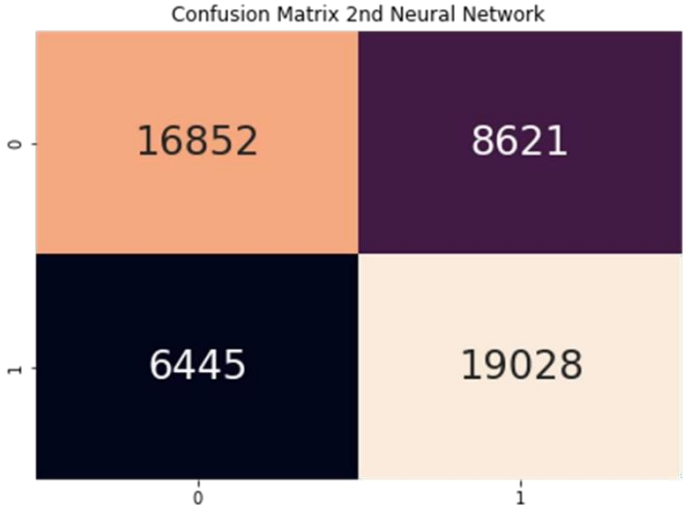
Classification Report:

	Precision	Recall	f1 - score	Support
0	0.72	0.66	0.69	25473
1	0.69	0.75	0.72	25473
accuracy			0.70	50946
macro avg	0.71	0.70	0.70	50946
weighted avg	0.71	0.70	0.70	50946

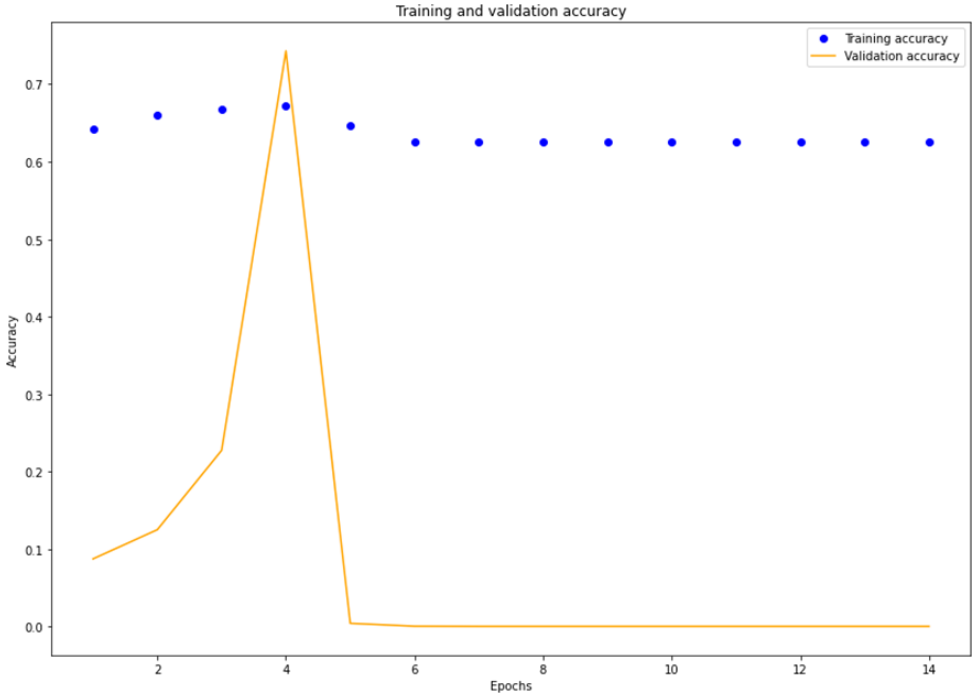
Accuracy score & ROC accuracy score:

Accuracy score	ROC accuracy score
70.43.60 %	70.43 %

Confusion Matrix:



Accuracy Learning Rate Curve:



5.7 Best Model – Καλύτερο μοντέλο

Το καλύτερο μοντέλο από όσα κατασκευάσαμε και παρουσιάσαμε στην παρούσα εργασία διαφαίνεται πως είναι ο αλγόριθμος Random Forest, ο οποίος μάλιστα έχει την ικανότητα να ταξινομεί σωστά, δηλαδή να προβλέπει σωστά, με πιθανότητα 92%. Αυτό σημαίνει πως το μοντέλο έχει την ευχέρεια να προβλέπει από τους 100 δανειολήπτες σωστά τους 92, για το αν θα είναι αξιόπιστοι ή όχι στην αποπληρωμή του δανείου τους. Και έτσι να ελαχιστοποιεί το χρηματοπιστωτικό ρίσκο του φορέα, ο οποίος τον χρησιμοποιεί.

Πως όμως μπορούμε να επιχειρήσουμε να ερμηνεύσουμε την υπεροχή του Random Forest μοντέλου έναντι των άλλων? Ένα σημαντικό πλεονέκτημα του Random Forest αλγορίθμου στο οποίο μπορούμε εν μέρει να αποδώσουμε την υπεροχή του, είναι η ικανότητα του να περιορίζει το overfitting (ύπερ-προσαρμογή) και το bias (μεροληψία) σε δεδομένα μεγάλων διαστάσεων όπως ήταν και το data set που είχαμε, αρετές που φάνηκαν και στην υλοποίηση μας.

Αυτά τα χαρακτηριστικά τα οποία είναι παρόντα και στα Decision Trees αλλά σε όχι τόσο υψηλό βαθμό, ερμηνεύουν και τις επίσης πολύ καλές επιδόσεις του Decision Tree μοντέλου που χρησιμοποιήσαμε, το οποίο είχε και τις δεύτερες καλύτερες επιδόσεις συνολικά και μάλιστα με πολλές ομοιότητες με το Random Forest, ειδικά αν λάβουμε υπόψιν τα Confusion Matrix των δύο όπου με την εξαίρεση των FP (False Positive), έχει παραπλήσιες αλλά και πάλι ξεκάθαρα υποδεέστερες του Random Forest επιδόσεις.

Από την θεωρία γνωρίζουμε για αυτές τις ιδιότητες και πλεονεκτήματα των Random Forests. Όσον αφορά τώρα τα βασικά τους μειονεκτήματα, αυτά είναι η υπολογιστική πολυπλοκότητα (complexity) και ο μεγάλος χρόνος εκπαίδευσης (training period) τα οποία σε αρκετές εφαρμογές της πραγματικής ζωής που απαιτούν ταχεία λήψη απόφασης δεν επιτρέπουν την εφαρμογή τους. Και φυσικά ας μην ξεχνάμε ότι το Random Forest είναι ένα προβλεπτικό (predictive) και όχι περιγραφικό (descriptive) μοντέλο, άρα εκ των πραγμάτων αποκλείεται η χρήση του σε μια ευρεία γκάμα εφαρμογών.

Όμως στην εφαρμογή που υλοποιήσαμε, η οποία φυσικά είναι της προβλεπτικής κατηγορίας, δεν είχαμε περιορισμό στον χρόνο εκπαίδευσης ενώ και ως προς την υπολογιστική πολυπλοκότητα απολαμβάναμε πλεονεκτήματα έναντι άλλων εφαρμογών. Έτσι μετριάσαμε τις όποιες δυσκολίες και μεγιστοποιήσαμε τα θετικά της χρήσης αλγορίθμων Random Forest.

Η αστοχία του Νευρωνικού Δικτύου στην πρώτη του υλοποίηση θα πρέπει να αποδοθεί στην ελλιπή του εκπαίδευση ενώ και τα τελικά αποτελέσματα της δεύτερης εκδοχής του δεν είναι όσο καλά είναι αυτά των καλύτερων αλγορίθμων μας. Σε ένα σημαντικά μεγαλύτερο δείγμα όμως πιστεύουμε ότι αυτή η εικόνα θα άλλαζε με το Νευρωνικό Δίκτυο να επιτυγχάνει σημαντικά πιο αξιόπιστες επιδόσεις.

6 ΔΙΑΔΙΚΑΣΙΑ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Ένας άλλος σημαντικός τομέας της Ανάλυσης Δεδομένων και της Τεχνητής Νοημοσύνης, που έχει αναδειχθεί τα τελευταία χρόνια είναι αυτός της επιλογής των σημαντικότερων μεταβλητών και ονομάζεται Feature Engineering (Μηχανική των Μεταβλητών). Στόχος του είναι η μελέτη και η χρήση υπολογιστικών αλγορίθμων σε ένα σύνολο δεδομένων, βάσει κάποιου τεχνικού ή μαθηματικού κριτηρίου, ώστε να μπορεί κάποιος να συμπεράνει ποιες μεταβλητές από το αρχικό σύνολο προσδίδουν την περισσότερη πληροφορία για το dataset (πλαίσιο δεδομένων) και είναι οι πιο σημαντικές (features importance) για την βελτιστοποίηση της απόδοσης των αναπτυσσόμενων μοντέλων (μηχανικής μάθησης) [48]. Για τον λόγο αυτό αναλύσαμε την σημαντικότητα της εκάστοτε μεταβλητής με τέσσερις διαφορετικές τεχνικές:

- Correlation Importance: βασίζεται στα ποσοστά συσχέτισης μεταξύ των μεταβλητών
- Permutation Importance: βασίζεται στον αλγόριθμο KNN
- Decision Tree Classification Importance: βασίζεται στον αλγόριθμο Decision Tree
- Random Forest Classification Importance: βασίζεται στον αλγόριθμο Random Forest

Και πήραμε τα κάτωθι αποτελέσματα:

Όνομα μεταβλητής	**Correlation Importance**	**Permutation Importance with KNN**	**Decision Tree Classification Importance**	**Random Forest Importance**
person_age				
person_income	X	X	X	X
person_home_ownership	X			
person_emp_length			X	
loan_intent	X			
loan_grade	X			
loan_amnt		X		
loan_int_rate	X		X	X
loan_status				
loan_percent_income	X		X	X
cb_person_default_on_file				

Συμπεραίνουμε, λοιπόν, πως η σημαντικότερη μεταβλητή αυτής της έρευνας και η οποία παίζει τον καθοριστικότερο ρόλο στον αν ένας δανειολήπτης είναι αξιόπιστος ή όχι αποτελεί

το ετήσιο εισόδημα του. Άλλες σημαντικές μεταβλητές που παίζουν ξεχωριστό ρόλο είναι οι τόκοι του κατεχόμενου δανείου και το ποσοστιαίο εισόδημα.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Το μέλλον της μηχανικής μάθησης στον τραπεζικό και χρηματοπιστωτικό τομέα αναμένεται να είναι μεγάλο, ιδίως στον τομέα της διαχείρισης πιστωτικών κινδύνων. Οι τεχνικές μηχανικής μάθησης και νευρωνικών δικτύων αναπτύσσονται και θα συνεχίσουν να εξελίσσονται ραγδαία για εφαρμογή πάνω σε τραπεζικά δεδομένα σε μια προσπάθεια βελτίωσης των λειτουργιών τους. Η ικανότητα των μοντέλων μηχανικής και βαθιάς μάθησης να αναλύουν μεγάλο όγκο δεδομένων με απλό και προγραμματιστικό τρόπο και με μεγαλύτερη αξιοπιστία είναι πολύ σημαντική. Η μηχανική μάθηση, έχοντας σημαντικές εφαρμογές στη διαχείριση κινδύνων, μπορεί να επιτρέψει τη δημιουργία ακριβέστερων μοντέλων κινδύνου με τον εντοπισμό πολύπλοκων, μη γραμμικών προτύπων σε μεγάλα σύνολα δεδομένων.

Ο πιστωτικός κίνδυνος θεωρείται ένας από τους σημαντικότερους κίνδυνους για έναν χρηματοπιστωτικό οργανισμό. Πιο συγκεκριμένα, τα προβλήματα διαχείρισης του πιστωτικού κινδύνου που έχουν διερευνηθεί αφορούσαν την πιστοληπτική ικανότητα. Στην παρούσα διπλωματική διατριβή μελετήθηκαν εκτενώς, τα πλεονεκτήματα και τα μειονεκτήματα των διαφόρων τεχνικών μηχανικής μάθησης στην επίλυση συγκεκριμένων προβλημάτων διαχείρισης κινδύνου, τα οποία μπορούν να μελετηθούν και να αξιολογηθούν περαιτέρω.

Τέλος, αξίζει να σημειωθεί πως σε μικρότερο όγκο δεδομένων, λόγου χάρη μέχρι 100000 παρατηρήσεις, οι παραδοσιακές τεχνικές μηχανικής μάθησης υπερτερούν με ξεχωριστά αποτελέσματα, με τρανταχτά παραδείγματα τους αλγορίθμους Random Forest και Support Vector Machines. Ωστόσο, τα νευρωνικά δίκτυα προτιμώνται, πλέον, όταν το πρόβλημα εμπεριέχει πολύ μεγαλύτερο όγκο δεδομένων (συνήθως δεκάδες εκατομμύρια), καθώς η υπολογιστική τους ισχύ είναι τεράστια, λειτουργούν επαναληπτικά - απεριόριστα για όσο ορίσει ο χρήστης και βρίσκονται σε μια κατάσταση διαρκούς μάθησης έως ότου προσφέρουν το επιθυμητό αποτέλεσμα.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. Management science, 49(3), 312-329.

[2] Bărbuță-Mișu, N., & Madaleno, M. (2020). Assessment of bankruptcy risk of large companies: European countries evolution analysis. Journal of Risk and Financial Management, 13(3), 58.

[3] Camska, D., & Klecka, J. (2020). Comparison of prediction models applied in economic recession and expansion. Journal of Risk and Financial Management, 13(3), 52.

[4] Ogachi, D., Ndege, R., Gaturu, P., & Zoltan, Z. (2020). Corporate bankruptcy prediction model, a special focus on listed companies in Kenya. Journal of Risk and Financial Management, 13(3), 47.

[5] Wiggins, R., Piontek, T., & Metrick, A. (2014). The Lehman brothers bankruptcy a: overview. Yale program on financial stability case study.

[6] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

[7] Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.

[8] Provenzano, A. R., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., ... & Nordio, C. (2020). Machine learning approach for credit scoring. arXiv preprint arXiv:2008.01687.

[9] Kubat, M. (2017). A Simple Machine-Learning Task. An Introduction to Machine Learning. Chapter 1, pp: 1-18. Springer.

[10] Kristóf, T., & Virág, M. (2020). A comprehensive review of corporate bankruptcy prediction in Hungary. Journal of Risk and Financial Management, 13(2), 35.

[11] Charalambakis, E., Dendramis, Y., & Tzavalis, E. (2017). On the determinants of NPLs: lessons from Greece. In Political Economy Perspectives on the Greek Crisis (pp. 289-309). Palgrave Macmillan, Cham.

[12] Pisula, T. (2020). An ensemble classifier-based scoring model for predicting bankruptcy of polish companies in the Podkarpackie Voivodeship. Journal of Risk and Financial Management, 13(2), 37.

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

[14] Gyamfi, N. K., & Abdulai, J. D. (2018, November). Bank fraud detection using support vector machine. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 37-41). IEEE.

[15] Mariani, M. C., Tweneboah, O. K., & Bhuiyan, M. A. M. (2019). Supervised machine learning models applied to disease diagnosis and prognosis. AIMS Public Health, 6(4), 405.

[16] Kleinbaum, D. G., & Klein, M. (2010). Introduction to logistic regression. In Logistic regression (pp. 1-39). Springer, New York, NY.

[17] Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2019). Logistic Regression. Data mining for business analytics: concepts, techniques and applications in Python. Chapter 10. John Wiley & Sons.

[18] Nikhil, A. (2020). A Predictive Analytic study on stock market trend by supervised machine learning algorithms. Alochana Chakra Journal. Volume IX, Issue V, May/2020, pp: 4911-4916.

[19] Kubat, M. (2017). Similarities: Nearest-Neighbor Classifiers. An Introduction to Machine Learning. Chapter 3, pp: 43-64. Springer.

[20] Chen, G. H., & Shah, D. (2018). Explaining the success of nearest neighbor methods in prediction. Boston, MA, USA: Now Publishers.

[21] Surya, V. B., Haneen, P., Ahmad, A. A., Omar, B. A., & Ahmad, L. (2019). Effects of Distance Measure Choice on KNN Classifier Performance-A Review. Mary Ann Liebert.

[22] Ye, R., Le, Z., & Suganthan, P. N. (2013, April). K-nearest neighbor based bagging SVM pruning. In 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL) (pp. 25-30). IEEE.

[23] Kubat, M. (2017). Decision Trees. An Introduction to Machine Learning. Chapter 6, pp: 113-136. Springer.

[24] Niculaescu, O. (2018). Classifying data with decision trees. XRDS: Crossroads, The ACM Magazine for Students, 24(4), 55-57.

[25] Misra, S., & Wu, Y. (2019). Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. Machine Learning for Subsurface Characterization. Chapter 10, pp: 289-314. Elsevier Inc.

[26] Mohajon, J. (2020). Confusion matrix for your multi-class machine learning model. [Διαδίκτυο] Towards Data Science. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826 (Πρόσβαση την 17 ΦΕΒ 2022).

[27] Chang, H., Zhang, H., Qin, Q. M., Zhang, T., Zhang, T., Liu, Y., ... & Shen, B. (2020). Identification of novel Phytophthora infestans small RNAs involved in potato late blight reveals potential cross-kingdom regulation to facilitate oomycete infection. International Journal of Data Mining and Bioinformatics, 23(2), 119-141.

[28] Kubat, M. (2017). Unsupervised Learning. An Introduction to Machine Learning. Chapter 14, pp: 273-296. Springer.

[29] Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. Journal of Internet Services and Applications, 9(1), 1-99.

[30] Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11), e00938.

[31] Berner, J., Grohs, P., Kutyniok, G., & Petersen, P. (2021). The modern mathematics of deep learning. arXiv preprint arXiv:2105.04026.

[32] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An introductory review of deep learning for prediction models with big data. Frontiers in Artificial Intelligence, 3, 4.

[33] Antar, M. A., Elassiouti, I., & Allam, M. N. (2006). Rainfall-runoff modelling using artificial neural networks technique: a Blue Nile catchment case study. Hydrological Processes: An International Journal, 20(5), 1201-1216.

[34] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), 1-74.

[35] Chang, J. Z. (2018). Training Neural Networks to Pilot Autonomous Vehicles: Scaled Self-Driving Car. Senior Projects Spring 2018. 402.

[36] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[37] Hartnett, K. (2019). Foundations built for a general theory of neural networks. [Διαδίκτυο] Quanta magazine. https://www.quantamagazine.org/foundations-built-for-a-general-theory-of-neural-networks-20190131/ (Πρόσβαση την 20 ΦΕΒ 2022).

[38] Kubat, M. (2017). Performance Evaluation. An Introduction to Machine Learning. Chapter 11, pp: 211-230. Springer.

[39] Feki, A., Ishak, A. B., & Feki, S. (2012). Feature selection using Bayesian and multiclass Support Vector Machines approaches: Application to bank risk prediction. Expert Systems with Applications, 39(3), 3087-3099.

[40] Viering, T., & Loog, M. (2021). The shape of learning curves: a review. arXiv preprint arXiv:2103.10948.

[41] Khan, R. A., Crenn, A., Meyer, A., & Bouakaz, S. (2019). A novel database of children's spontaneous facial expressions (LIRIS-CSE). Image and Vision Computing, 83, 61-69.

[42] Singh, H. (2019). Image Processing Using Machine Learning. Practical Machine Learning and Image Processing - For Facial Recognition, Object Detection, and Pattern Recognition Using Python. Chapter 5, pp: 89-132. New York, NY, USA: Apress.

[43] Rodriguez, L., Pardo, C., Cepeda, J., Gomez, J., & Rivera, S. (2021). Python notebook usefullness, case study: Optimization techniques course. Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, 37(4).

[44] Yalçın, O. G. (2019). 4 Reasons Why You Should Use Google Colab for Your Next Pro-ject. [Διαδίκτυο] Towards Data Science, Medium. https://towardsdatascience.com/4-reasons-why-you-should-use-google-colab-for-your-next-project-b0c4aaad39ed (Πρόσβαση την 12 ΔΕΚ 2021).

[45] Yadav, D. (2019). Categorical encoding using Label-Encoding and One-Hot-Encoder. [Διαδίκτυο] Towards Data Science, Medium. https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd (Πρόσβαση την 16 ΔΕΚ 2021).

[46] Aguilar F. (2019). SMOTE-NC in ML Categorization Models for Imbalanced Datasets. [Διαδίκτυο] Medium. https://medium.com/analytics-vidhya/smote-nc-in-ml-categorization-models-fo-imbalanced-datasets-8adbdcf08c25 (Πρόσβαση την 20 ΔΕΚ 2021).

[47] Navlani, A., Fandango, A., & Idris, I. (2021). Data Visualization. Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python. Chapter 5, pp: 135-189. Packt Publishing Ltd.

[48] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1-21.