ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«ΠΡΟΗΓΜΕΝΑ ΣΥΣΤΗΜΑΤΑ ΠΛΗΡΟΦΟΡΙΚΗΣ – ΑΝΑΠΤΥΞΗ ΛΟΓΙΣΜΙΚΟΥ ΚΑΙ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Συγκριτική μελέτη μεθόδων Μηχανικής Μάθησης για πρόβλεψη πιστωτικού κινδύνου** **A comparative study of Machine Learning approaches for credit risk prediction**
Ονοματεπώνυμο Φοιτητή	**Αριστείδης Βασιλάκης**
Πατρώνυμο	**Γεώργιος**
Αριθμός Μητρώου	**ΜΠΣΠ18005**
Επιβλέπων	**Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής**

Ιούνιος 2023

Τριμελής Εξεταστική Επιτροπή

Διονύσιος Σωτηρόπουλος	Ευθύμιος Αλέπης	Γεώργιος Τσιχριντζής
Επίκουρος Καθηγητής	Αναπληρωτής Καθηγητής	Καθηγητής

Περίληψη

Η πρόβλεψη πιστωτικού κινδύνου είναι ένας σημαντικός τομέας στη χρηματοοικονομική, καθώς επιτρέπει στις χρηματοπιστωτικές εταιρείες να αξιολογούν τον κίνδυνο που συνδέεται με τη χορήγηση πιστώσεων σε πελάτες. Η πιστωτική ανάλυση περιλαμβάνει το μέτρο για τη διερεύνηση της πιθανότητας του αιτούντος να αποπληρώσει έγκαιρα το δάνειο και να προβλέψει την αθέτηση ή αδυναμία αποπληρωμής του. Υπάρχουν δύο βασικοί κίνδυνοι, η απώλεια πιθανών εσόδων που προκύπτει από την μη έγκριση ενός καλού υποψηφίου ή από την απόρριψη πολλών καθώς και η οικονομική ζημία που προκύπτει από την έγκριση ενός υποψηφίου που καταλήγει να μην αποπληρώσει το δάνειο.

Οι μέθοδοι μηχανικής μάθησης έχουν αναπτυχθεί και εφαρμοστεί με επιτυχία σε αυτόν τον τομέα για να προβλέψουν την πιθανότητα πρόκλησης πιστωτικού προβλήματος από τους δανειολήπτες. Στην εργασία μας θα χρησιμοποιήσουμε τρεις μεθόδους, τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Networks), τα Support Vector Machines (SVM) και τα Random Forests. Με τη βοήθεια αυτών των αλγορίθμων θα εκπαιδεύσουμε ένα σύνολο δεδομένων (dataset) το οποίο περιλαμβάνει πληροφορίες για τους δανειολήπτες και το αν έχουν ανταποκριθεί στις πληρωμές των πιστώσεών τους. Αυτός ο τρόπος πρόβλεψης πιστωτικού κινδύνου μπορεί να παρέχει σημαντική πληροφορία για την απόδοση των δανείων και τη δυνατότητα πληρωμής των δανειοληπτών.

Abstract

Credit risk forecasting is an important area in finance as it allows financial companies to assess the risk associated with extending credit to customers. Credit risk analysis includes the measure to investigate the applicant's likelihood of repaying the loan on time and predicting his inability to repay. There are two main risks, the loss of potential revenue resulting from not approving one good candidate or rejecting many, and the financial loss resulting from approving a candidate who ends up defaulting on the loan.

Machine learning methods have been developed and successfully applied in this field to predict the likelihood of borrowers becoming credit distressed. In our work, we will use three methods, Artificial Neural Networks, Support Vector Machines (SVM) and Random Forests. With the help of these algorithms we will train a dataset that includes information about borrowers and whether they have responded to their loan payments. This way of predicting credit risk can provide important information on loan performance and borrowers' ability to pay.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΠΡΟΓΝΩΣΗΣ ΤΟΥ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ

1.1 Η έννοια του Πιστωτικού Κινδύνου

Ο πιστωτικός κίνδυνος αναφέρεται στην πιθανή οικονομική ζημία που μπορεί να αντιμετωπίσει ένας δανειστής ή επενδυτής εάν ένας δανειολήπτης δεν αποπληρώσει το χρέος του ή δεν εκπληρώσει τις οικονομικές του υποχρεώσεις. Είναι ο κίνδυνος που προκύπτει από την αβεβαιότητα της ικανότητας των δανειοληπτών να εκπληρώσουν έγκαιρα τις υποχρεώσεις αποπληρωμής τους *[1].* Όταν ένας δανειολήπτης αθετήσει ή αντιμετωπίζει οικονομική δυσπραγία, ο δανειστής μπορεί να υποστεί ζημίες όσον αφορά το αρχικό ποσό, τις πληρωμές τόκων ή οποιεσδήποτε άλλες οικονομικές δεσμεύσεις που σχετίζονται με το δάνειο. Με άλλα λόγια, ο πιστωτικός κίνδυνος είναι ο κίνδυνος που σχετίζεται με τον δανεισμό χρημάτων ή την παροχή πίστωσης σε ιδιώτες, εταιρείες ή οντότητες. Περιλαμβάνει την πιθανότητα οι δανειολήπτες να αθετήσουν τις πληρωμές των δανείων τους ή να μην εκπληρώσουν τις οικονομικές τους υποχρεώσεις όπως έχουν συμφωνηθεί.

1.2 Τύποι Πιστωτικού Κινδύνου

Παρακάτω παρουσιάζονται οι διάφοροι τύποι που αφορούν τον Πιστωτικό Κίνδυνο. Η κατανόηση και η διαχείριση αυτών των τύπων πιστωτικού κινδύνου είναι ουσιαστικής σημασίας για τους δανειστές και τους επενδυτές για να έχουν τη δυνατότητα να λαμβάνουν τεκμηριωμένες αποφάσεις, να ορίζουν κατάλληλες στρατηγικές διαχείρισης κινδύνου και να μετριάζουν πιθανές απώλειες *[2]* .

1. Κίνδυνος αθέτησης (Default Risk)

Ο κίνδυνος αθέτησης είναι ο κίνδυνος ο δανειολήπτης να αποτύχει να πραγματοποιήσει έγκαιρες πληρωμές ή να αθετήσει τις χρεωστικές του υποχρεώσεις. Είναι ο πιο σημαντικός και ευρέως αναγνωρισμένος τύπος πιστωτικού κινδύνου. Όταν ένας δανειολήπτης αθετήσει, ο δανειστής μπορεί να υποστεί οικονομικές απώλειες όσον αφορά το αρχικό ποσό, τις πληρωμές τόκων ή οποιεσδήποτε άλλες οικονομικές δεσμεύσεις που σχετίζονται με το δάνειο.

2. Κίνδυνος φερεγγυότητας (Creditworthiness Risk)

Ο κίνδυνος πιστοληπτικής ικανότητας αναφέρεται στον κίνδυνο που σχετίζεται με την αξιολόγηση της πιστοληπτικής ικανότητας ενός δανειολήπτη πριν από τη χορήγηση της πίστωσης. Περιλαμβάνει την αξιολόγηση της οικονομικής κατάστασης του δανειολήπτη, του πιστωτικού ιστορικού, της σταθερότητας του εισοδήματος και άλλων σχετικών παραγόντων για τον προσδιορισμό της ικανότητάς του να αποπληρώσει. Αυτός ο τύπος πιστωτικού κινδύνου βοηθά τους δανειστές να

καθορίσουν εάν θα εγκρίνουν ένα δάνειο, να ορίσουν κατάλληλα επιτόκια και να καθορίσουν πιστωτικά όρια.

3. Κίνδυνος συγκέντρωσης (Concentration Risk)

Ο κίνδυνος συγκέντρωσης προκύπτει από την ύπαρξη μεγάλης έκθεσης (exposure) σε έναν μεμονωμένο δανειολήπτη, κλάδο ή γεωγραφική περιοχή. Εάν ένας δανειστής έχει ένα σημαντικό μέρος του χαρτοφυλακίου δανείων του συγκεντρωμένο σε έναν δανειολήπτη ή έναν τομέα, τυχόν δυσμενείς εξελίξεις που επηρεάζουν αυτόν τον δανειολήπτη ή τον τομέα μπορεί να έχουν ουσιαστικό αντίκτυπο στη συνολική έκθεση του δανειστή σε πιστωτικό κίνδυνο. Για τον μετριασμό του κινδύνου συγκέντρωσης χρησιμοποιούνται τεχνικές διαφοροποίησης και διαχείρισης χαρτοφυλακίου.

4. Κυρίαρχος Κίνδυνος (Sovereign Risk)

Ο κυρίαρχος κίνδυνος είναι ο κίνδυνος που σχετίζεται με τον δανεισμό σε κυβερνήσεις ή κυβερνητικές οντότητες. Περιλαμβάνει τη δυνατότητα αθέτησης υποχρεώσεων ή καθυστερήσεων πληρωμών από την κυβέρνηση. Παράγοντες που επηρεάζουν τον κυρίαρχο κίνδυνο περιλαμβάνουν την πολιτική σταθερότητα της χώρας, την οικονομική της δύναμη, τις δημοσιονομικές πολιτικές και την αστάθεια του συναλλάγματος. Οι δανειστές αξιολογούν τον κρατικό κίνδυνο όταν χορηγούν πιστώσεις σε ξένες κυβερνήσεις ή επενδύουν σε κρατικά ομόλογα.

5. Κίνδυνος αντισυμβαλλομένου (Counterparty Risk)

Ο κίνδυνος αντισυμβαλλομένου αναφέρεται στον κίνδυνο ο αντισυμβαλλόμενος σε μια χρηματοοικονομική συναλλαγή να μην εκπληρώσει τις υποχρεώσεις του. Είναι σύνηθες σε διάφορες χρηματοοικονομικές συναλλαγές, όπως συμβόλαια παραγώγων (derivatives contracts), διαπραγμάτευση τίτλων (securities trading) και διατραπεζικός δανεισμός (interbank lending). Ο κίνδυνος αντισυμβαλλομένου προκύπτει λόγω παραγόντων όπως η χρεοκοπία, η πτώχευση ή οι λειτουργικές αποτυχίες του αντισυμβαλλομένου. Τα πιστωτικά παράγωγα (credit derivatives) και οι εξασφαλίσεις (collateralization) χρησιμοποιούνται για τον μετριασμό του κινδύνου αντισυμβαλλομένου.

6. Κίνδυνος χώρας (Country Risk)

Ο κίνδυνος μιας χώρας είναι ο κίνδυνος που σχετίζεται με τον δανεισμό ή την επένδυση σε μια ξένη χώρα. Περιλαμβάνει παράγοντες όπως η πολιτική σταθερότητα, το νομικό και ρυθμιστικό περιβάλλον, οι οικονομικές συνθήκες και η αστάθεια των συναλλαγματικών ισοτιμιών. Η αξιολόγηση κινδύνου μιας χώρας βοηθά στην αξιολόγηση του πιθανού πιστωτικού κινδύνου που σχετίζεται με διασυνοριακές συναλλαγές και επενδύσεις.

7. Κίνδυνος κλάδου (Sector Risk)

Ο κίνδυνος κλάδου αναφέρεται στον πιστωτικό κίνδυνο που σχετίζεται με συγκεκριμένους κλάδους ή τομείς. Ορισμένοι κλάδοι ενδέχεται να είναι πιο επιρρεπείς σε χρηματοπιστωτική αστάθεια ή οικονομική ύφεση, αυξάνοντας τον πιστωτικό κίνδυνο των δανειοληπτών σε αυτούς τους τομείς. Οι δανειστές αξιολογούν ειδικούς παράγοντες του κλάδου, όπως οι συνθήκες της αγοράς, ο ανταγωνισμός, οι τεχνολογικές εξελίξεις, οι κανονιστικές αλλαγές και η δυναμική ζήτησης-προσφοράς κατά την αξιολόγηση του πιστωτικού κινδύνου.

1.3 Παράγοντες που επηρεάζουν τον πιστωτικό κίνδυνο

Ακολουθούν οι βασικοί παράγοντες που μπορούν να επηρεάσουν τον πιστωτικό κίνδυνο. Είναι σημαντικό να σημειωθεί ότι αυτοί οι παράγοντες είναι αλληλένδετοι και η σχετική σημασία τους μπορεί να ποικίλλει ανάλογα με το πλαίσιο και τον τύπο της πίστωσης που εξετάζεται *[2]*.

1. Οικονομική σταθερότητα (Financial Stability)

- **Εισόδημα και ταμειακές ροές**

Το επίπεδο εισοδήματος και η σταθερότητα του δανειολήπτη είναι σημαντικοί δείκτες της ικανότητάς τους να δημιουργούν επαρκή κεφάλαια για την αποπληρωμή των χρεών.

- **Αναλογία χρέους προς εισόδημα**

Ο λόγος χρέους προς εισόδημα του δανειολήπτη, ο οποίος συγκρίνει τις συνολικές υποχρεώσεις του χρέους με το εισόδημά του, παρέχει πληροφορίες για το βάρος του χρέους και την ικανότητά τους να αναλάβουν πρόσθετο χρέος.

- **Περιουσιακά στοιχεία και υποχρεώσεις**

Η αξιολόγηση των περιουσιακών στοιχείων του δανειολήπτη (όπως ακίνητα, επενδύσεις ή αποταμιεύσεις) και των υποχρεώσεων (όπως υφιστάμενα χρέη και οικονομικές υποχρεώσεις) βοηθά στον προσδιορισμό της συνολικής οικονομικής του θέσης και της ικανότητάς του να ανταποκρίνεται στις μελλοντικές υποχρεώσεις πληρωμών.

- **Χρηματοοικονομικοί δείκτες**

Η ανάλυση χρηματοοικονομικών δεικτών όπως οι δείκτες ρευστότητας (liquidity ratios), οι δείκτες κερδοφορίας και οι δείκτες μόχλευσης βοηθά στην αξιολόγηση της οικονομικής υγείας του δανειολήπτη και της ικανότητάς του να διαχειρίζεται το χρέος.

2. Πιστωτικό ιστορικό (Credit History)

- **Πιστωτικό αποτέλεσμα**

Το πιστωτικό αποτέλεσμα (credit score) του δανειολήπτη, που δημιουργείται από τα πιστωτικά γραφεία, παρέχει μια αριθμητική αναπαράσταση της πιστοληπτικής του ικανότητας με βάση παράγοντες όπως το ιστορικό πληρωμών, η χρήση της πίστωσης, η διάρκεια του πιστωτικού ιστορικού και οι τύποι πίστωσης.

- **Συμπεριφορά πληρωμών**

Η επανεξέταση του ιστορικού πληρωμών του δανειολήπτη αποκαλύπτει το ιστορικό του (track record) στην εκπλήρωση προηγούμενων οικονομικών υποχρεώσεων, συμπεριλαμβανομένων τυχόν καθυστερήσεων πληρωμών, αθετήσεων ή καθυστερήσεων.

- **Πιστωτική χρήση**

Η αξιολόγηση του δείκτη πιστωτικής χρήσης (credit utilization) του δανειολήπτη, ο οποίος συγκρίνει τα υπόλοιπα των πιστωτικών καρτών του με τα πιστωτικά του όρια, βοηθά στη μέτρηση της εξάρτησής του από την πίστωση και της ικανότητάς του να τη διαχειρίζεται υπεύθυνα.

- **Ερωτήματα πίστωσης**

Ο αριθμός και η συχνότητα των πρόσφατων ερωτημάτων για πίστωση (credit inquiries) μπορεί να υποδηλώνει την πιθανή οικονομική πίεση ή την υπερβολική δανειοληπτική συμπεριφορά του δανειολήπτη.

3. Οικονομικές συνθήκες (Economic Conditions)

- **Επιτόκια**

Οι διακυμάνσεις των επιτοκίων (interest rates) μπορεί να επηρεάσουν τις δυνατότητες αποπληρωμής των δανειοληπτών, ειδικά για δάνεια με κυμαινόμενο επιτόκιο. Τα υψηλότερα επιτόκια μπορεί να αυξήσουν το κόστος δανεισμού και ενδεχομένως να επιβαρύνουν τα οικονομικά των δανειοληπτών.

- **Ποσοστά ανεργίας**

Τα υψηλά ποσοστά ανεργίας μπορεί να αυξήσουν την πιθανότητα οι δανειολήπτες να αντιμετωπίζουν οικονομικές δυσκολίες και να μην μπορούν να αποπληρώσουν τα χρέη τους.

- **Ανάπτυξη ΑΕΠ**

Ο συνολικός ρυθμός οικονομικής ανάπτυξης μπορεί να επηρεάσει τις επιχειρήσεις και τα άτομα, επηρεάζοντας την ικανότητά τους να δημιουργούν εισόδημα και να αποπληρώνουν χρέη.

4. Κίνδυνοι κλάδου (Industry Risks)

- **Συνθήκες αγοράς**

Η αξιολόγηση των τάσεων του κλάδου, της ανταγωνιστικότητας της αγοράς, των ρυθμιστικών παραγόντων και των τεχνολογικών εξελίξεων βοηθά στη μέτρηση της οικονομικής υγείας και σταθερότητας των δανειοληπτών σε συγκεκριμένους κλάδους.

- **Επιχειρηματικές ιδιαιτερότητες**

Η αξιολόγηση του επιχειρηματικού μοντέλου, της θέσης στην αγορά, του ανταγωνιστικού πλεονεκτήματος και της πελατειακής βάσης ενός δανειολήπτη μπορεί να παρέχει πληροφορίες για τις δυνατότητές του για επιτυχία και πιστοληπτική ικανότητα.

- **Βιομηχανικοί κανονισμοί**

Ρυθμιστικές αλλαγές ή διαταραχές σε συγκεκριμένους κλάδους μπορεί να επηρεάσουν την ικανότητα των δανειοληπτών να παράγουν έσοδα και να εκπληρώσουν τις οικονομικές τους υποχρεώσεις.

5. Νομικοί και πολιτικοί παράγοντες (Legal and Political factors)

- **Νομικό πλαίσιο**

Το νομικό πλαίσιο (legal framework) και οι κανονισμοί που διέπουν τον δανεισμό καθώς και την ανάκτηση του χρέους (debt recovery), μπορούν να επηρεάσουν τον πιστωτικό κίνδυνο.

- **Πολιτική σταθερότητα**

Η πολιτική αστάθεια ή οι αλλαγές στις κυβερνητικές πολιτικές μπορεί να επηρεάσουν τις οικονομικές συνθήκες, οι οποίες, με τη σειρά τους, επηρεάζουν τον πιστωτικό κίνδυνο.

6. Εξασφάλιση και Εγγυήσεις (Collateral and Guarantees)

- **Εξασφάλιση**

Η αποδοχή περιουσιακών στοιχείων που μπορούν να δεσμευθούν σε περίπτωση αθέτησης υποχρεώσεων παρέχει πρόσθετη ασφάλεια και μειώνει τις πιθανές ζημίες για τους δανειστές.

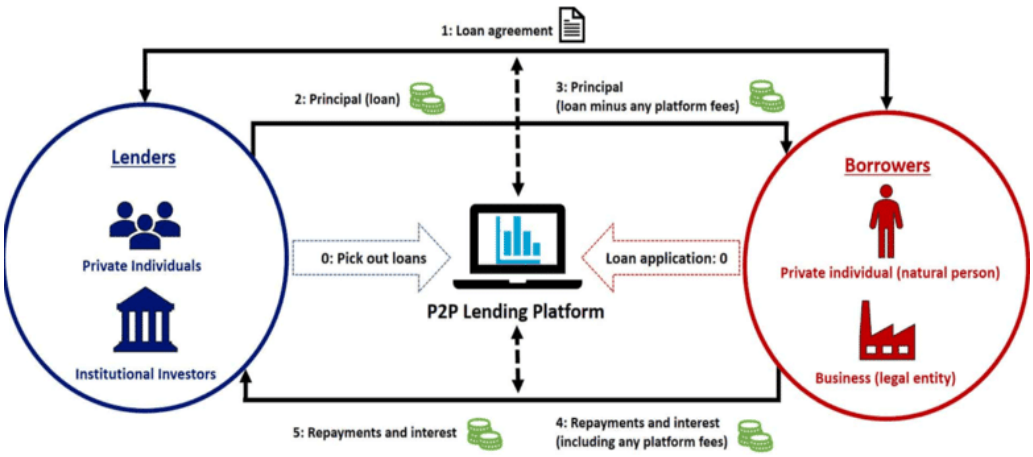
- **Εγγυήσεις**

Η απόκτηση εγγυήσεων από τρίτους, όπως συνυπογράφοντες ή εγγυητές, μπορεί να μειώσει τον πιστωτικό κίνδυνο διασφαλίζοντας εναλλακτικές πηγές αποπληρωμής.

7. Πιστωτική αξιολόγηση και αναδοχή (Credit Assessment and Underwriting)

Η διεξοδική αξιολόγηση των οικονομικών πληροφοριών, του πιστωτικού ιστορικού των δανειοληπτών και η διενέργεια αξιολογήσεων κινδύνου πριν από τη χορήγηση πίστωσης συμβάλλει στον μετριασμό του πιστωτικού κινδύνου.

Οι αποτελεσματικές πρακτικές αναδοχής, συμπεριλαμβανομένων των αναλογιών δανείου προς αξία, των συστημάτων αξιολόγησης κινδύνου και των μοντέλων πιστωτικού κινδύνου, βοηθούν στην αξιολόγηση και διαχείριση του πιστωτικού κινδύνου.



Εικόνα 1 *Επισκόπηση πλαισίου δανεισμού*

1.4 Η πρόβλεψη του Πιστωτικού Κινδύνου

Η πρόβλεψη πιστωτικού κινδύνου είναι ένα κρίσιμο έργο στον χρηματοπιστωτικό κλάδο που περιλαμβάνει την αξιολόγηση της πιθανότητας οι δανειολήπτες να αθετήσουν τις δανειακές τους υποχρεώσεις. Οι δανειστές και τα χρηματοπιστωτικά ιδρύματα χρησιμοποιούν μοντέλα πιστωτικού κινδύνου για να αξιολογήσουν την πιστοληπτική ικανότητα ιδιωτών, επιχειρήσεων ή άλλων οντοτήτων που αναζητούν δάνεια ή πιστωτικές διευκολύνσεις. Ο στόχος είναι να ελαχιστοποιηθούν οι πιθανές ζημίες με τον εντοπισμό δανειοληπτών υψηλού κινδύνου και τη λήψη τεκμηριωμένων αποφάσεων δανεισμού *[3]*. Είναι σημαντικό να σημειωθεί ότι η πρόβλεψη πιστωτικού κινδύνου είναι ένα σύνθετο και εξελισσόμενο πεδίο και υπάρχουν πολλές παραλλαγές και πρόσθετες εκτιμήσεις ανάλογα με το συγκεκριμένο πλαίσιο (context), τις απαιτήσεις και τα διαθέσιμα δεδομένα.

Ακολουθεί μια ανάλυση και καταγραφή του προβλήματος πρόβλεψης πιστωτικού κινδύνου *[3]*:

1. Συλλογή δεδομένων (Data Collection)

Το πρώτο βήμα στην πρόβλεψη πιστωτικού κινδύνου είναι η συλλογή σχετικών δεδομένων. Αυτό περιλαμβάνει συνήθως οικονομικές πληροφορίες, όπως εισόδημα, περιουσιακά στοιχεία, υποχρεώσεις, πιστωτικό ιστορικό και συμπεριφορά πληρωμών. Οι πρόσθετες πηγές δεδομένων μπορεί να περιλαμβάνουν ιστορικό απασχόλησης, δημογραφικούς παράγοντες και μακροοικονομικούς δείκτες.

2. Προεπεξεργασία δεδομένων (Data Preprocessing)

Μόλις συλλεχθούν τα δεδομένα, πρέπει να υποβληθούν σε επεξεργασία και να προετοιμαστούν για ανάλυση. Αυτό περιλαμβάνει καθαρισμό δεδομένων (data cleaning), χειρισμό τιμών που λείπουν (missing values), αντιμετώπιση ακραίων τιμών (outliers) και μετασχηματισμό των μεταβλητών εάν αυτό είναι απαραίτητο. Η προεπεξεργασία δεδομένων αποσκοπεί στη διασφάλιση της ποιότητας και της συνέπειας των δεδομένων.

3. Επιλογή Χαρακτηριστικών (Feature Selection)

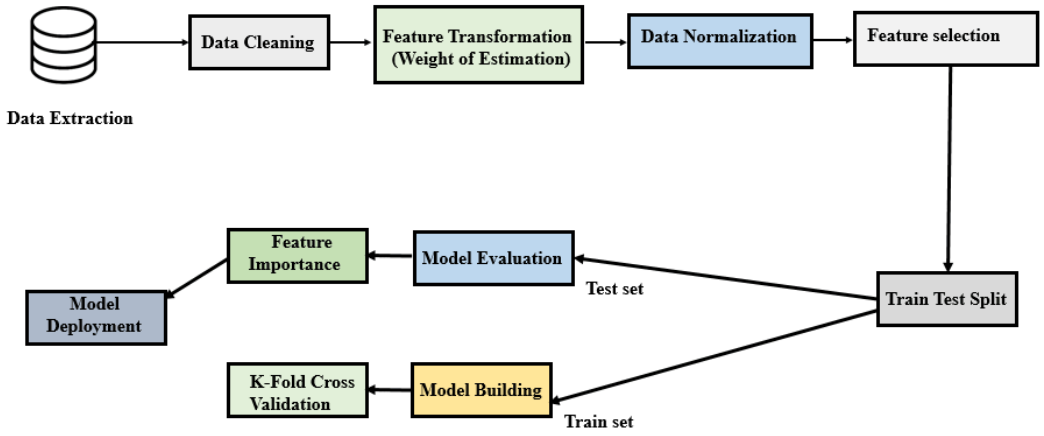
Σε αυτό το βήμα, επιλέγονται ή δημιουργούνται τα πιο ενημερωτικά χαρακτηριστικά (μεταβλητές) για την εκπαίδευση του μοντέλου πρόβλεψης πιστωτικού κινδύνου. Οι σχετικές γνώσεις και οι στατιστικές τεχνικές μπορούν να εφαρμοστούν για τον εντοπισμό κατάλληλων χαρακτηριστικών. Η επιλογή χαρακτηριστικών μπορεί να περιλαμβάνει τη δημιουργία νέων μεταβλητών, τον συνδυασμό ή τον μετασχηματισμό υπαρχουσών ή τη μείωση της διάστασης μέσω τεχνικών όπως η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA).

4. Επιλογή μοντέλου (Model Selection)

Μπορούν να χρησιμοποιηθούν διάφορα μοντέλα μηχανικής μάθησης και αντιστοίχως κατάλληλα στατιστικά μοντέλα για την πρόβλεψη πιστωτικού κινδύνου. Τα μοντέλα που χρησιμοποιούνται συνήθως περιλαμβάνουν λογιστική παλινδρόμηση (Logistic Regression), δέντρα αποφάσεων (Decision Trees), τυχαία δάση (Random Forests), Support Vector Machines (SVM), αλγόριθμους ενίσχυσης κλίσης (π.χ. XGBoost, LightGBM) και νευρωνικά δίκτυα (Neural Networks). Η επιλογή του μοντέλου εξαρτάται από διάφορους παράγοντες όπως το μέγεθος των δεδομένων, οι απαιτήσεις του προβλήματος και η επιθυμητή απόδοση πρόγνωσης.

5. Εκπαίδευση μοντέλου (Model Training)

Σε αυτό το βήμα, το επιλεγμένο μοντέλο εκπαιδεύεται χρησιμοποιώντας ιστορικά δεδομένα με γνωστά αποτελέσματα (π.χ. αθετήσεις δανείων). Το σύνολο δεδομένων χωρίζεται σε ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο επικύρωσης (validation set). Το μοντέλο μαθαίνει μοτίβα και σχέσεις στα δεδομένα εκπαίδευσης, βελτιστοποιώντας τις παραμέτρους του για να ελαχιστοποιήσει το σφάλμα πρόβλεψης (prediction error). Το σύνολο επικύρωσης χρησιμοποιείται για να τελειοποιήσει το μοντέλο και να αξιολογήσει την απόδοσή του.



Εικόνα 2 *Επιλογή κατάλληλου μοντέλου Μηχανικής Μάθησης*

6. Αξιολόγηση μοντέλου (Model Evaluation)

Μόλις το μοντέλο εκπαιδευτεί, η απόδοσή του αξιολογείται χρησιμοποιώντας διάφορες μετρήσεις, όπως ακρίβεια (accuracy), ανάκληση (recall) και καμπύλη χαρακτηριστικών λειτουργίας δέκτη (ROC curve). Αυτές οι μετρικές μετρούν την ικανότητα του μοντέλου να προβλέπει σωστά τους κακοπληρωτές και τους μη προεπιλεγμένους. Οι τεχνικές της διασταυρούμενης επικύρωσης (cross validation techniques), όπως η διασταυρούμενη επικύρωση k-fold cross validation, μπορούν να παρέχουν πιο αξιόπιστες εκτιμήσεις για την απόδοση του μοντέλου.

7. Ανάπτυξη μοντέλου (Model Deployment)

Μετά από ικανοποιητική αξιολόγηση, το μοντέλο μπορεί να αναπτυχθεί σε κατάλληλο περιβάλλον παραγωγής. Οι νέες πιστωτικές αιτήσεις μπορούν να βαθμολογηθούν χρησιμοποιώντας το εκπαιδευμένο μοντέλο για την πρόβλεψη της πιθανότητας αθέτησης υποχρεώσεων. Η έξοδος του μοντέλου μπορεί να χρησιμοποιηθεί ως εισροή για διαδικασίες λήψης αποφάσεων, όπως η έγκριση ή η απόρριψη αιτήσεων δανείου ή ο καθορισμός των όρων και των προϋποθέσεων ενός δανείου.

8. Παρακολούθηση και Συντήρηση Μοντέλου (Monitoring and Model Maintenance)

Τα μοντέλα πρόβλεψης πιστωτικού κινδύνου θα πρέπει να παρακολουθούνται τακτικά για να διασφαλίζεται η διαρκής ακρίβεια (ongoing accuracy) και συνάφειά τους. Καθώς οι οικονομικές συνθήκες ή η συμπεριφορά των δανειοληπτών αλλάζουν, τα μοντέλα ενδέχεται να απαιτούν ενημερώσεις (updates) ή επανεκπαίδευση (retraining). Η παρακολούθηση μπορεί να περιλαμβάνει την καταγραφή μετρήσεων απόδοσης του μοντέλου, την αξιολόγηση της βαθμονόμησης του μοντέλου και τον εντοπισμό και την αντιμετώπιση πιθανών προκαταλήψεων (biases) ή μετατόπισης δεδομένων (data drift).

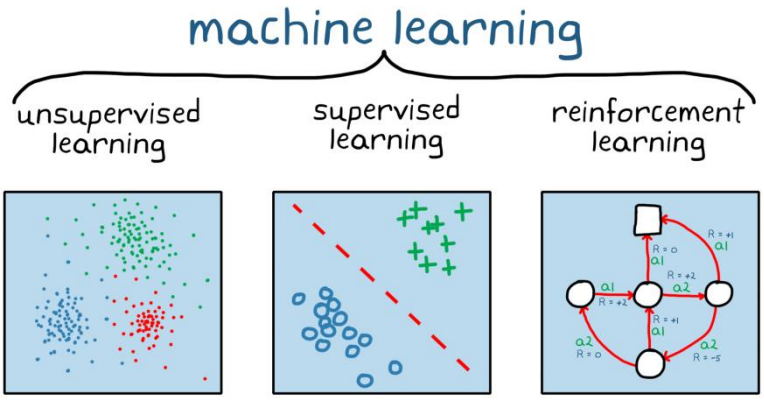
ΚΕΦΑΛΑΙΟ 2

ΜΕΘΟΔΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

2.1 Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση είναι ένας κλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence) που εστιάζει σε αλγόριθμους εκπαίδευσης για να μαθαίνει από δεδομένα και να κάνει προβλέψεις ή αποφάσεις χωρίς να έχει προγραμματιστεί ρητά. Παρέχει στον αλγόριθμο ένα σύνολο δεδομένων (dataset), όπου οι είσοδοι συνδέονται με τις αντίστοιχες εξόδους *[5]*. Μέσω μιας διαδικασίας ανάλυσης των δεδομένων και προσαρμογής εσωτερικών παραμέτρων, ο αλγόριθμος μαθαίνει να αναγνωρίζει μοτίβα και σχέσεις, επιτρέποντάς του να κάνει ακριβείς προβλέψεις για νέα δεδομένα. Η μηχανική μάθηση μπορεί να κατηγοριοποιηθεί σε εποπτευόμενη μάθηση (supervised learning) χρησιμοποιώντας labeled data, μάθηση χωρίς επίβλεψη (unsupervised learning) και ενισχυτική μάθηση (reinforcement learning) δηλαδή μάθηση μέσω αλληλεπίδρασης με ένα περιβάλλον και ανταμοιβές ή ποινές.

Η διαδικασία της μηχανικής μάθησης περιλαμβάνει πολλά βασικά βήματα. Αρχικά, συλλέγονται και προετοιμάζονται τα δεδομένα, συμπεριλαμβανομένου του καθαρισμού (cleaning), της προεπεξεργασίας (preprocessing) και της εξαγωγής χαρακτηριστικών (feature extraction) *[6]*. Στη συνέχεια, επιλέγεται ένα κατάλληλο μοντέλο μηχανικής μάθησης βάσει του προβλήματος και των διαθέσιμων δεδομένων. Το μοντέλο αυτό εκπαιδεύεται στο σύνολο δεδομένων, όπου μαθαίνει να αντιστοιχίζει εισόδους σε εξόδους προσαρμόζοντας τις εσωτερικές του παραμέτρους. Το εκπαιδευμένο μοντέλο αξιολογείται σε ξεχωριστό σύνολο δεδομένων για να αξιολογηθεί η απόδοσή του και η ικανότητα γενίκευσης. Μόλις το μοντέλο κριθεί ικανοποιητικό, μπορεί να αναπτυχθεί για να γίνουν προβλέψεις ή αποφάσεις σχετικά με νέα δεδομένα.



Εικόνα 3 *Κατηγορίες Μηχανικής Μάθησης*

Η μηχανική μάθηση έχει γίνει διάχυτη (pervasive) σε διάφορους τομείς και εφαρμογές. Χρησιμοποιείται για αναγνώριση εικόνων και αντικειμένων, επεξεργασία φυσικής γλώσσας, ανίχνευση απάτης (fraud detection), συστήματα συστάσεων (recommendation systems), χρηματοοικονομικές προβλέψεις, διαγνωστικά υγειονομικής περίθαλψης, αυτόνομα οχήματα και άλλα. Η δύναμη της μηχανικής μάθησης έγκειται στην ικανότητά της να αποκαλύπτει μοτίβα και ιδέες από τεράστιες ποσότητες δεδομένων, να αυτοματοποιεί πολύπλοκες εργασίες και να βελτιώνει τις διαδικασίες λήψης αποφάσεων. Καθώς η τεχνολογία προχωρά και περισσότερα δεδομένα γίνονται διαθέσιμα, η μηχανική μάθηση συνεχίζει να εξελίσσεται και να βρίσκει νέες εφαρμογές σε διάφορους τομείς, οδηγώντας την καινοτομία και διαμορφώνοντας το μέλλον.

2.2 Μέθοδοι Μηχανικής Μάθησης στην πρόβλεψη πιστωτικού κινδύνου

Οι μέθοδοι μηχανικής μάθησης έχουν φέρει επανάσταση στην πρόβλεψη πιστωτικού κινδύνου παρέχοντας πιο ακριβή και αποτελεσματικά μοντέλα για την αξιολόγηση της πιστοληπτικής ικανότητας των δανειοληπτών. Αυτές οι μέθοδοι αξιοποιούν προηγμένους αλγόριθμους και μεγάλα σύνολα δεδομένων (datasets) για τον εντοπισμό προτύπων και την πρόβλεψη της πιθανότητας αθέτησης υποχρεώσεων ή παραβατικότητας.

Οι τεχνικές μηχανικής μάθησης υπερέχουν στην πρόβλεψη πιστωτικού κινδύνου λόγω της ικανότητάς τους να χειρίζονται μεγάλους όγκους διαφορετικών δεδομένων. Μπορούν να ενσωματώσουν διάφορους τύπους πληροφοριών, όπως πιστωτικό ιστορικό (credit history), οικονομικές καταστάσεις, λεπτομέρειες απασχόλησης και μακροοικονομικούς δείκτες. Αξιοποιώντας αυτά τα πλούσια δεδομένα, τα μοντέλα μηχανικής μάθησης μπορούν να καταγράψουν πολύπλοκες σχέσεις και κρυφά μοτίβα που μπορεί να χάνουν τα παραδοσιακά μοντέλα. Αυτή η προσέγγιση βάσει δεδομένων ενισχύει την ακρίβεια και την ευρωστία των μοντέλων πρόβλεψης πιστωτικού κινδύνου.

Οι αλγόριθμοι μηχανικής μάθησης είναι εξαιρετικά ευέλικτοι και προσαρμόσιμοι, επιτρέποντας στα μοντέλα να προσαρμόζονται στις μεταβαλλόμενες δυναμικές του πιστωτικού κινδύνου. Τα μοντέλα μπορούν να εκπαιδευτούν ώστε να μαθαίνουν από ιστορικά δεδομένα και να προσαρμόζονται σε νέα πρότυπα ή τάσεις καθώς εμφανίζονται. Αυτή η ικανότητα είναι ιδιαίτερα πολύτιμη σε δυναμικές πιστωτικές αγορές, όπου οι παράγοντες κινδύνου μπορεί να εξελιχθούν με την πάροδο του χρόνου. Επιπλέον, τα μοντέλα μηχανικής εκμάθησης μπορούν εύκολα να επανεκπαιδευτούν και να ενημερωθούν για να ενσωματώσουν νέα δεδομένα, διασφαλίζοντας τη συνεχή ακρίβεια και συνάφειά τους *[7]*.

Οι μέθοδοι μηχανικής μάθησης έχουν επιδείξει ανώτερη προγνωστική απόδοση σε σύγκριση με τα παραδοσιακά μοντέλα αξιολόγησης πιστοληπτικής ικανότητας. Χρησιμοποιώντας προηγμένους αλγόριθμους όπως τυχαία δάση, μηχανές υποστήριξης διανυσμάτων ή νευρωνικά δίκτυα, αυτά τα μοντέλα μπορούν να καταγράψουν πολύπλοκες, μη γραμμικές σχέσεις μεταξύ μεταβλητών. Η βελτιωμένη προγνωστική ακρίβεια επιτρέπει στους δανειστές να λαμβάνουν πιο ενημερωμένες αποφάσεις, με αποτέλεσμα καλύτερη διαχείριση κινδύνου, μειωμένα ποσοστά αθέτησης υποχρεώσεων και βελτιωμένη κερδοφορία *[6]*.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι οι μέθοδοι μηχανικής μάθησης παρουσιάζουν επίσης προκλήσεις στην πρόβλεψη πιστωτικού κινδύνου. Απαιτούν σημαντικούς υπολογιστικούς πόρους και τεχνογνωσία για ανάπτυξη, ανάπτυξη και συντήρηση. Επιπλέον, μπορεί να είναι επιρρεπείς στο overfitting αν δεν έχουν επικυρωθεί σωστά. Ο πολλαπλός τρόπος ερμηνείας (interpretability) είναι μια άλλη ανησυχία, καθώς ορισμένα μοντέλα μηχανικής μάθησης λειτουργούν ως «μαύρα κουτιά (black boxes)» που στερούνται διαφάνειας στην εξήγηση των προβλέψεών τους. Ωστόσο, η συνεχιζόμενη έρευνα και η πρόοδος στην Τεχνητή Νοημοσύνη στοχεύουν στην αντιμετώπιση αυτών των προκλήσεων και στη διασφάλιση της υπεύθυνης και αποτελεσματικής χρήσης της μηχανικής μάθησης στην πρόβλεψη πιστωτικού κινδύνου.

2.3 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks - ANN) είναι υπολογιστικά μοντέλα εμπνευσμένα από τη δομή και τη λειτουργία των βιολογικών νευρωνικών δικτύων στον εγκέφαλο. Είναι ένα υποσύνολο αλγορίθμων μηχανικής μάθησης που βρίσκουν εφαρμογή στην επίλυση σύνθετων προβλημάτων κάνοντας χρήση από μεγάλες ποσότητες δεδομένων. Τα ANN αποτελούνται από διασυνδεδεμένους κόμβους ή τεχνητούς νευρώνες, οργανωμένους σε στρώματα (layers) *[9].*

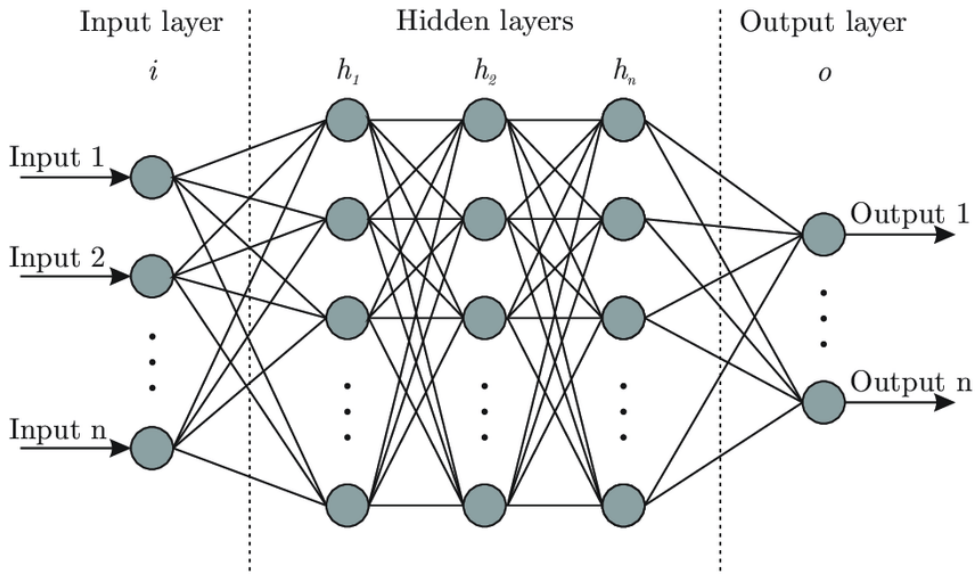
Η δομή ενός Τεχνητού Νευρωνικού Δικτύου (ANN) συνήθως περιλαμβάνει ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα (hidden layers) και ένα επίπεδο εξόδου (output layer). Το επίπεδο εισόδου λαμβάνει ακατέργαστα δεδομένα, όπως εικόνες, κείμενο ή αριθμητικές τιμές. Κάθε κόμβος στο επίπεδο εισόδου αντιπροσωπεύει ένα χαρακτηριστικό (attribute) που αφορά τα δεδομένα. Τα κρυφά επίπεδα, που βρίσκονται μεταξύ των επιπέδων εισόδου και εξόδου, εκτελούν υπολογισμούς στα δεδομένα εισόδου. Το επίπεδο εξόδου παρέχει το τελικό αποτέλεσμα ή την πρόβλεψη.

Κάθε κόμβος σε ένα Τεχνητό Νευρωνικό Δίκτυο σχετίζεται με ένα βάρος (weight), το οποίο καθορίζει την ισχύ ή τη σημασία της εισόδου του. Κατά τη διάρκεια της εκπαίδευσης (training), αυτά τα βάρη προσαρμόζονται για τη βελτιστοποίηση της απόδοσης του δικτύου. Η συνάρτηση ενεργοποίησης (activation function), που εφαρμόζεται στο σταθμισμένο άθροισμα των εισόδων σε κάθε κόμβο, εισάγει μη γραμμικότητα στο μοντέλο. Καθορίζει την έξοδο ή την ενεργοποίηση του κόμβου, βάσει του οποίου οι πληροφορίες μεταδίδονται στο επόμενο επίπεδο.

Η προς τα εμπρός διάδοση (forward propagation) είναι η διαδικασία με την οποία τα δεδομένα ρέουν μέσω του δικτύου από το επίπεδο εισόδου στο επίπεδο εξόδου. Σε κάθε επίπεδο, οι κόμβοι λαμβάνουν εισόδους από το προηγούμενο επίπεδο, υπολογίζουν το σταθμισμένο άθροισμα, εφαρμόζουν τη συνάρτηση ενεργοποίησης και περνούν την έξοδο στο επόμενο επίπεδο. Αυτή η διαδοχική επεξεργασία επιτρέπει στο δίκτυο να μετασχηματίσει και να εξάγει σχετικά χαρακτηριστικά από τα δεδομένα εισόδου.

Η εκπαίδευση ενός Τεχνητού Νευρωνικού Δικτύου περιλαμβάνει την ελαχιστοποίηση του σφάλματος ή της ασυμφωνίας μεταξύ της προβλεπόμενης εξόδου του και της επιθυμητής εξόδου. Αυτό επιτυγχάνεται μέσω μιας διαδικασίας που ονομάζεται backpropagation. Κατά τη διάρκεια της backpropagation, το δίκτυο υπολογίζει την κλίση (gradient) της συνάρτησης σφάλματος σε σχέση με κάθε βάρος (weight) στο

δίκτυο. Αυτή η κλίση χρησιμοποιείται στη συνέχεια για την ενημέρωση των βαρών χρησιμοποιώντας αλγόριθμους βελτιστοποίησης όπως η στοχαστική κλίση κατάβασης (SGD) ή οι παραλλαγές της. Προσαρμόζοντας επαναληπτικά τα βάρη με βάση το σφάλμα, το δίκτυο βελτιώνει σταδιακά την προγνωστική του ικανότητα *[12].*



Εικόνα 4 *Τεχνητό Νευρωνικό Δίκτυο (ANN)*

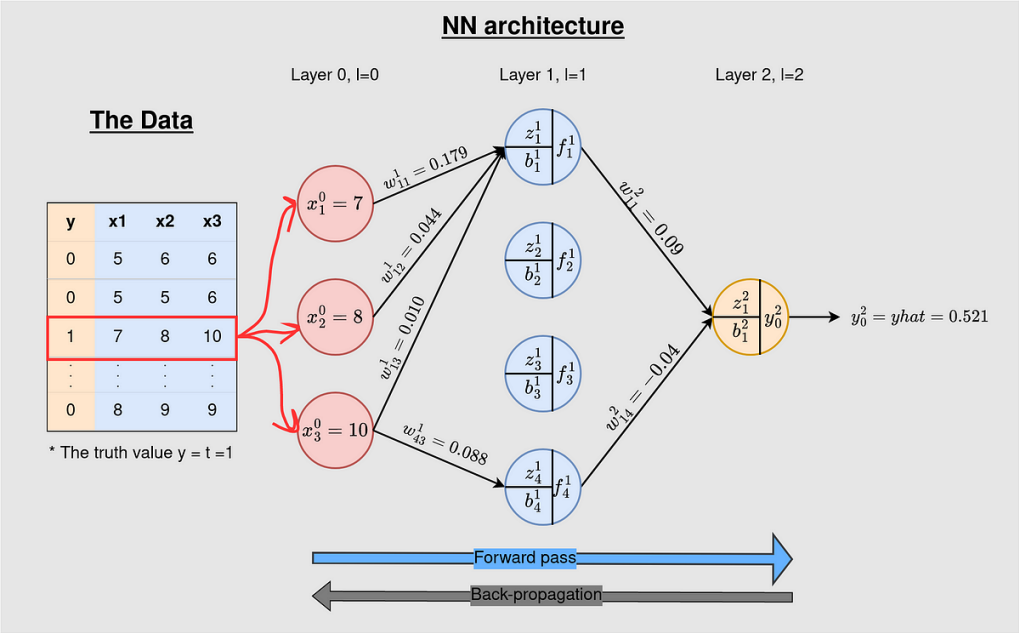
Τα δεδομένα εκπαίδευσης (training data) διαδραματίζουν κρίσιμο ρόλο στη διαδικασία μάθησης των Τεχνητών Νευρωνικών Δικτύων. Αποτελούνται από παραδείγματα (labeled examples), όπου τα δεδομένα εισόδου αντιστοιχίζονται με την αντίστοιχη σωστή έξοδο. Το δίκτυο συγκρίνει την προβλεπόμενη έξοδο του με τη γνωστή έξοδο και προσαρμόζει ανάλογα τα βάρη του. Αυτή η διαδικασία επιτρέπει στο δίκτυο να γενικεύει και να κάνει ακριβείς προβλέψεις για νέα δεδομένα. Τα κρυφά επίπεδα (hidden layers) αποτελούν βασικό συστατικό των Τεχνητών Νευρωνικών Δικτύων (ANN), επιτρέποντάς τους να καταγράφουν πολύπλοκα μοτίβα και σχέσεις στα δεδομένα. Κάθε επίπεδο μαθαίνει να αναπαριστά όλο και πιο αφηρημένα χαρακτηριστικά της εισόδου. Τα Βαθιά Νευρωνικά Δίκτυα (Deep Neural Networks), με πολλαπλά κρυφά επίπεδα, έχουν τη δυνατότητα να μαθαίνουν ιεραρχικές αναπαραστάσεις, επιτρέποντάς τους να χειρίζονται πιο περίπλοκες εργασίες.

Οι συναρτήσεις ενεργοποίησης (activation functions) που χρησιμοποιούνται στα Τεχνητά Νευρωνικά Δίκτυα εισάγουν μη γραμμικότητα στο μοντέλο. Οι κοινές συναρτήσεις ενεργοποίησης περιλαμβάνουν τη σιγμοειδή, την υπερβολική εφαπτομένη (tanh) και την ανορθωμένη γραμμική μονάδα (ReLU). Αυτές οι συναρτήσεις επιτρέπουν στο δίκτυο να μοντελοποιεί περίπλοκες σχέσεις και να συλλαμβάνει μη γραμμικά μοτίβα στα δεδομένα. Τα βάρη σε ένα Τεχνητό Νευρωνικό Δίκτυο καθορίζουν την ισχύ των συνδέσεων μεταξύ των κόμβων. Αρχικά, σε αυτά τα βάρη εκχωρούνται τυχαίες τιμές. Κατά τη διάρκεια της εκπαίδευσης (training), το δίκτυο προσαρμόζει αυτά τα βάρη με βάση το σφάλμα μεταξύ της προβλεπόμενης εξόδου και της επιθυμητής εξόδου. Το μέγεθος των προσαρμογών βάρους

καθορίζεται από τον ρυθμό εκμάθησης, ο οποίος ελέγχει το μέγεθος του βήματος κατά τη διαδικασία βελτιστοποίησης.

Τα Τεχνητά Νευρωνικά Δίκτυα μπορούν να έχουν διάφορες αρχιτεκτονικές, συμπεριλαμβανομένων των δικτύων προώθησης (feedforward networks), επαναλαμβανόμενων δικτύων (recurrent networks) και συνελικτικών δικτύων (convolutional networks). Τα δίκτυα προώθησης επεξεργάζονται δεδομένα σε μία μόνο κατεύθυνση, από την είσοδο στην έξοδο, ενώ τα επαναλαμβανόμενα δίκτυα έχουν συνδέσεις που σχηματίζουν βρόχους, επιτρέποντάς τους να επεξεργάζονται διαδοχικά ή εξαρτώμενα από το χρόνο δεδομένα. Τα συνελικτικά δίκτυα είναι εξειδικευμένα για την ανάλυση δεδομένων που μοιάζουν με πλέγμα, όπως εικόνες, εφαρμόζοντας λειτουργίες συνέλιξης *[10]*.

Η διαδικασία εκμάθησης των Τεχνητών Νευρωνικών Δικτύων περιλαμβάνει έναν επαναληπτικό βρόχο ανάδρασης. Το δίκτυο κάνει προβλέψεις, τις συγκρίνει με τις γνωστές εξόδους, προσαρμόζει τα βάρη και επαναλαμβάνει αυτή τη διαδικασία μέχρι να επιτευχθεί το επιθυμητό επίπεδο ακρίβειας. Αυτή η επαναληπτική διαδικασία είναι υπολογιστικά εντατική και συχνά απαιτεί μεγάλες ποσότητες δεδομένων για αποτελεσματική εκπαίδευση. Τα Τεχνητά Νευρωνικά Δίκτυα απαιτούν σημαντικούς υπολογιστικούς πόρους, ειδικά για την εκπαίδευση σε βαθιά δίκτυα (deep networks) με πολλούς κόμβους και επίπεδα. Υψηλής απόδοσης υπολογιστική υποδομή, όπως μονάδες γραφικής επεξεργασίας (GPU) ή εξειδικευμένο υλικό όπως μονάδες επεξεργασίας τανυστών (tensors) (TPUs), χρησιμοποιείται συχνά για την επιτάχυνση των διαδικασιών εκπαίδευσης και συμπερασμάτων.



Εικόνα 5 *Η διαδικασία backpropagation*

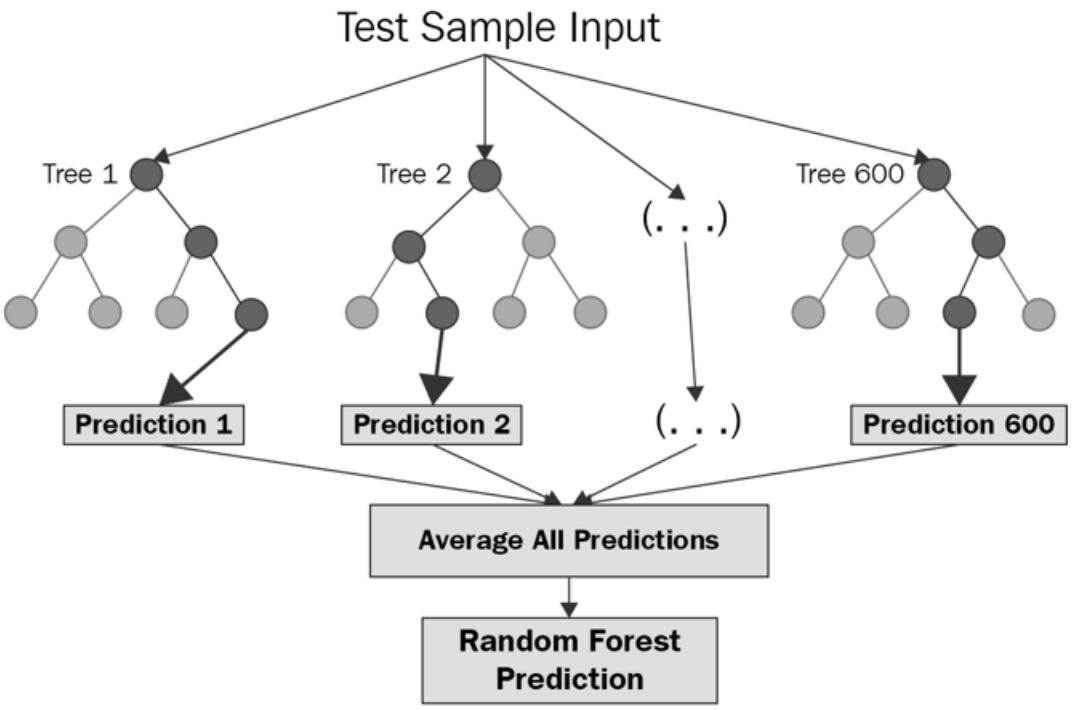
Τα Τεχνητά Νευρωνικά Δίκτυα βρίσκουν εφαρμογή σε ένα ευρύ φάσμα πεδίων. Έχει αποδειχθεί βρίσκουν επιτυχία στην αναγνώριση εικόνας και ομιλίας, στην επεξεργασία φυσικής γλώσσας, στα συστήματα συστάσεων (recommendation systems), στα αυτόνομα οχήματα και σε πολλούς άλλους τομείς. Τα Τεχνητά

Νευρωνικά Δίκτυα υπερέχουν σε εργασίες που περιλαμβάνουν αναγνώριση προτύπων (pattern recognition), ταξινόμηση (classification), παλινδρόμηση (regression) και λήψη αποφάσεων με βάση μεγάλα σύνολα δεδομένων.

2.4 Τυχαία Δάση (Random Forests)

Τα Random Forests είναι ένας δημοφιλής αλγόριθμος μηχανικής μάθησης που ανήκει στην οικογένεια εκμάθησης συνόλου (ensemble learning family). Χρησιμοποιούνται ευρέως για εργασίες ταξινόμησης (classification) και παλινδρόμησης (regression) λόγω της ικανότητάς τους να χειρίζονται πολύπλοκα σύνολα δεδομένων *[12]*. Τα Random Forests λειτουργούν συνδυάζοντας πολλαπλά δέντρα αποφάσεων για να κάνουν προβλέψεις.

Η εκμάθηση συνόλου (ensemble learning) περιλαμβάνει το συνδυασμό των προβλέψεων πολλαπλών μοντέλων για τη βελτίωση της συνολικής απόδοσης. Στην περίπτωση των Random Forests, το σύνολο αποτελείται από δέντρα απόφασης (decision trees). Κάθε δέντρο απόφασης στο Random Forest εκπαιδεύεται σε διαφορετικό υποσύνολο των δεδομένων εκπαίδευσης (training data), δειγματοληπτείται τυχαία με αντικατάσταση. Αυτή η διαδικασία είναι γνωστή ως bootstrapping.



Εικόνα 6 *Αναπαράσταση ενός Random Forest*

Κατά την κατασκευή κάθε δέντρου απόφασης, ένα τυχαίο υποσύνολο χαρακτηριστικών λαμβάνεται υπόψη σε κάθε διαχωρισμό. Επιλέγοντας μόνο ένα υποσύνολο χαρακτηριστικών, τα Random Forests εισάγουν την τυχαιότητα και την ποικιλομορφία μεταξύ των δέντρων, μειώνοντας τον κίνδυνο υπερβολικής προσαρμογής και αυξάνοντας την ικανότητα της γενίκευσης του μοντέλου.

Η εκπαίδευση των δέντρων απόφασης σε ένα Random Forest περιλαμβάνει την αναδρομική κατάτμηση των δεδομένων με βάση τα επιλεγμένα χαρακτηριστικά και τους κανόνες απόφασης. Κάθε δέντρο κατασκευάζεται ανεξάρτητα, με στόχο τη βελτιστοποίηση ενός συγκεκριμένου κριτηρίου, όπως το Gini impurity για ταξινόμηση ή το μέσο τετραγωνικό σφάλμα (Mean Squared Error – MSE) για παλινδρόμηση.

Όταν πρόκειται για προβλέψεις, κάθε δέντρο απόφασης στο Random Forest παράγει ανεξάρτητα τη δική του πρόβλεψη. Για τις εργασίες ταξινόμησης, η τελική πρόβλεψη καθορίζεται με πλειοψηφία, όπου επιλέγεται η τάξη (class) με τις περισσότερες ψήφους από όλα τα δέντρα. Στις εργασίες παλινδρόμησης, οι προβλέψεις από όλα τα δέντρα υπολογίζονται κατά μέσο όρο για να ληφθεί η τελική πρόβλεψη.

Τα τυχαία δάση προσφέρουν πολλά πλεονεκτήματα. Πρώτον, είναι ανθεκτικά σε θόρυβο και ακραίες τιμές στα δεδομένα. Το σύνολο των δέντρων αποφάσεων βοηθά στον υπολογισμό του μέσου όρου των μεμονωμένων σφαλμάτων (individual errors), με αποτέλεσμα να προκύπτει ένα πιο σταθερό και αξιόπιστο μοντέλο. Δεύτερον, μπορούν να χειριστούν χώρους υψηλών διαστάσεων (high dimensional spaces) και να καταγράψουν πολύπλοκες αλληλεπιδράσεις μεταξύ των χαρακτηριστικών *[13]*.

Ένα από τα οφέλη των Random Forests είναι η ικανότητά τους να εκτιμούν τη σημασία των χαρακτηριστικών. Με την ανάλυση της απόδοσης διαφορετικών χαρακτηριστικών σε όλο το σύνολο, η σχετική σημασία κάθε χαρακτηριστικού μπορεί να ποσοτικοποιηθεί (quantified). Αυτές οι πληροφορίες είναι πολύτιμες για την επιλογή χαρακτηριστικών και την κατανόηση των υποκείμενων δεδομένων.

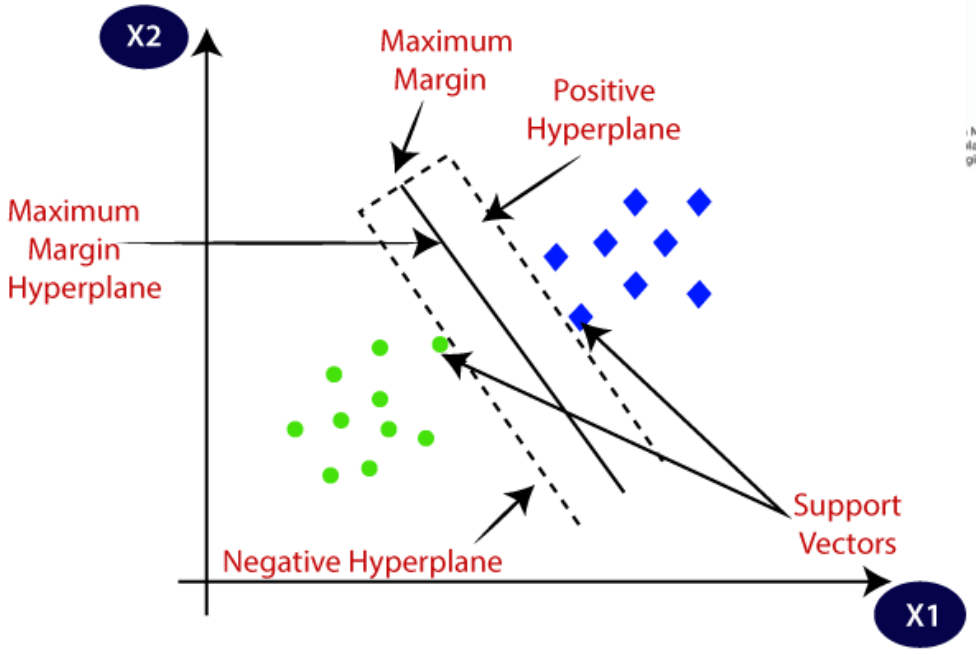
Τα Random Forests παρέχουν επίσης έναν εσωτερικό μηχανισμό αξιολόγησης που ονομάζεται Out Of Bag evaluation - OOB. Δεδομένου ότι κάθε δέντρο απόφασης εκπαιδεύεται σε ένα υποσύνολο δεδομένων με bootstrap, υπάρχουν περιπτώσεις (instances) που παραλείπονται κατά τη διάρκεια της εκπαίδευσης. Αυτές οι περιπτώσεις OOB μπορούν να χρησιμοποιηθούν για την εκτίμηση της απόδοσης του μοντέλου χωρίς να απαιτείται ξεχωριστό σύνολο επικύρωσης. Τα Random Forests είναι υπολογιστικά αποδοτικά και μπορούν να χειριστούν μεγάλα σύνολα δεδομένων. Τα μεμονωμένα δέντρα αποφάσεων μπορούν να εκπαιδευτούν και να αξιολογηθούν ανεξάρτητα, επιτρέποντας παράλληλους υπολογισμούς σε επεξεργαστές πολλαπλών πυρήνων ή κατανεμημένα συστήματα.

2.5 Support Vector Machines (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) είναι ένας ισχυρός αλγόριθμος Μηχανικής Μάθησης που χρησιμοποιείται συνήθως για εργασίες ταξινόμησης και παλινδρόμησης. Λειτουργούν βρίσκοντας ένα βέλτιστο υπερεπίπεδο (hyperplane) που διαχωρίζει στο μέγιστο διαφορετικές κλάσεις ή προβλέπει συνεχείς τιμές. Τα SVM είναι αποτελεσματικά στο χειρισμό γραμμικά διαχωρίσιμων καθώς και μη γραμμικά διαχωρίσιμων δεδομένων, καθιστώντας τα ευέλικτα και ευρέως εφαρμόσιμα σε διάφορους τομείς *[10]*.

Ο πρωταρχικός στόχος των SVM είναι να βρουν ένα υπερεπίπεδο με το μεγαλύτερο περιθώριο (margin), το οποίο είναι η απόσταση μεταξύ του υπερεπίπεδου και των πλησιέστερων σημείων δεδομένων κάθε κατηγορίας. Αυτό το περιθώριο διασφαλίζει έναν σαφή διαχωρισμό μεταξύ των κλάσεων και βοηθά στην καλή γενίκευση σε μη ορατά δεδομένα. Τα σημεία δεδομένων (data points) που βρίσκονται πιο κοντά στο υπερεπίπεδο, γνωστά ως διανύσματα υποστήριξης (support vectors), παίζουν κρίσιμο ρόλο στον καθορισμό του ορίου απόφασης.

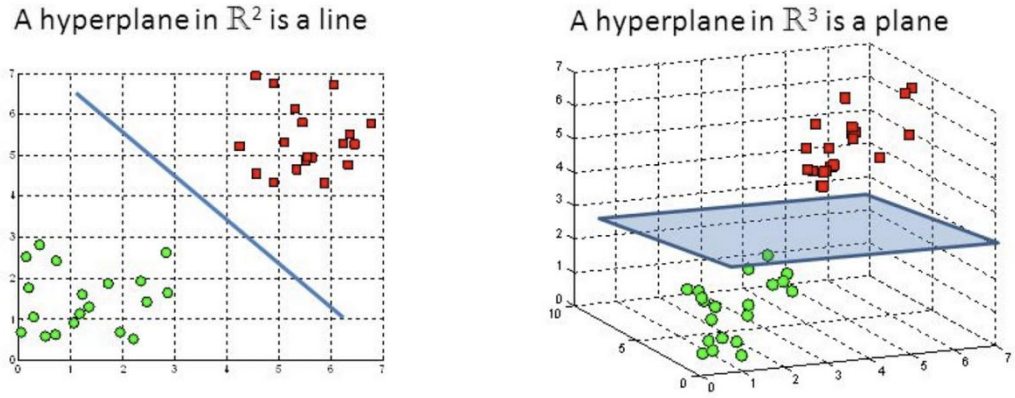
Τα SVM μπορούν να χειριστούν μη γραμμικά διαχωρίσιμα δεδομένα χρησιμοποιώντας συναρτήσεις πυρήνα (kernel functions). Μια συνάρτηση πυρήνα μετασχηματίζει τα δεδομένα εισόδου σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης, όπου γίνεται δυνητικά διαχωρίσιμο. Εφαρμόζοντας την τεχνική του πυρήνα (kernel trick), τα SVM είναι σε θέση να καταγράφουν περίπλοκες σχέσεις μεταξύ των data points και να κάνουν ακριβείς προβλέψεις.



Εικόνα 7 *Κατάταξη δεδομένων με SVM*

Η εκπαίδευση των SVM περιλαμβάνει την επίλυση ενός προβλήματος βελτιστοποίησης που στοχεύει στην ελαχιστοποίηση του σφάλματος ταξινόμησης μεγιστοποιώντας παράλληλα το περιθώριο. Αυτό το πρόβλημα βελτιστοποίησης συνήθως επιλύεται χρησιμοποιώντας τεχνικές τετραγωνικού προγραμματισμού. Η λύση στο πρόβλημα βελτιστοποίησης παρέχει τις βέλτιστες τιμές για τις παραμέτρους του υπερεπίπεδου, δηλαδή τα βάρη και τα biases, που ορίζουν το όριο απόφασης (decision boundary).

Για τον χειρισμό μη γραμμικά διαχωρίσιμων δεδομένων, τα SVM χρησιμοποιούν συναρτήσεις πυρήνα (kernel functions) όπως ο γραμμικός πυρήνας, ο πολυωνυμικός πυρήνας ή ο πυρήνας της συνάρτησης ακτινικής βάσης (RBF). Αυτές οι συναρτήσεις του πυρήνα αντιστοιχίζουν έμμεσα τα δεδομένα σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης, όπου γίνονται γραμμικά διαχωρίσιμα. Η επιλογή της συνάρτησης πυρήνα εξαρτάται από τα χαρακτηριστικά των δεδομένων και το πρόβλημα.



Εικόνα 8 *Υπερεπίπεδα στον δισδιάστατο και τρισδιάστατο χώρο*

Στην περίπτωση της δυαδικής ταξινόμησης (binary classification), τα SVM ταξινομούν νέα data points καθορίζοντας σε ποια πλευρά του ορίου απόφασης εμπίπτουν. Εάν ένα data point βρίσκεται στη μία πλευρά του υπερεπίπεδου, εκχωρείται στη μία κλάση και εάν βρίσκεται στην άλλη πλευρά, εκχωρείται στην άλλη κλάση. Τα SVM μπορούν επίσης να επεκταθούν για να χειριστούν προβλήματα ταξινόμησης πολλών κατηγοριών χρησιμοποιώντας τεχνικές όπως one vs one ή one vs rest.

Τα SVM έχουν μια παράμετρο τακτοποίησης (regularization), που συχνά υποδηλώνεται ως C, που ελέγχει την αντιστάθμιση (trade-off) μεταξύ της μεγιστοποίησης του περιθωρίου και της ελαχιστοποίησης του σφάλματος ταξινόμησης. Μια μικρότερη τιμή του C επιτρέπει μεγαλύτερο περιθώριο, αλλά μπορεί να οδηγήσει σε περισσότερες εσφαλμένες ταξινομήσεις. Αντίθετα, μια μεγαλύτερη τιμή του C δίνει έμφαση στην ελαχιστοποίηση του σφάλματος ταξινόμησης, με αποτέλεσμα ενδεχομένως ένα μικρότερο περιθώριο *[13]*.

Τα SVM προσφέρουν πολλά πλεονεκτήματα. Είναι αποτελεσματικά στο χειρισμό δεδομένων υψηλών διαστάσεων, καθώς βασίζονται μόνο σε ένα υποσύνολο των support vectors για την κατασκευή του ορίου απόφασης. Αυτή η ιδιότητα καθιστά τα SVM λιγότερο ευαίσθητα στην παρουσία μη σχετικών χαρακτηριστικών. Επιπλέον, τα SVM είναι ανθεκτικά σε ακραίες τιμές (outliers), καθώς η επιρροή τους στο όριο απόφασης περιορίζεται στα support vectors.

Ένας περιορισμός των SVM είναι η υπολογιστική τους πολυπλοκότητα, ειδικά όταν έχουμε να κάνουμε με μεγάλα σύνολα δεδομένων. Ο χρόνος εκπαίδευσης μπορεί να είναι σημαντικός, ιδιαίτερα για μη γραμμικά προβλήματα με πολύπλοκες συναρτήσεις πυρήνα. Υπάρχουν επίσης ορισμένες προκλήσεις στην επιλογή κατάλληλων συναρτήσεων πυρήνα και συντονισμού παραμέτρων για την επίτευξη βέλτιστης απόδοσης.

Τα SVM έχουν βρει εφαρμογές σε διάφορους τομείς, συμπεριλαμβανομένης της ταξινόμησης κειμένου, της ταξινόμησης εικόνων, της βιοπληροφορικής και των οικονομικών. Είναι ιδιαίτερα κατάλληλα όταν υπάρχει ανάγκη για ένα σαφές όριο απόφασης (decision boundary) και χειρισμό μη γραμμικών σχέσεων. Τα SVM έχουν μελετηθεί εκτενώς και έχουν μια σταθερή θεωρητική βάση, καθιστώντας τα μια αξιόπιστη επιλογή για πολλές εργασίες μηχανικής μάθησης.

ΚΕΦΑΛΑΙΟ 3

ΠΛΑΙΣΙΟ ΔΕΔΟΜΕΝΩΝ (DATASET)

3.1 Περιγραφή dataset

Το dataset "All Lending Club loan data" από τον ιστότοπο kaggle.com, που περιέχεται στο αρχείο **`accepted_2007_to_2018Q4.csv.gz`**, είναι ένα ολοκληρωμένο πλαίσιο δεδομένων που παρέχει πληροφορίες σχετικά με δάνεια που εκδίδονται από το Lending Club, μια πλατφόρμα δανεισμού peer-to-peer. Η χρονική κάλυψη εκτείνεται από το 2007 έως το τέταρτο τρίμηνο του 2018. Όσον αφορά τη γεωγραφική κάλυψη, το dataset περιλαμβάνει δάνεια από διάφορες πολιτείες στις Ηνωμένες Πολιτείες *[15]*.

Η επεξεργασία του dataset έγινε από το Google Colaboratory, γνωστό και ως Colab, ένα web IDE κατάλληλο για εκτέλεση κώδικα Python για εφαρμογές σχετικές με την Επιστήμη Δεδομένων (Data Science) και τη Μηχανική Μάθηση (Machine Learning) *[22]*.

Το πλαίσιο δεδομένων αποτελείται από 2260701 γραμμές και 151 στήλες, είναι δηλαδή ένα αρκετά μεγάλο dataset αφού αφορά πάνω από δύο εκατομμύρια δανειολήπτες. Σε ορισμένα πεδία ενδέχεται να λείπουν τιμές, οι οποίες θα μπορούσαν να επηρεάσουν ορισμένες αναλύσεις. Έτσι απαιτείται καθαρισμός των δεδομένων (data cleaning) πριν από τη διεξαγωγή συγκεκριμένων συμπερασμάτων.

Μέσα από την επεξεργασία που θα προκύψει είναι η διερεύνηση της κατάστασης του δανείου, των αθετήσεων πληρωμών και των χρεώσεων με την πάροδο του χρόνου και η αξιολόγηση του κινδύνου, δηλαδή η διερεύνηση της σχέσης μεταξύ των χαρακτηριστικών του δανείου (π.χ. βαθμολογία, επιτόκιο) και των επιτοκίων αθέτησης. Επίσης, θα μπορούσε να γίνει ανάλυση της συμπεριφοράς των δανειοληπτών, η εξέταση προτύπων και τάσεων στους σκοπούς των δανείων, η διάρκεια απασχόλησης και τα επίπεδα εισοδήματος. Τέλος θα μπορούσαμε να κάνουμε μια γεωγραφική ανάλυση, δηλαδή να αναλύσουμε την κατανομή των δανείων μεταξύ διαφορετικών πολιτειών.

Ακολουθεί μια λεπτομερής περιγραφή του συνόλου δεδομένων:

- **Επισκόπηση dataset:**

  Όνομα αρχείου: **`accepted_2007_to_2018Q4.csv.gz`**

  Μορφή: Συμπιεσμένο αρχείο CSV (gzip)

  Μέγεθος: 1.55 Gb

- **Πεδία (fields):**

Το σύνολο δεδομένων περιέχει πολλά πεδία (στήλες) που παρέχουν πληροφορίες για κάθε δάνειο. Μερικές από τις πιο σημαντικές στήλες είναι:

id	Ένα μοναδικό αναγνωριστικό που εκχωρείται σε κάθε δάνειο
loan_amnt	Το χρηματικό ποσό που αιτείται ο δανειολήπτης
term	Η διάρκεια αποπληρωμής του δανείου σε μήνες που μπορεί να είναι 36 ή 60
int_rate	Το επιτόκιο που αποδίδεται στο δάνειο.
grade	Το Lending Club όρισε μια βαθμολογία για το δάνειο με βάση την πιστοληπτική ικανότητα του δανειολήπτη. Πιθανές τιμές: A, B, C, D, E, F, G
sub_grade	Πιο αναλυτική ταξινόμηση εντός της βαθμίδας του δανείου. Οι υποβαθμίσεις κυμαίνονται από 1 έως 5 σε κάθε βαθμό. Για παράδειγμα, εντός του βαθμού A, οι υποβαθμίσεις μπορεί να είναι A1, A2, A3 κ.λπ.
loan_status	Η τρέχουσα κατάσταση του δανείου (π.χ. πλήρως εξοφλημένο, εξοφλημένο κ.λπ.). Οι πιθανές τιμές είναι: Fully Paid, Charged Off, Current, Late (16-30 days), Late (31-120 days)
purpose	Ο σκοπός για τον οποίο λήφθηκε το δάνειο (π.χ. ενοποίηση χρέους, αναχρηματοδότηση πιστωτικής κάρτας , βελτίωση κατοικίας, μεγάλη αγορά, μικρές επιχειρήσεις, ιατρικά έξοδα, χρηματοδότηση αυτοκινήτου, διακοπές κ.λπ.).
application type	Υποδεικνύει εάν η αίτηση δανείου ήταν ατομική ή κοινή αίτηση. Πιθανές τιμές: individual, joint
emp_length	Η διάρκεια της απασχόλησης του δανειολήπτη σε χρόνια. Πιθανές τιμές: < 1 year, 1 year, 2 years, 3 years, 4 years, 5 years, 6 years, 7 years, 8 years, 9 years, 10+ years, or "N/A" εάν δεν είναι διαθέσιμο.
issue_date	Η ημερομηνία έκδοσης του δανείου. Ημερομηνίες με τη μορφή "month-YY (Dec-15)”

verification_status	Υποδεικνύει εάν το εισόδημα του δανειολήπτη επαληθεύτηκε. Πιθανές τιμές: Verified, Source Verified, Not Verified
home_ownership	Ο τύπος ιδιοκτησίας κατοικίας που αναφέρεται από τον δανειολήπτη. Οι τιμές που μπορεί να πάρει είναι RENT, OWN, MORTGAGE, κ.α
annual_inc	Το ετήσιο εισόδημα του δανειολήπτη που δηλώνει ο ίδιος.
dti (debt-to-income ratio)	Ένας λόγος που υπολογίζεται υπολογίζοντας τις συνολικές μηνιαίες πληρωμές χρέους του δανειολήπτη προς τις συνολικές υποχρεώσεις χρέους, εξαιρουμένων των στεγαστικών δανείων, διαιρούμενο με το μηνιαίο εισόδημα του δανειολήπτη
addr_state	Το πολιτεία στην οποία διαμένει ο δανειολήπτης. Οι πιθανές τιμές είναι συντμήσεις δύο γραμμάτων, όπως NY για τη Νέα Υόρκη, CA για Καλιφόρνια, κ.α.
mths_since_recent_inq	Μήνες από την πιο πρόσφατη έρευνα
revol_util	Το ποσό της πίστωσης που χρησιμοποιεί ο δανειολήπτης σε σχέση με όλη τη διαθέσιμη ανακυκλούμενη πίστωση
bc_open_to_buy	Σύνολο ανοιχτό για αγορά σε ανακυκλούμενες τραπεζικές κάρτες
bc_util	Αναλογία συνολικού τρέχοντος υπολοίπου προς υψηλό πιστωτικό όριο για όλους τους λογαριασμούς τραπεζικών καρτών
num_op_rev_tl	Αριθμός ανοιχτών ανακυκλούμενων λογαριασμών

3.2 Προεπεξεργασία (Preprocessing)

Από το έτος 2007 έως το τέταρτο τρίμηνο του 2018, έχουν καταγραφεί στοιχεία για περισσότερους από 2 εκατομμύρια δανειολήπτες σε αυτό το dataset. Ο στόχος είναι να εντοπιστούν πρότυπα που υποδεικνύουν εάν ένα άτομο είναι πιθανό να χρεοκοπήσει, τα οποία μπορούν να χρησιμοποιηθούν για τη λήψη μέτρων όπως η άρνηση του δανείου, η μείωση του ποσού του δανείου και ο δανεισμός σε επικίνδυνους αιτούντες με υψηλότερο επιτόκιο.

Αρχικά φορτώνουμε τα κατάλληλα πακέτα και βιβλιοθήκες στο Colab (**`sklearn`**, **`pandas`**, **`numpy`**, **`matplotlib`**, **`seaborn`**, κ.α) *[4]* και διαβάζουμε το αρχείο που περιέχει τα δεδομένα μας (dataset). Μπορούμε να παρατηρήσουμε ότι το dataset περιέχει πάνω από 2 εκατομμύρια γραμμές που αφορούν τα δάνεια και 150 στήλες. Αυτός είναι ένας αρκετά σημαντικός όγκος δεδομένων και συνήθως υπάρχει πολύς «θόρυβος» (noise) που ενδέχεται να δημιουργήσει πολλά προβλήματα κατά την επεξεργασία επομένως πρέπει με κάποιο τρόπο να βελτιώσουμε την ποιότητα και την καταλληλότητα των δεδομένων.

```
[4]  from google.colab import drive
     drive.mount('/content/drive')

     Mounted at /content/drive

[5]  path='/content/drive/MyDrive/accepted_2007_to_2018Q4.csv'

[6]  loan = pd.read_csv(path, low_memory=False)

[7]  loan.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 2260701 entries, 0 to 2260700
     Columns: 151 entries, id to settlement_term
     dtypes: float64(113), object(38)
     memory usage: 2.5+ GB
```

Εικόνα 9 *Πληροφορίες για το dataset*

Στη συνέχεια προχωράμε στη διαδικασία να μειώσουμε τον αριθμό των μεταβλητών (στήλες) επιλέγοντας τις πιο σημαντικές ιδιότητες που αντιστοιχούν στις στήλες. Παρατηρούμε ότι τώρα το dataset έχει γίνει πιο ξεκάθαρο και φαίνονται καλύτερα οι μεμονωμένες τιμές των μεταβλητών που κρατήσαμε. Μπορούμε επίσης να χρησιμοποιήσουμε τη συνάρτηση describe για να δούμε πιο αναλυτικά τα στατιστικά μεγέθη. Μπορούμε να δούμε το πλήθος των τιμών κάθε χαρακτηριστικού (count), τον μέσο όρο κάθε μεταβλητής (mean), τη μέγιστη και την ελάχιστη τιμή (min-max), την τυπική απόκλιση (std) ως μέτρο διασποράς, καθώς και το τα τεταρτημόρια Q1 το οποίο αντιστοιχεί στο 25% των μετρήσεων αν αυτές διαταχθούν σε αύξουσα σειρά, Q2 το οποίο ταυτίζεται με τη διάμεσο και αντιστοιχεί στο 50% των μετρήσεων και Q3 το οποίο αντιστοιχεί στο 75% των μετρήσεων.

```
loans = loan[['loan_amnt', 'term','int_rate', 'sub_grade','emp_title',
              'emp_length','home_ownership', 'annual_inc', 'loan_status', 'addr_state',
              'dti','mths_since_recent_inq', 'revol_util', 'bc_open_to_buy', 'bc_util', 'num_op_rev_tl']]
```

```
loans.head()
```

	loan_amnt	term	int_rate	sub_grade	emp_title	emp_length	home_ownership	annual_inc	loan_status
0	3600.0	36 months	13.99	C4	leadman	10+ years	MORTGAGE	55000.0	Fully Paid
1	24700.0	36 months	11.99	C1	Engineer	10+ years	MORTGAGE	65000.0	Fully Paid
2	20000.0	60 months	10.78	B4	truck driver	10+ years	MORTGAGE	63000.0	Fully Paid
3	35000.0	60 months	14.85	C5	Information Systems Officer	10+ years	MORTGAGE	110000.0	Current
4	10400.0	60 months	22.45	F1	Contract Specialist	3 years	MORTGAGE	104433.0	Fully Paid

```
pd.set_option('display.float_format', lambda x: '%.0f' % x)
```

```
loans.describe()
```

	loan_amnt	int_rate	annual_inc	dti	mths_since_recent_inq	revol_util	bc_open_to_buy	bc_util	num_op_rev_tl
count	2260668	2260668	2260664	2258957	1965233	2258866	2185733	2184597	2190392
mean	15047	13	77992	19	7	50	11394	58	8
std	9190	5	112696	14	6	25	16600	29	5
min	500	5	0	-1	0	0	0	0	0
25%	8000	9	46000	12	2	32	1722	35	5
50%	12900	13	65000	18	5	50	5442	60	7
75%	20000	16	93000	24	11	69	14187	83	10
max	40000	31	110000000	999	25	892	711140	340	91

Εικόνα 10 *Το dataset μετά την μείωση των μεταβλητών*

Ακολούθως θα ασχοληθούμε με τις τιμές που λείπουν (missing data).Οι τιμές αυτές προκύπτουν όταν δεν αποθηκεύεται καμία τιμή δεδομένων σε κάποιο πεδίο. Είναι σημαντικό να εντοπίσουμε τα missing data γιατί ειδάλλως τα πιθανά αποτελέσματα θα υστερούν σε ποιότητα και θα μοιάζουν αφύσικα και πιθανόν να προκύψουν και άλλα προβλήματα καθώς προχωράει η επεξεργασία. Επιπροσθέτως, θα πρέπει να λάβουμε υπόψη ότι κάποια πακέτα όπως το scikit-learn, θεωρούν ότι όλες οι τιμές είναι αριθμητικές, οπότε θα πρέπει να είμαστε πολύ προσεκτικοί αν έχουμε διαφορετικού τύπου δεδομένα.

Όπως μπορούμε να δούμε στην *Εικόνα 11*, οι περισσότερες από τις μεταβλητές δεν περιέχουν πολλές τιμές που λείπουν, οι περισσότερες από αυτές είναι πολύ κάτω από 8%. Δεν θα χρειαστεί να αφαιρέσουμε καμία μεταβλητή, αλλά πρέπει να χειριστούμε αυτές τις τιμές που λείπουν, πριν προσαρμόσουμε το μοντέλο. Σε πρώτη φάση θα διαγράψουμε τις σειρές με τιμές που λείπουν. Αυτή η ενέργεια θα ισχυροποιήσει το μοντέλο μας παρόλο που θα χαθούν ορισμένα δεδομένα, όμως το dataset είναι αρκετά μεγάλο οπότε αυτές οι αλλαγές δεν θα το επηρεάσουν αισθητά.

```
missing_data = pd.DataFrame({'total_missing': loans.isnull().sum(), '%_missing':
(loans.isnull().sum()/2260701)*100})
missing_data
```

	total_missing	%_missing
loan_amnt	33	0
term	33	0
int_rate	33	0
sub_grade	33	0
emp_title	167002	7
emp_length	146940	6
home_ownership	33	0
annual_inc	37	0
loan_status	33	0
addr_state	33	0
dti	1744	0
mths_since_recent_inq	295468	13
revol_util	1835	0
bc_open_to_buy	74968	3
bc_util	76104	3
num_op_rev_tl	70309	3

Εικόνα 11 *Ποσοστό missing data για κάθε μεταβλητή*

```
loans = loans.dropna()
loans.info()
```

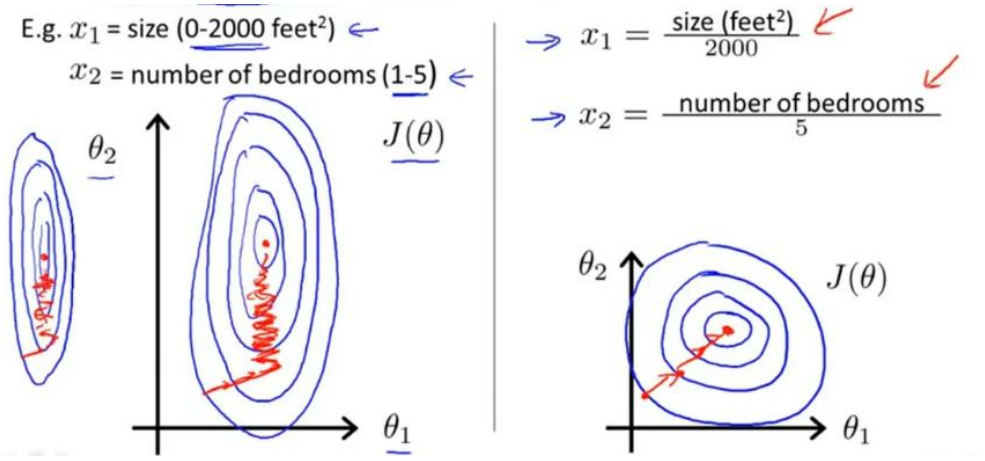
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1787734 entries, 0 to 2260698
Data columns (total 16 columns):
 #   Column                 Dtype
---  ------                 -----
 0   loan_amnt              float64
 1   term                   object
 2   int_rate               float64
 3   sub_grade              object
 4   emp_title              object
 5   emp_length             object
 6   home_ownership         object
 7   annual_inc             float64
 8   loan_status            object
 9   addr_state             object
 10  dti                    float64
 11  mths_since_recent_inq  float64
 12  revol_util             float64
 13  bc_open_to_buy         float64
 14  bc_util                float64
 15  num_op_rev_tl          float64
dtypes: float64(9), object(7)
memory usage: 231.9+ MB
```

Εικόνα 12 *Το dataset μετά την αφαίρεση των missing values*

3.3 Normalization - Scaling

Τις περισσότερες φορές, το σύνολο δεδομένων (dataset) θα περιέχει χαρακτηριστικά που ποικίλλουν πολύ σε μεγέθη, μονάδες και εύρος. Αλλά επειδή, οι περισσότεροι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν Ευκλείδεια απόσταση μεταξύ δύο σημείων (data points) στους υπολογισμούς τους, αυτό είναι ένα πρόβλημα *[14]*.

Αν δεν πάρουμε τις κατάλληλες πρωτοβουλίες για κλιμάκωση, οι αλγόριθμοι αυτοί θα λάβουν μονάχα το μέγεθος των χαρακτηριστικών και θα παραμελήσουν τις μονάδες. Έτσι, τα αποτελέσματα θα διαφέρουν πολύ μεταξύ των διαφορετικών μονάδων για παράδειγμα 5kg και 5000 gr Τα χαρακτηριστικά με υψηλά μεγέθη θα μετράνε πολύ περισσότερο στους υπολογισμούς της απόστασης από τα χαρακτηριστικά με πιο χαμηλά μεγέθη. Για να καταστείλουμε αυτό το φαινόμενο, πρέπει να φέρουμε όλα τα χαρακτηριστικά στο ίδιο επίπεδο μεγεθών. Αυτό μπορεί να επιτευχθεί με κλιμάκωση (scaling).



Εικόνα 13 *Κλιμάκωση (scaling) δεδομένων*

Η κανονικοποίηση (normalization) έχει βασικό σκοπό την αλλαγή των παρατηρήσεων ώστε να μπορούν να περιγραφούν ως κανονική κατανομή (normal distribution). Η κανονική κατανομή (κατανομή Gauss), γνωστή και ως καμπύλη καμπάνας (bell curve), είναι μια ειδική στατιστική κατανομή όπου περίπου ίσος αριθμός παρατηρήσεων βρίσκονται πάνω και κάτω από το μέσο όρο, η μέση και η διάμεση τιμή ταυτίζονται και υπάρχουν περισσότερες παρατηρήσεις πιο κοντά στον μέσο όρο *(Εικόνα 14)*.

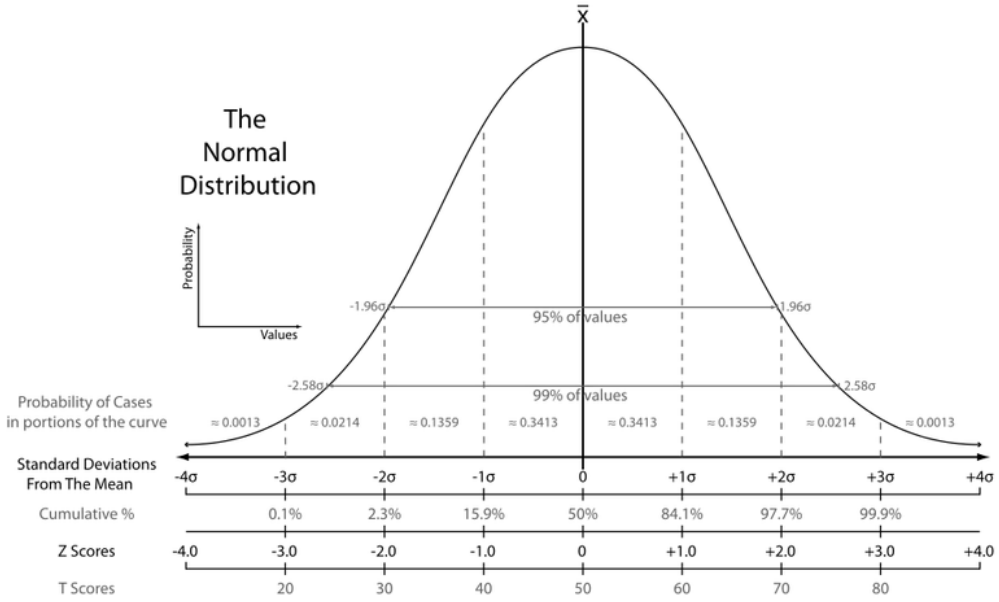
Υπάρχουν διάφοροι τρόποι για να επιτευχθεί η κανονικοποίηση (normalization):

1. Υπολογίζω την μέγιστη (max) και την ελάχιστη (min) τιμή του δείγματος και μετασχηματίζω τα δεδομένα ώστε να βρίσκονται σε ένα συγκεκριμένο εύρος π.χ.[0,1]

$$\boldsymbol{y = \frac{x - \min(x)}{\max(x) - \min(x)}}$$

όπου $\boldsymbol{y}$ είναι η κανονικοποιημένη μεταβλητή. Εύκολα μπορούμε να παρατηρήσουμε ότι αν $\boldsymbol{x} = \boldsymbol{min}(\boldsymbol{x})$ τότε το $\boldsymbol{y} = \mathbf{0}$, ενώ όταν $\boldsymbol{x} = \boldsymbol{max}(\boldsymbol{x})$ το $\boldsymbol{y} = \mathbf{1}$.

Η κλιμάκωση αυτού του τύπου θεωρείται σημαντική στους αλγόριθμους όπως ο Support Vector Machines (SVM) και ο k-nearest neighbors (KNN), όπου η απόσταση μεταξύ των σημείων (data points) είναι σημαντική.



Εικόνα 14 *Κανονική κατανομή*

2. Υπολογίζω την μέση τιμή (mean) του δείγματος και μετασχηματίζω τα δεδομένα ώστε να βρίσκονται στο επιθυμητό εύρος, χρησιμοποιώντας αντίστοιχα τον μετασχηματισμό:

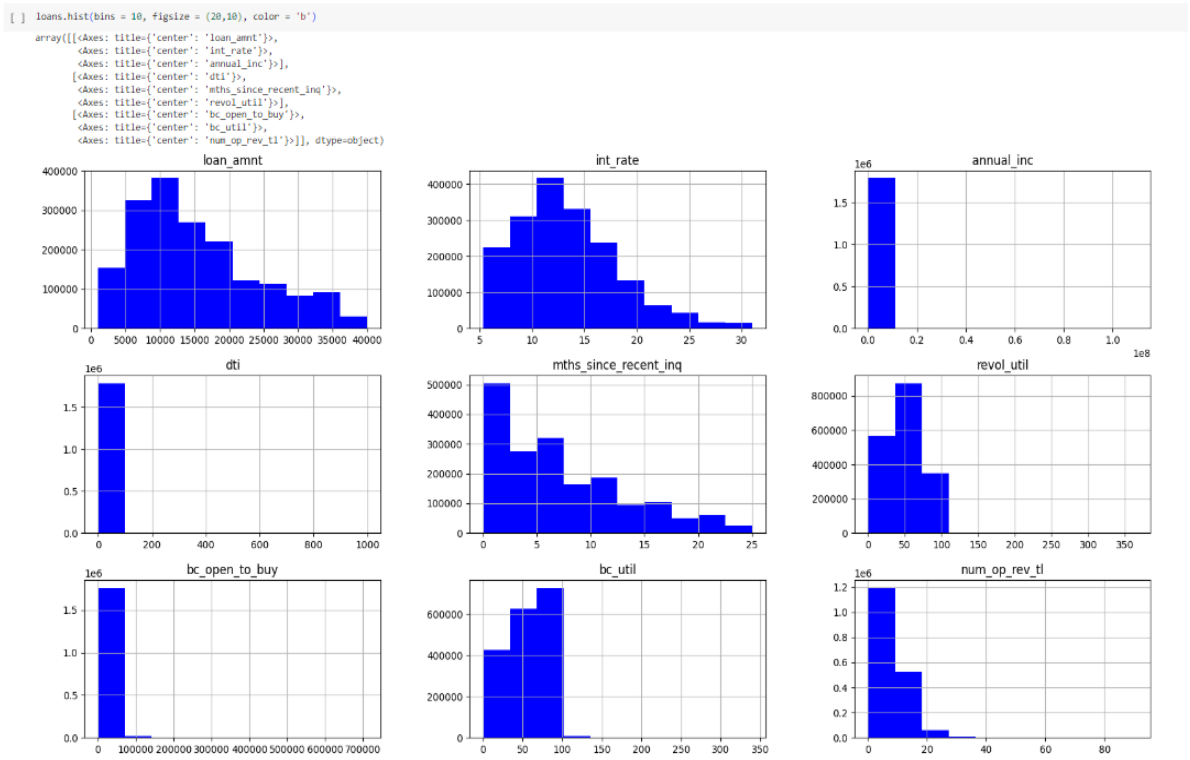
$$\boldsymbol{y} = \frac{\boldsymbol{x} - \mathbf{average}\ (\boldsymbol{x})}{\mathbf{max}(\boldsymbol{x}) - \mathbf{min}\ (\boldsymbol{x})}$$

3. Η τυποποίηση (standardization) που αναφέρεται και ως z-score normalization, μετασχηματίζει τα δεδομένα έτσι ώστε η κατανομή που προκύπτει να έχει μέσο όρο 0 και τυπική απόκλιση 1. Χρησιμοποιείται ο μετασχηματισμός:

$$\boldsymbol{y} = \frac{\boldsymbol{x} - \mathbf{average}\ (\boldsymbol{x})}{\boldsymbol{variance}(\boldsymbol{x})}$$

Συνοψίζοντας, κατά την κλιμάκωση (scaling), αλλάζει το εύρος των δεδομένων, ενώ στην κανονικοποίηση (normalization) αλλάζει κυρίως η κατανομή των δεδομένων.

Ας οπτικοποιήσουμε στη συνέχεια την κατανομή κάποιων παρατηρήσεων του δείγματος, χρησιμοποιώντας τη βιβλιοθήκη matplotlib:



Εικόνα 15 *Ιστογράμματα κατανομής μεταβλητών του δείγματος*

Παρατηρώ ότι κάποιες μεταβλητές έχουν ακραίες τιμές (outliers) και παρουσιάζουν μέτρια έως θετική κυρτότητα (skewness). Εφόσον οι ακραίες τιμές επηρεάζουν τη διακύμανση θα πρέπει να ασχοληθούμε με αυτές.

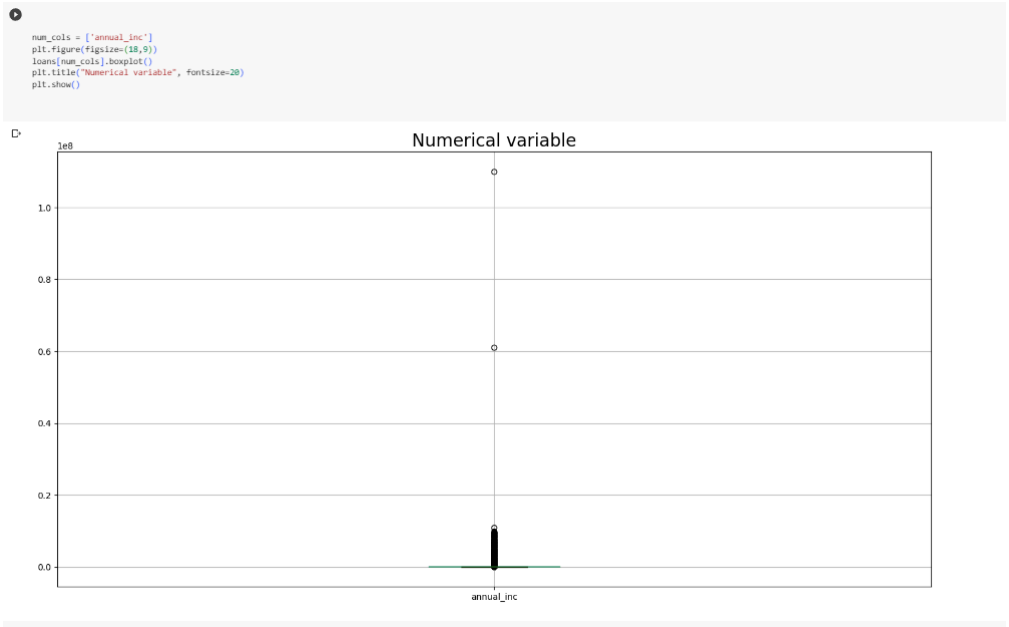
Ας θεωρήσουμε τη μεταβλητή για το ετήσιο εισόδημα (**`annual_inc`**):

```
[ ]  loans.annual_inc.describe()

count     1787734
mean        81316
std        121941
min             0
25%         50000
50%         69000
75%         96000
max     110000000
Name: annual_inc, dtype: float64
```

```
[ ]  loans.annual_inc.unique()

array([ 55000.,  65000.,  63000., ..., 136799., 131209., 180792.])
```

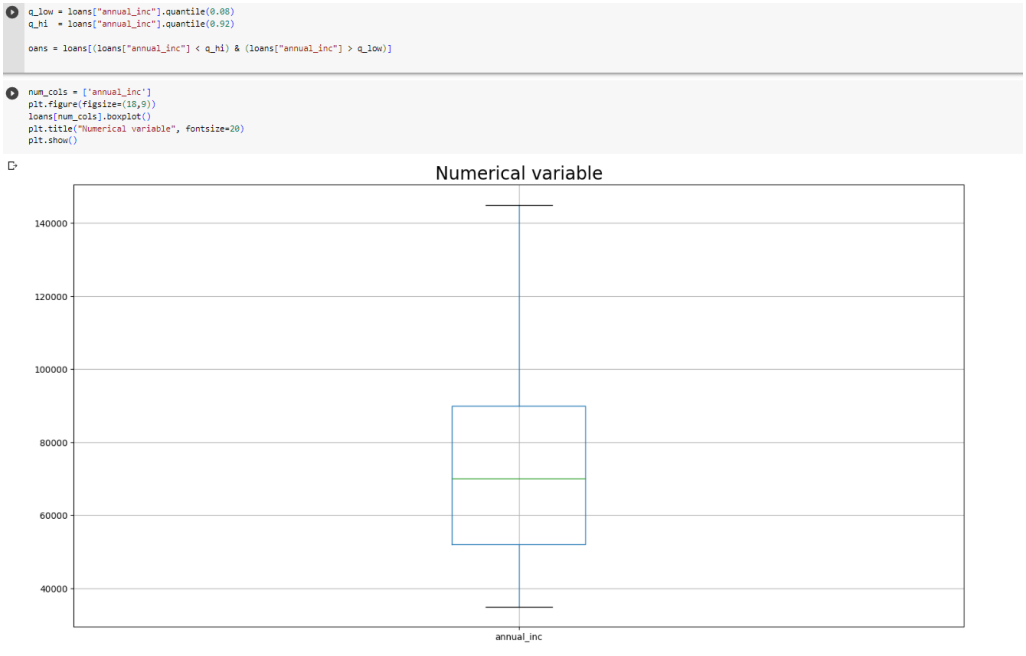
Εικόνα 16 *Στατιστικά στοιχεία της μεταβλητής annual_inc*

Από τα στατιστικά στοιχεία που πήραμε *(Εικόνα 16)*, παρατηρούμε ότι η μέγιστη τιμή για το ετήσιο εισόδημα είναι 11 εκατομμύρια δολάρια Αμερικής, που προφανώς δεν μπορεί να περιγράψει τον μέσο άνθρωπο που αιτείται για δάνειο. Με αυτήν τη λογική μπορούμε να παραλείψουμε αυτήν την τιμή. Ένας άλλος τρόπος για να δούμε ξεκάθαρα τις ακραίες τιμές είναι να σχεδιάσουμε ένα θηκόγραμμα (box plot) για τη μεταβλητή annual_inc:



Εικόνα 17 *Οι ακραίες τιμές της μεταβλητής annual_inc*

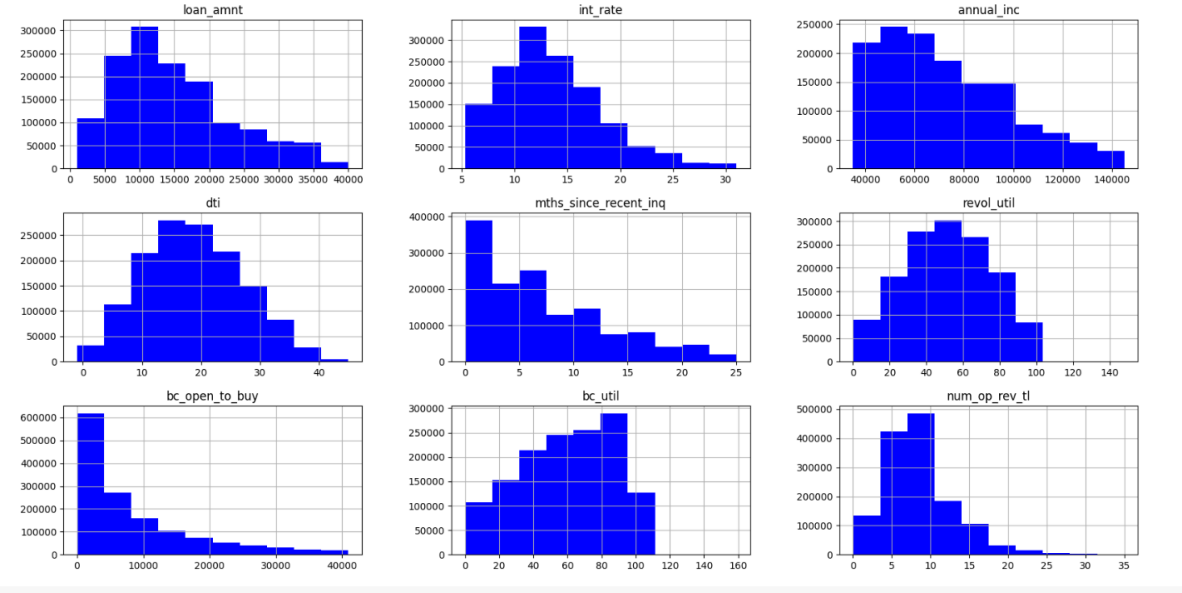
Παρατηρώ ότι υπάρχουν δύο (2) ακραίες τιμές, 11 και 6 εκατομμύρια δολάρια αντίστοιχα οι οποίες, όπως φαίνεται και στο θηκόγραμμα *(Εικόνα 17)* συμπιέζουν το εύρος των τιμών για το ετήσιο εισόδημα. Γι' αυτόν το λόγο θα τις αφαιρέσω και θα σχεδιάσω εκ νέου το θηκόγραμμα, χωρίς τις ακραίες τιμές.



Εικόνα 18 *Το box plot χωρίς τους outliers*

Το θηκόγραμμα (boxplot) τώρα φαίνεται πιο κανονικοποιημένο καθώς δεν έχει ακραίες τιμές (outliers). Την ίδια μέθοδο εφαρμόζουμε και στις άλλες μεταβλητές με σκοπό να απαλείψουμε τις τυχόν ακραίες τιμές.

Τελικά, μετά την αφαίρεση των ακραίων τιμών από τις μεταβλητές του δείγματος, η κατανομή των παρατηρήσεων διορθώθηκε, όπως παρουσιάζεται στην παρακάτω εικόνα:



Εικόνα 19 *Ιστογράμματα κατανομής μεταβλητών του δείγματος χωρίς outliers*

Ακολούθως, θα ασχοληθούμε με τα κατηγορικά δεδομένα (categorical data) του δείγματος, τα οποία ακολουθούν μια διαφορετική προσέγγιση από τα αριθμητικά δεδομένα (numerical data). Οι υπολογιστικές μηχανές δεν κατανοούν τα δεδομένα κειμένου, οπότε πρέπει να μετατρέψουμε τα κατηγορικά δεδομένα σε αριθμητικά για να επεξεργαστούν.

Ένας τρόπος είναι να χρησιμοποιήσουμε την τακτική κωδικοποίηση (ordinal encoding), δηλαδή θα αποδώσουμε αριθμούς σε τακτοποιημένα κατηγορικά δεδομένα (π.χ για το επίπεδο εκπαίδευσης θα ,μπορούσαμε να δώσουμε τις τιμές 1 για απόφοιτους Δημοτικού, 2 για απόφοιτους Γυμνασίου, 3 για απόφοιτους Λυκείου, 4 για απόφοιτους Πανεπιστημίου και 5 για κατόχους μεταπτυχιακού τίτλου).

Για να μετατρέψουμε τα δεδομένα μας από κατηγορικά σε αριθμητικά, ώστε να μπορούμε να τα επεξεργαστούμε στη συνέχεια, θα κάνουμε χρήση της συνάρτησης **`replace()`** όπως φαίνεται στην *Εικόνα 20*.

```
cleaner_app_type = {"term": {" 36 months": 1.0, " 60 months": 2.0},
                    "sub_grade": {"A1": 1.0, "A2": 2.0, "A3": 3.0, "A4": 4.0, "A5": 5.0,
                                  "B1": 11.0, "B2": 12.0, "B3": 13.0, "B4": 14.0, "B5": 15.0,
                                  "C1": 21.0, "C2": 22.0, "C3": 23.0, "C4": 24.0, "C5": 25.0,
                                  "D1": 31.0, "D2": 32.0, "D3": 33.0, "D4": 34.0, "D5": 35.0,
                                  "E1": 41.0, "E2": 42.0, "E3": 43.0, "E4": 44.0, "E5": 45.0,
                                  "F1": 51.0, "F2": 52.0, "F3": 53.0, "F4": 54.0, "F5": 55.0,
                                  "G1": 61.0, "G2": 62.0, "G3": 63.0, "G4": 64.0, "G5": 65.0,
                                   },"emp_length": {"< 1 year": 0.0, '1 year': 1.0, '2 years': 2.0, '3 years': 3.0, '4 years': 4.0,
                                   '5 years': 5.0, '6 years': 6.0, '7 years': 7.0, '8 years': 8.0, '9 years': 9.0,
                                   '10+ years': 10.0 }
                    }
loans = loans.replace(cleaner_app_type)
loans.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1391217 entries, 0 to 2260698
Data columns (total 13 columns):
 #   Column                 Non-Null Count    Dtype
---  ------                 --------------    -----
 0   loan_amnt              1391217 non-null  float64
 1   term                   1391217 non-null  float64
 2   sub_grade              1391217 non-null  float64
 3   emp_title              1391217 non-null  object
 4   emp_length             1391217 non-null  float64
 5   home_ownership         1391217 non-null  object
 6   annual_inc             1391217 non-null  float64
 7   loan_status            1391217 non-null  object
 8   addr_state             1391217 non-null  object
 9   dti                    1391217 non-null  float64
 10  mths_since_recent_inq  1391217 non-null  float64
 11  revol_util             1391217 non-null  float64
 12  num_op_rev_tl          1391217 non-null  float64
dtypes: float64(9), object(4)
memory usage: 148.6+ MB
```

Εικόνα 20 *Μετατροπή κατηγορικών δεδομένων σε αριθμητικά*

ΚΕΦΑΛΑΙΟ 4

ΑΝΑΛΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

4.1 Περιγραφική ανάλυση

Σε αυτό το σημείο θα προσπαθήσουμε να κατανοήσουμε την αιτία για την οποία οι δανειολήπτες αδυνατούν να αποπληρώσουν το δάνειό τους. Μέχρι τώρα ασχοληθήκαμε με μεταβλητές και δεδομένα εισόδου (input data) και ως επιθυμητή έξοδο θέλουμε μια απάντηση για το αν ο πελάτης θα αποπληρώσει τελικά το δάνειό του ή όχι. Η μεταβλητή **`loan_status`** περιλαμβάνει τα παρακάτω χαρακτηριστικά, από τα οποία θα μελετήσουμε τις τιμές Fully_Paid και Charhed_Off :

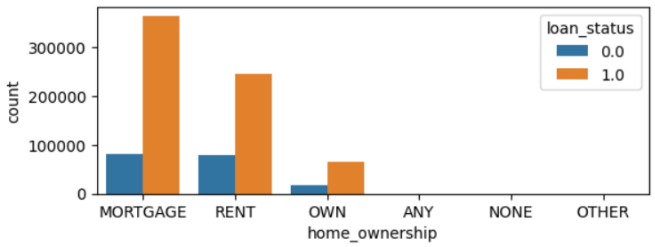
```
[ ] loans['loan_status'].value_counts()

    Fully Paid            672377
    Current               520605
    Charged Off           176077
    Late (31-120 days)     13923
    In Grace Period         5512
    Late (16-30 days)       2701
    Default                   22
    Name: loan_status, dtype: int64

[ ] array = ['Charged Off', 'Fully Paid']
    loans = loans.loc[loans['loan_status'].isin(array)]
    loans.head()
```

	loan_amnt	term	sub_grade	emp_title	emp_length	home_ownership	annual_inc	loan_status	addr_state	dti	mths_since_recent_inq	revol_util	num_op_rev_tl
0	3600	1	24	leadman	10	MORTGAGE	55000	Fully Paid	PA	6	4	30	4
2	20000	2	14	truck driver	10	MORTGAGE	63000	Fully Paid	IL	11	10	56	4
4	10400	2	51	Contract Specialist	3	MORTGAGE	104433	Fully Paid	PA	25	1	64	7
7	20000	1	11	road driver	10	MORTGAGE	85000	Fully Paid	SC	18	8	6	3
8	10000	1	2	SERVICE MANAGER	6	RENT	85000	Fully Paid	PA	13	1	34	13

Εικόνα 21 *Οι τιμές της μεταβλητής loan_status*

Παρακάτω φαίνεται ότι το 79.2% των υπόλοιπων δανείων έχει εξοφληθεί πλήρως και το 20.8% έχει χρεωθεί:

```
[ ] loans['loan_status'].value_counts()

    1    672377
    0    176077
    Name: loan_status, dtype: int64

    plt.subplot(2, 1, 1)
    sns.countplot(x='home_ownership', data=loans, hue='loan_status')

    <Axes: xlabel='home_ownership', ylabel='count'>
```



Εικόνα 22 *Ποσοστά αποπληρωμένων δανείων*

Για να δούμε στη συνέχεια αν οι αιτούντες που δεν έχουν περιουσία είναι πιο πιθανό να χρεώσουν καθώς και με ποιον τρόπο ο τύπος ιδιοκτησίας επηρεάζει το αποτέλεσμα του **`loan_status`**.

```
[ ] analyse_home_ownership = loans.groupby(['home_ownership','loan_status'])['loan_status'].count()
    analyse_home_ownership = analyse_home_ownership.groupby(level=0).apply(lambda x:100 * x / float(x.sum()))

    analyse_home_ownership
```

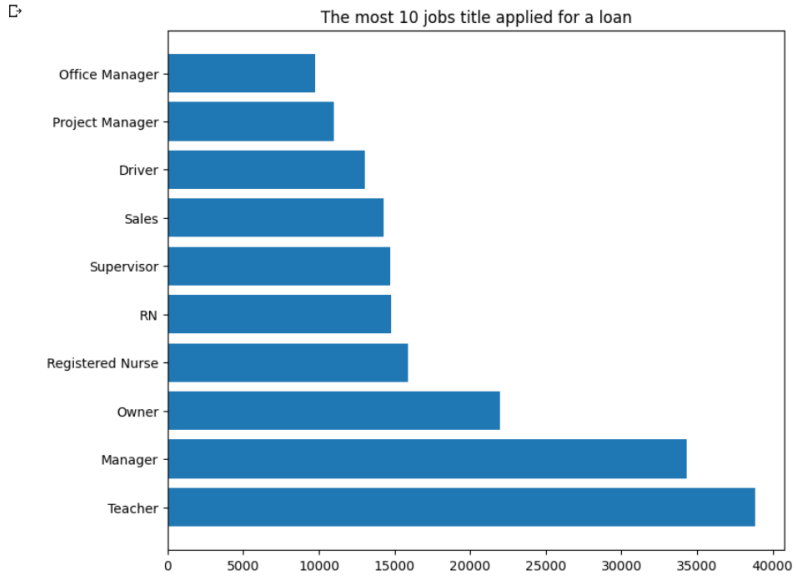
```
<ipython-input-48-50f464ff5598>:2: FutureWarning: Not prepending group keys to the result index of transform-like
To preserve the previous behavior, use

        >>> .groupby(..., group_keys=False)

To adopt the future behavior and silence this warning, use

        >>> .groupby(..., group_keys=True)
  analyse_home_ownership = analyse_home_ownership.groupby(level=0).apply(lambda x:
home_ownership  loan_status
ANY             0              22
                1              78
MORTGAGE        0              18
                1              82
NONE            0              14
                1              86
OTHER           0              22
                1              78
OWN             0              21
                1              79
RENT            0              24
                1              76
Name: loan_status, dtype: float64
```

Εικόνα 23 *Σχέση μεταξύ ιδιοκτησίας και αποπληρωμής του δανείου*

Παρατηρούμε ότι οι αιτούντες δάνειο που ενοικιάζουν αντί να κατέχουν το ακίνητο έχουν 15-30% περισσότερες πιθανότητες να χρεώσουν σε σύγκριση με εκείνους που έχουν υποθήκη ή κατέχουν ακίνητο.

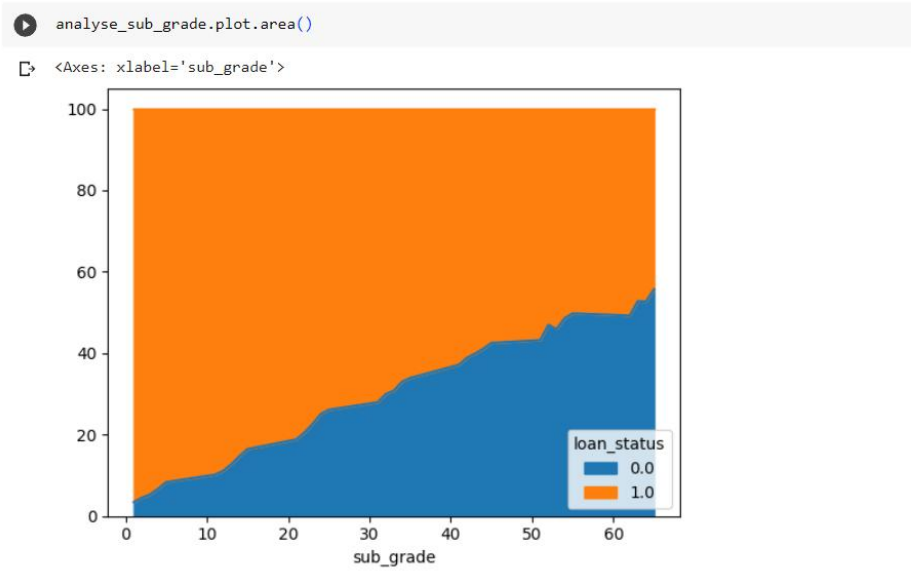
Υπάρχουν άλλοι τύποι ανάλυσης που μπορούμε να κάνουμε, που μπορούν να μας βοηθήσουν στην ερμηνεία του μοντέλου σε μεταγενέστερο στάδιο. Για παράδειγμα, θα μπορούσαμε να δούμε ποιοι τύποι εργασίας καταλήγουν πιο συχνά να μην πληρώνουν δάνειο *(Εικόνα 24)*.

```
plt.barh(loan_emp_title.emp_title.value_counts()[:10].index, loan.emp_title.value_counts()[:10])
plt.title("The most 10 jobs title applied for a loan")
plt.tight_layout()
```



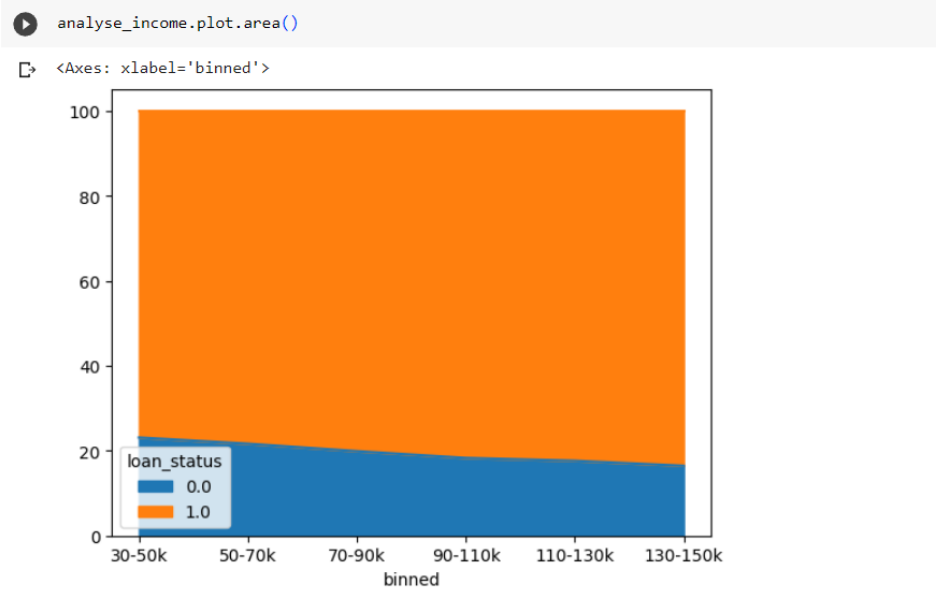
Εικόνα 24 *Σχέση μεταξύ εργασίας και αποπληρωμής του δανείου*

Στη συνέχεια ας δούμε πως συσχετίζονται μεταξύ τους διάφορα ζεύγη μεταβλητών:



Εικόνα 25 *Συσχέτιση loan_status και sub_grade*

Παρατηρούμε μια ισχυρή συσχέτιση μεταξύ loan_status και sub_grade.

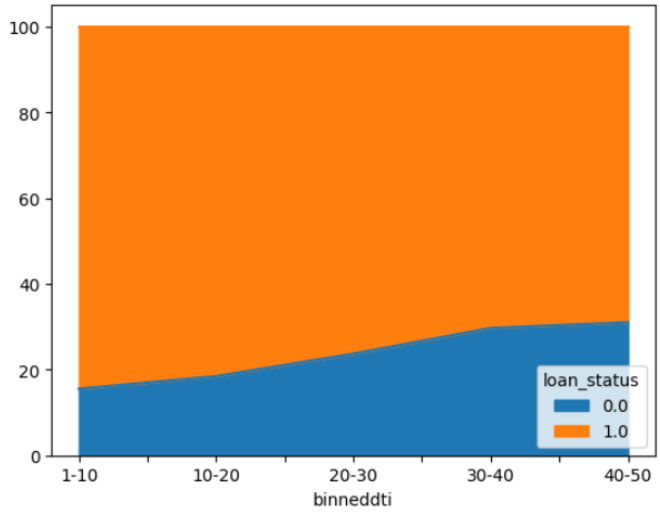


Εικόνα 26 *Συσχέτιση loan_status και annual_inc*

Η συσχέτιση μεταξύ loan_status και annual_inc *(Εικόνα 26)* είναι πολύ καλύτερη, καθώς μπορούμε να δούμε μια άμεση σχέση μεταξύ του εισοδήματος του αιτούντος και της κατάστασης του δανείου. Όσο χαμηλότερο είναι το εισόδημα τόσο πιο πιθανό είναι ο αιτών να μην το επιστρέψει, στην πραγματικότητα, είναι σχεδόν 1.5 φορές πιο πιθανό. Αυτές οι πληροφορίες θα μπορούσαν να χρησιμοποιηθούν από τους υπαλλήλους δανείων για τη δημιουργία πρόσθετων επιχειρηματικών κανόνων για τη μείωση των εταιρικών κινδύνων.

```
analyse_dti = analyse_dti.unstack()
analyse_dti.plot.area()
```

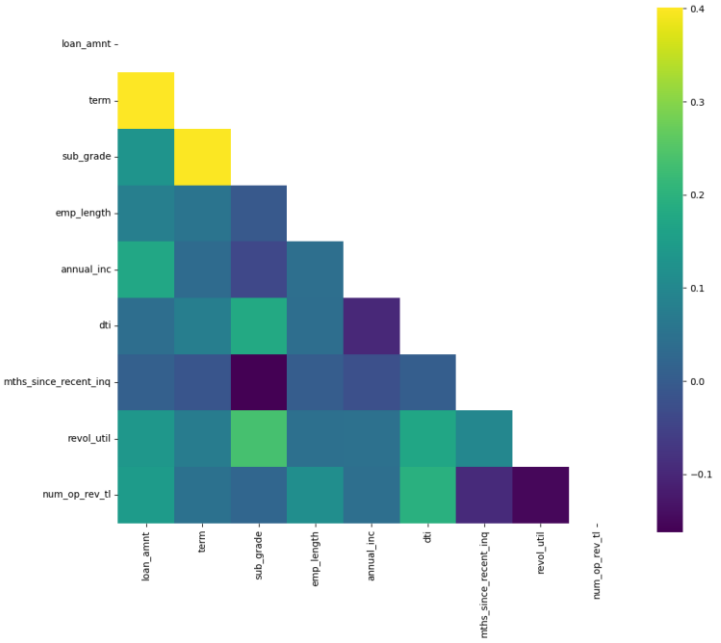
```
<Axes: xlabel='binneddti'>
```



Εικόνα 27 *Συσχέτιση loan_status και ddi*

Οι αιτούντες με έως και 20% υπάρχον χρέος (ddi) επιστρέφουν δάνεια πιο συχνά *(Εικόνα 27)*. Συγκρίνοντας τους αιτούντες με χαμηλό ddi και με υψηλό ddi, μπορούμε να δούμε ότι οι αιτούντες υψηλό ddi έχουν έως και 200% περισσότερες πιθανότητες να χρεώσουν.

Ας κατασκευάσουμε στη συνέχεια τον πίνακα συσχέτισης όλων των μεταβλητών του dataset. Ο πίνακας αυτός *(Εικόνα 28)* παρέχει μια οπτική αναπαράσταση των σχέσεων μεταξύ των διαφορετικών μεταβλητών στο σύνολο των δεδομένων με βάση τις τιμές συσχέτισής τους.



Εικόνα 28 *Συσχέτιση των μεταβλητών του dataset*

4.2 Αποτελέσματα με βάση τα μοντέλα Μηχανικής Μάθησης.

- **Νευρωνικά Δίκτυα (Neural Networks)**

Θα ξεκινήσουμε τη φάση μοντελοποίησης. Αρχικά θα εφαρμόσουμε ένα train-test-split στα δεδομένα πριν προχωρήσουμε στη φάση της εκπαίδευσης *[17] (Εικόνα 29).* Δημιουργούμε τις μεταβλητές X και y για την αποθήκευση των χαρακτηριστικών εισόδου και την αποθήκευση των τιμών της μεταβλητής στόχου (target variable). Στη συνέχεια τα σύνολα εκπαίδευσης εκχωρούνται στα **`X_train`** και **`y_train`** ενώ τα σύνολα δοκιμών εκχωρούνται στα **`X_test`** και **`y_test`** αντίστοιχα.

Ο κώδικας δημιουργεί ένα στιγμιότυπο του **`MinMaxScaler`** και το χρησιμοποιεί για να κλιμακώσει τα δεδομένα εκπαίδευσης και δοκιμής ξεχωριστά. Τα δεδομένα εκπαίδευσης **`X_train`**, προσαρμόζονται και μετασχηματίζονται χρησιμοποιώντας **`fit_transform`**, ενώ τα δεδομένα δοκιμής **`X_test`**, μετασχηματίζονται χρησιμοποιώντας **`tranform`**. Οι εντολές εκτύπωσης παρέχουν την shape information των μετασχηματισμένων δεδομένων εκπαίδευσης και δοκιμής, αντίστοιχα. Έπειτα κατασκευάζεται ένα μοντέλο νευρωνικού δικτύου με πολλαπλά πυκνά στρώματα. Κάθε πυκνό στρώμα περιέχει έναν καθορισμένο αριθμό νευρώνων και χρησιμοποιεί τη συνάρτηση ενεργοποίησης **`relu`**, εκτός από το στρώμα εξόδου που χρησιμοποιεί τη συνάρτηση ενεργοποίησης **`sigmoid.`**

```
[ ] df_accepted = df_accepted.drop(columns=['earliest_cr_line'])
```

```
df = loans.copy()

X = df.loc[:, df.columns != 'loan_paid'].values
y = df.loan_paid.values
```

```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[ ] scaler = MinMaxScaler()

    X_train= scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)

    print(X_train.shape)
    print(X_test.shape)

    (1072820, 81)
    (268206, 81)
```

```
# building the model

model = Sequential()
model.add(Dense(units=78,activation='relu'))
model.add(Dense(units=39,activation='relu'))
model.add(Dense(units=19,activation='relu'))
model.add(Dense(units=8,activation='relu'))
model.add(Dense(units=4,activation='relu'))
model.add(Dense(units=1,activation='sigmoid'))
```

```
[ ] model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Εικόνα 29 *Φάση μοντελοποίησης Νευρωνικού Δικτύου*

Τελικά, η μέθοδος **`compile`** διαμορφώνει το μοντέλο για εκπαίδευση προσδιορίζοντας τη συνάρτηση απώλειας (loss function), τον αλγόριθμο βελτιστοποίησης (optimizer algorithm) και τη μετρική αξιολόγησης (evaluation metric).

Όπως φαίνεται παρακάτω *(Εικόνα 30)* η μέθοδος **`fit`** εκπαιδεύει το μοντέλο χρησιμοποιώντας τα παρεχόμενα δεδομένα εκπαίδευσης. Καθορίζει τα χαρακτηριστικά εισόδου και τη μεταβλητή στόχου (target variable) για εκπαίδευση, τον αριθμό των epochs, το batch size και τα δεδομένα επικύρωσης (validation data). Το όρισμα verbose καθορίζει το επίπεδο των πληροφοριών που εμφανίζονται κατά τη διάρκεια της εκπαίδευσης.

```
[ ]  model.fit(x=X_train,
                y=y_train,
                epochs=40,
                batch_size=512,
                validation_data=(X_test, y_test), verbose=1)
```

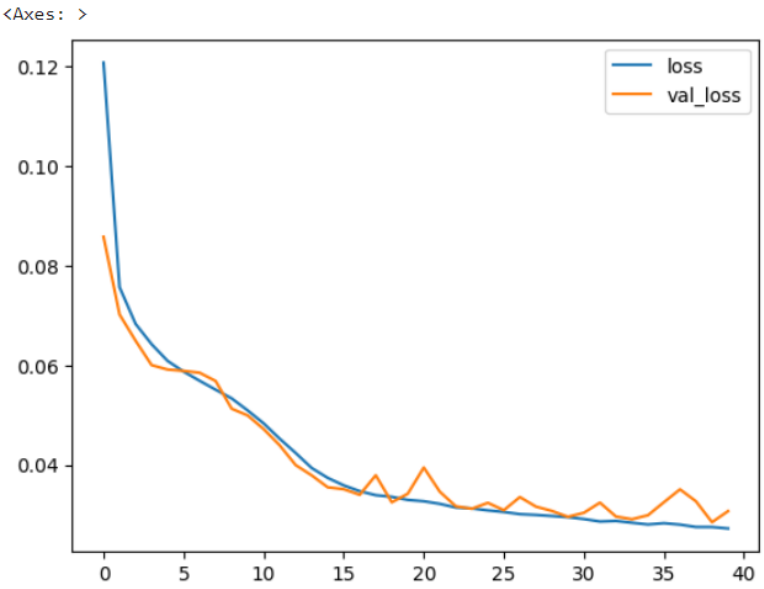
```
Epoch 1/40
2096/2096 [==============================] - 13s 5ms/step - loss: 0.1207 - accuracy: 0.9508 - val_loss: 0.0858 - val_accuracy: 0.9659
Epoch 2/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0757 - accuracy: 0.9705 - val_loss: 0.0702 - val_accuracy: 0.9725
Epoch 3/40
2096/2096 [==============================] - 12s 6ms/step - loss: 0.0683 - accuracy: 0.9738 - val_loss: 0.0649 - val_accuracy: 0.9752
Epoch 4/40
2096/2096 [==============================] - 10s 5ms/step - loss: 0.0642 - accuracy: 0.9757 - val_loss: 0.0600 - val_accuracy: 0.9777
Epoch 5/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0609 - accuracy: 0.9772 - val_loss: 0.0591 - val_accuracy: 0.9782
Epoch 6/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0587 - accuracy: 0.9781 - val_loss: 0.0589 - val_accuracy: 0.9787
Epoch 7/40
2096/2096 [==============================] - 12s 6ms/step - loss: 0.0569 - accuracy: 0.9790 - val_loss: 0.0585 - val_accuracy: 0.9783
Epoch 8/40
2096/2096 [==============================] - 9s 4ms/step - loss: 0.0551 - accuracy: 0.9798 - val_loss: 0.0568 - val_accuracy: 0.9786
Epoch 9/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0534 - accuracy: 0.9806 - val_loss: 0.0513 - val_accuracy: 0.9815
Epoch 10/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0509 - accuracy: 0.9815 - val_loss: 0.0499 - val_accuracy: 0.9817
Epoch 11/40
2096/2096 [==============================] - 12s 6ms/step - loss: 0.0483 - accuracy: 0.9826 - val_loss: 0.0472 - val_accuracy: 0.9833
Epoch 12/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0453 - accuracy: 0.9836 - val_loss: 0.0439 - val_accuracy: 0.9844
Epoch 13/40
2096/2096 [==============================] - 10s 5ms/step - loss: 0.0424 - accuracy: 0.9847 - val_loss: 0.0400 - val_accuracy: 0.9857
Epoch 14/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0394 - accuracy: 0.9857 - val_loss: 0.0379 - val_accuracy: 0.9863
Epoch 15/40
2096/2096 [==============================] - 11s 5ms/step - loss: 0.0374 - accuracy: 0.9864 - val_loss: 0.0355 - val_accuracy: 0.9870
Epoch 16/40
```

Εικόνα 30 *Φάση εκπαίδευσης Νευρωνικού Δικτύου*

Αποθηκεύουμε το ιστορικό εκπαίδευσης του μοντέλου στο dataframe **`losses`**, ώστε να μπορούμε πιο εύκολα να αναλύσουμε και να οπτικοποιήσουμε τη διαδικασία εκπαίδευσης.

Ο κώδικας που ακολουθεί *(Εικόνα 31)* δημιουργεί μια γραφική παράσταση που δείχνει την απώλεια εκπαίδευσης (training loss) και την απώλεια επικύρωσης (validation loss) κατά τη διάρκεια της εκπαίδευσης. Η γραμμή **`loss`** αντιπροσωπεύει την απώλεια εκπαίδευσης, η οποία υποδεικνύει πόσο καλά ταιριάζει το μοντέλο στα δεδομένα εκπαίδευσης. Η γραμμή **`val_loss`** αντιπροσωπεύει την απώλεια επικύρωσης, η οποία δείχνει την απόδοση του μοντέλου στα δεδομένα επικύρωσης κατά τη διάρκεια της εκπαίδευσης. Η γραφική παράσταση μπορεί να βοηθήσει στην οπτικοποίηση της σύγκλισης ή της απόκλισης της απόδοσης του μοντέλου κατά τη διάρκεια της εκπαίδευσης και μπορεί να παρέχει πληροφορίες για πιθανά ζητήματα overfitting ή κακής προσαρμογής (underfitting).

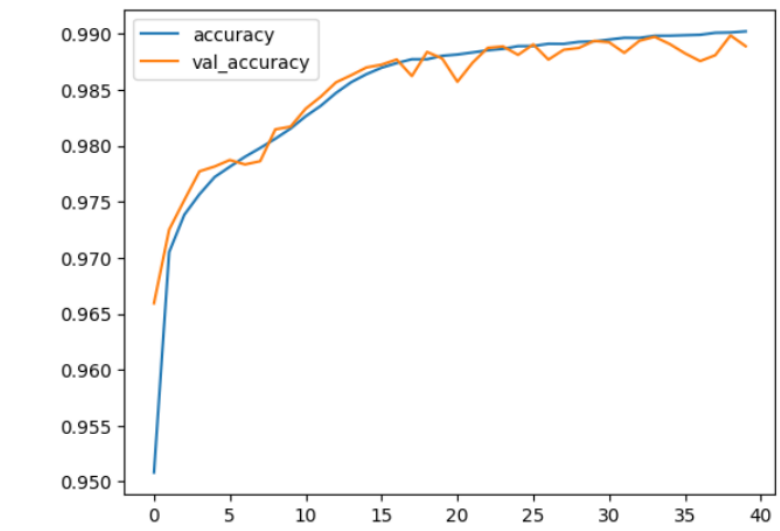
```
losses[['loss','val_loss']].plot()
```



Εικόνα 31 *Γράφημα loss και val_loss*

Στη συνέχεια, ο κώδικας δημιουργεί μια γραφική παράσταση που δείχνει την ακρίβεια της εκπαίδευσης (training accuracy) και την ακρίβεια επικύρωσης (validation accuracy) κατά τη διάρκεια της εκπαίδευσης *(Εικόνα 32).* Η γραμμή **`accuracy`** αντιπροσωπεύει την ακρίβεια εκπαίδευσης, η οποία δείχνει πόσο καλά το μοντέλο προβλέπει τα δεδομένα εκπαίδευσης. Η γραμμή **`val_accuracy`** αντιπροσωπεύει την ακρίβεια επικύρωσης, η οποία δείχνει την απόδοση του μοντέλου στα δεδομένα επικύρωσης κατά τη διάρκεια της εκπαίδευσης.

```
losses[['accuracy','val_accuracy']].plot()
```



Εικόνα 32 *Γράφημα accuracy και val_accuracy*

Ο κώδικας που ακολουθεί δημιουργεί προβλέψεις για τα δεδομένα δοκιμής (test data) χρησιμοποιώντας το εκπαιδευμένο μοντέλο. Στη συνέχεια, ο κώδικας προχωρά στην εκτύπωση μιας αναφοράς ταξινόμησης χρησιμοποιώντας τη συνάρτηση **`classification_report`** από το module **`sklearn.metrics`**. Έπειτα κατασκευάζει τον πίνακα σύγχυσης confusion matrix. Το dataframe που θα προκύψει θα εμφανίσει τις προβλέψεις True Positive (TP), True Negative (TN), False Positive (FP) και False Negative για κάθε κλάση του προβλήματος ταξινόμησης.

```
[ ]  predictions = (model.predict(X_test) > 0.5).astype("int32")

     print(classification_report(y_test, predictions))
```

```
8382/8382 [==============================] - 12s 1ms/step
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     53089
           1       0.99      0.99      0.99    215117

    accuracy                           0.99    268206
   macro avg       0.98      0.98      0.98    268206
weighted avg       0.99      0.99      0.99    268206
```

```
[ ]  pd.DataFrame(confusion_matrix(y_test, predictions))
```

	0	1
0	51552	1537
1	1439	213678

Εικόνα 33 *Confusion Matrix για το μοντέλο ANN*

Με βάση τα προηγούμενα αποτελέσματα *(Εικόνα 33),* μπορούμε να αναλύσουμε την απόδοση του μοντέλου χρησιμοποιώντας διάφορες μετρικές (metrics). Η ακρίβεια (precision) μετρά την αναλογία των σωστά προβλεπόμενων θετικών δειγμάτων από όλα τα δείγματα που προβλέπονται ως θετικά *[12]*. Για την κλάση 0, η ακρίβεια είναι 0.97, πράγμα που σημαίνει ότι το 97% των δειγμάτων που προβλέφθηκαν ως κατηγορία 0 ήταν στην πραγματικότητα κατηγορία 0. Για την κλάση 1, η ακρίβεια είναι 0.99, υποδεικνύοντας ότι το 99% των δειγμάτων που προβλέφθηκαν ως κατηγορία 1 ήταν στην πραγματικότητα κατηγορία 1. Οι πολύ υψηλές τιμές της ακρίβειας (precision) υποδεικνύουν καλύτερη απόδοση όσον αφορά στη σωστή αναγνώριση θετικών δειγμάτων.

Η ανάκληση (recall), γνωστή και ως ευαισθησία ή πραγματικό θετικό ποσοστό, μετρά την αναλογία των σωστά προβλεπόμενων θετικών δειγμάτων από όλα τα πραγματικά θετικά δείγματα. Για την κλάση 0, η τιμή της ανάκλησης είναι 0.97, που σημαίνει ότι το 97% των πραγματικών δειγμάτων κατηγορίας 0 αναγνωρίστηκαν σωστά ως κατηγορίας 0. Για την κατηγορία 1, η ανάκληση είναι 0.99, υποδεικνύοντας ότι το 99% των πραγματικών δειγμάτων κατηγορίας 1 αναγνωρίστηκαν σωστά ως κατηγορία 1.

Η βαθμολογία F1 (F1 score) είναι η αρμονική μέση τιμή της ακρίβειας και της ανάκλησης. Παρέχει ένα ισορροπημένο μέτρο που λαμβάνει υπόψη τόσο την ακρίβεια όσο και την ανάκληση. Για την κατηγορία 0, η βαθμολογία F1 είναι 0.97, ενώ για την κατηγορία 1 είναι 0.99. Οι υψηλότερες βαθμολογίες F1 υποδηλώνουν καλύτερη συνολική απόδοση στον σωστό προσδιορισμό τόσο των θετικών όσο και των αρνητικών δειγμάτων.

Η υποστήριξη (support) υποδεικνύει τον αριθμό των δειγμάτων σε κάθε κατηγορία στο σύνολο δεδομένων δοκιμής. Για την κλάση 0, η υποστήριξη είναι 53.089, ενώ για την κλάση 1, είναι 215.117.

Η συνολική ακρίβεια (accuracy) του μοντέλου υπολογίζεται ως η αναλογία των σωστά προβλεπόμενων δειγμάτων από όλα τα δείγματα. Η ακρίβεια είναι 0.99, που σημαίνει ότι το μοντέλο πέτυχε ακρίβεια 99% στο σύνολο δεδομένων δοκιμής επιβεβαιώνοντας με αυτόν τον τρόπο τη συνολική αποτελεσματικότητα του μοντέλου.

Με βάση τώρα τον confusion matrix, λαμβάνουμε την πληροφορία:

True Positive (TP): Το μοντέλο προέβλεψε σωστά 51.552 δείγματα ως κατηγορία 0

False Positive (FP): Το μοντέλο προέβλεψε εσφαλμένα 1.537 δείγματα ως κατηγορία 1 όταν ήταν στην πραγματικότητα κατηγορία 0.

False Negative (FN): Το μοντέλο προέβλεψε εσφαλμένα 1.439 δείγματα ως κατηγορία 0 όταν ήταν στην πραγματικότητα κατηγορία 1.

True Negative (TN): Το μοντέλο προέβλεψε σωστά 213.678 δείγματα ως κατηγορία 1.

Από αυτές τις τιμές, μπορούμε να αντλήσουμε πρόσθετες πληροφορίες:

Sensitivity για την κλάση 0: $\frac{TP}{TP+FN} = 0.973$

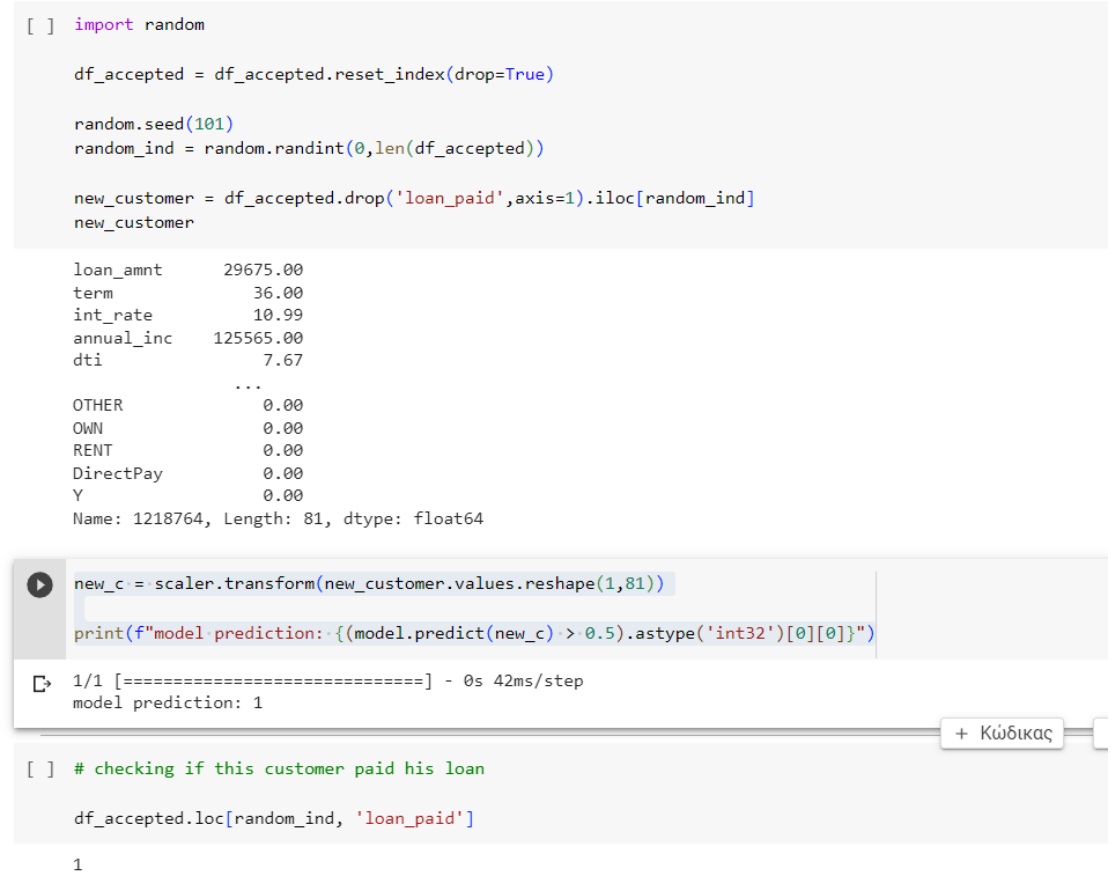
Specificity για την κλάση 1: $\frac{TN}{TN+FP} = 0.992$

False Positive Rate για την κλάση 0: $\frac{FP}{FP+TN} = 0.007$

False Negative Rate για την κλάση 1: $\frac{FN}{FN+TP} = 0.027$

Αυτές οι πρόσθετες μετρήσεις παρέχουν μια πιο λεπτομερή κατανόηση της απόδοσης του μοντέλου. Οι τιμές υψηλής **Sensitivity** και **Specificity** υποδηλώνουν ότι το μοντέλο είναι αποτελεσματικό στον σωστό εντοπισμό δειγμάτων κατηγορίας 0 και κατηγορίας 1. Τα χαμηλά ποσοστά **False Positive Rate** και **False Negative Rate** δείχνουν ότι το μοντέλο κάνει σχετικά λίγες εσφαλμένες προβλέψεις. Ο παρεχόμενος πίνακας σύγχυσης (confusion matrix) επιβεβαιώνει την υψηλή απόδοση του μοντέλου, με την πλειονότητα των δειγμάτων να ταξινομούνται σωστά όπως υποδεικνύεται από τις αντίστοιχες μετρήσεις.

Τέλος, μπορούμε να προσομοιώσουμε το παραπάνω μοντέλο για έναν τυχαίο πελάτη, όπως φαίνεται παρακάτω:



Εικόνα 34 *Προσομοίωση μοντέλου για έναν τυχαίο πελάτη*

Με βάση το αποτέλεσμα, η πρόβλεψη του μοντέλου για τον νέο πελάτη είναι 1. Αυτό δείχνει ότι το μοντέλο προβλέπει ότι ο νέος πελάτης ανήκει στη θετική κατηγορία, δηλαδή προέβη σε αποπληρωμή του δανείου του.

- **Τυχαία Δάση (Random Forests)**

Αρχικά κάνουμε import την κλάση **`RandomForestClassifier`** από το module **`sklearn.ensemble`** στο scikit-learn. Ο αλγόριθμος Random Forest είναι ένας εποπτευόμενος (supervised) αλγόριθμος μάθησης και συνδυάζει πολλαπλά δέντρα αποφάσεων (decision trees) για να δημιουργήσει ένα πιο ισχυρό και ακριβές μοντέλο. Κάθε δέντρο απόφασης στο τυχαίο δάσος δημιουργείται χρησιμοποιώντας ένα τυχαία επιλεγμένο υποσύνολο των δεδομένων εκπαίδευσης και ένα τυχαίο υποσύνολο χαρακτηριστικών. Η τελική πρόβλεψη του τυχαίου δάσους καθορίζεται με τη συγκέντρωση των προβλέψεων όλων των μεμονωμένων δέντρων για παράδειγμα χρησιμοποιώντας την πλειοψηφία για εργασίες ταξινόμησης (classification) ή τον μέσο όρο για εργασίες παλινδρόμησης (regression).

Στη συνέχεια, εισάγεται η συνάρτηση **`accuracy_score`** από το module **`sklearn.metrics`** στο scikit-learn. Η συνάρτηση **`accuracy_score`** χρησιμοποιείται ειδικά για τον υπολογισμό της ακρίβειας των μοντέλων ταξινόμησης. Συγκρίνει τις προβλεπόμενες ετικέτες (labels) ή τιμές που δημιουργούνται από ένα μοντέλο με τις πραγματικές ετικέτες για να προσδιορίσει την ακρίβεια των προβλέψεων του μοντέλου *[16]*.

```
[ ] from sklearn.preprocessing import MinMaxScaler
    scaler = MinMaxScaler()

[ ] x_train = scaler.fit_transform(x_train)
    x_test = scaler.transform(x_test)

[ ] x_train.shape,x_test.shape,y_train.shape,y_test.shape

    ((887550, 8), (457224, 8), (887550,), (457224,))

[ ] from sklearn.ensemble import RandomForestClassifier

[ ] model = RandomForestClassifier(max_depth=6)

[ ] model.fit(x_train,y_train)

    RandomForestClassifier
    RandomForestClassifier(max_depth=6)

[ ] pred_lr = model.predict(x_test)

[ ] from sklearn.metrics import accuracy_score  #for accuracy_score

    accuracy= accuracy_score(pred_lr,y_test)*100
    print('Accuracy of RandomForestClassifier: {:.2f}'.format(accuracy))

    Accuracy of RandomForestClassifier: 80.15
```

Εικόνα 35 *Μέτρηση ακρίβειας του μοντέλου Random Forests*

Η έξοδος Accuracy of RandomForestClassifier: 80.15 υποδεικνύει ότι το μοντέλο RandomForestClassifier πέτυχε ακρίβεια 80.15% στα δεδομένα τα οποία δοκιμάστηκε (test data). Αυτό σημαίνει ότι περίπου το 80.15% των προβλεπόμενων ετικετών που δημιουργούνται από το μοντέλο ταιριάζουν με τις πραγματικές ετικέτες από τα δεδομένα δοκιμής. Το accuracy score είναι μια μετρική αξιολόγησης (evaluation metric) που χρησιμοποιείται συνήθως στις εργασίες ταξινόμησης και οι υψηλότερες τιμές υποδεικνύουν καλύτερη απόδοση. Σε αυτήν την περίπτωση, μια ακρίβεια 80.15% υποδηλώνει ότι το μοντέλο RandomForestClassifier κάνει σωστές προβλέψεις για ένα σημαντικό μέρος των δειγμάτων που δοκιμάζονται (test samples). Είναι σημαντικό να σημειωθεί ότι η ακρίβεια (accuracy) από μόνη της μπορεί να μην παρέχει μια πλήρη εικόνα της απόδοσης του μοντέλου, ειδικά εάν το

σύνολο δεδομένων δεν είναι ισορροπημένο (unbalanced) ή εάν η εσφαλμένη ταξινόμηση ορισμένων κλάσεων είναι πιο κρίσιμη από άλλες.

Τέλος, για να πάρουμε μια πιο λεπτομερή αναφορά για την απόδοση του μοντέλου θα κάνουμε χρήση της συνάρτησης **`classification_report`**. Η συνάρτηση αυτή περιλαμβάνει μετρήσεις όπως η ακρίβεια (precision), η ανάκληση (recall) η f1-score και η support. Επιπλέον, παρέχει μακρο-μέσους (macro-averages) και σταθμισμένους μέσους (weighted averages) σε όλες τις κατηγορίες, οι οποίοι μπορούν να δώσουν μια συνολική περίληψη της απόδοσης του μοντέλου.

```
[ ] from sklearn.metrics import classification_report

    print(classification_report(y_test, pred_lr))
```

```
              precision    recall  f1-score   support

           0       0.59      0.01      0.02     91078
           1       0.80      1.00      0.89    366146

    accuracy                           0.80    457224
   macro avg       0.70      0.50      0.46    457224
weighted avg       0.76      0.80      0.72    457224
```

Εικόνα 36 *Classification Report του μοντέλου Random Forests*

Η κλάση 0 έχει χαμηλό precision 0.59, υποδεικνύοντας ότι μεταξύ των προβλεπόμενων περιπτώσεων που επισημαίνονται ως κατηγορία 0, μόνο το 59% είναι πραγματικά True Positives. Η ανάκληση (recall) είναι εξαιρετικά χαμηλή, στο 0.01, υποδηλώνοντας ότι το μοντέλο δυσκολεύεται να αναγνωρίσει σωστά τις περιπτώσεις που ανήκουν στην κλάση 0. Η f-1 score, που συνδυάζει precision και recall, είναι επίσης πολύ χαμηλή στο 0.02. Ωστόσο, η support υποδεικνύει έναν σημαντικό αριθμό περιπτώσεων για την κλάση 0 (91078).

Η κλάση 1 έχει υψηλό precision 0.80, υποδεικνύοντας ότι το μοντέλο προσδιορίζει με ακρίβεια ένα μεγάλο ποσοστό περιπτώσεων που επισημαίνονται ως κλάση 1. Η recall είναι τέλεια, στο 1.00, υποδεικνύοντας ότι το μοντέλο προσδιορίζει σωστά όλες τις περιπτώσεις που ανήκουν στην κλάση 1. Το f-1 score είναι υψηλό, στο 0.89, υποδηλώνοντας καλή ισορροπία μεταξύ precision και recall. Η support υποδεικνύει έναν σημαντικό αριθμό περιπτώσεων για την κλάση 1 (366146).

Η συνολική ακρίβεια (accuracy) του μοντέλου αναφέρεται ως 0.80, που σημαίνει ότι περίπου το 80% των προβλέψεων ταιριάζουν με τα true labels. Ο μέσος όρος macro παρέχει μέσο όρο precision, recall, f-1 score και support σε όλες τις κατηγορίες. Σε αυτήν την περίπτωση, ο μέσος όρος της macro δείχνει precision 0.70, recall 0,50 και f-1 score 0.46. Αυτές οι τιμές υποδεικνύουν ότι η απόδοση του μοντέλου είναι ασθενέστερη για την κλάση 0 σε σύγκριση με την κλάση 1.

Ο σταθμισμένος μέσος όρος (weighted average) λαμβάνει υπόψη την ανισορροπία της κατηγορίας και παρέχει έναν μέσο σταθμισμένο από το support της κάθε κλάσης. Η σταθμισμένη μέση precision είναι 0.76, η recall είναι 0.80 και το f-1 score είναι 0.72. Αυτές οι τιμές υποδεικνύουν μια εύλογα ισορροπημένη απόδοση μεταξύ των κλάσεων, λαμβάνοντας υπόψη το αντίστοιχο support.

- **Support Vector Machines (SVM)**

Εκτελώ τον ταξινομητή **`Support Vector Classifier (SVC)`** από τη βιβλιοθήκη **`sklearn`** της Python. Στη συνέχεια καλείται η μέθοδος **`fit`** στον ταξινομητή για να εκπαιδεύσει το μοντέλο. Η μέθοδος δέχεται δύο ορίσματα, τα X_train και y_train. Το X_train αντιπροσωπεύει τα χαρακτηριστικά εισόδου ή τα δεδομένα εκπαίδευσης, ενώ το y_train αντιπροσωπεύει τις αντίστοιχες ετικέτες (labels).

```
 # import svc
from sklearn.svm import SVC

# Creating an instance
classifier = SVC(kernel='rbf',random_state=None)

# Fitting the model
classifier.fit(X_train, y_train)
```

```
[ ] classifier.intercept_
```

```
[ ] classifier.n_support_
```

```
[ ] classifier.support_vectors_
```

```
[ ] # Predicting the values
    y_pred_train = classifier.predict(X_train)
    y_pred_test = classifier.predict(X_test)
```

Εικόνα 37 *Ταξινομητής SVM*

Το αποτέλεσμα είναι να πάρουμε έναν πίνακα (array) ο οποίος καθορίζει τη θέση του ορίου απόφασης (decision boundary) σε σχέση με τον χώρο χαρακτηριστικών (feature space). Στη συνέχεια παίρνουμε το πλήθος των support vectors. Ο αριθμός των support vectors για κάθε τάξη (class) παρέχει πληροφορίες για την πολυπλοκότητα και τα χαρακτηριστικά του προβλήματος ταξινόμησης. Ένας μεγάλος αριθμός διανυσμάτων μπορεί να υποδεικνύει ένα πιο σύνθετο όριο απόφασης (decision boundary) ή ένα σύνολο δεδομένων με επικαλυπτόμενες (overlapping) κλάσεις.

Στη συνέχεια, ο αλγόριθμος μας δίνει έναν πίνακα του οποίου η κάθε γραμμή αντιπροσωπεύει ένα support vector και οι στήλες αντιστοιχούν στα χαρακτηριστικά

(features) των δεδομένων *[19]*. Ο αριθμός των στηλών στον πίνακα αντιστοιχεί στον αριθμό των χαρακτηριστικών που χρησιμοποιούνται στον ταξινομητή SVM. Εξετάζοντας τα support vectors, μπορούμε να αποκτήσουμε πληροφορίες για τα συγκεκριμένα δείγματα (samples) από τα δεδομένα εκπαίδευσης που έχουν τη μεγαλύτερη επιρροή στο όριο απόφασης. Αυτά τα διανύσματα είναι ζωτικής σημασίας για τον ταξινομητή SVM για την πραγματοποίηση ακριβών προβλέψεων.

Ακολούθως, θα υπολογίσουμε τον πίνακα σύγχυσης (confusion matrix) καθώς και την ακρίβεια (accuracy) του ταξινομητή *(Εικόνα 38).* Ο πίνακας σύγχυσης παρέχει μια λεπτομερή ανάλυση των προβλέψεων του μοντέλου, επιτρέποντάς μας να αναλύσουμε την απόδοση για κάθε κατηγορία ξεχωριστά. Βοηθά στην αξιολόγηση της ακρίβειας του ταξινομητή και στον εντοπισμό τυχόν προκαταλήψεων (biases) ή σφαλμάτων στις προβλέψεις. Σε επόμενο βήμα, θα αξιολογήσουμε την απόδοση του ταξινομητή SVM στα μη ορατά δεδομένα στο σύνολο δοκιμής. Το accuracy score παρέχει μια ενιαία μέτρηση για την αξιολόγηση της συνολικής απόδοσης του ταξινομητή SVM. Αντιπροσωπεύει το ποσοστό των σωστά ταξινομημένων δειγμάτων και χρησιμοποιείται συνήθως για την αξιολόγηση της προγνωστικής ακρίβειας του μοντέλου. Εξετάζοντας τον confusion matrix και το accuracy score τόσο για το σετ εκπαίδευσης όσο και για τα σετ δοκιμών, μπορείτε να αποκτήσουμε πληροφορίες σχετικά με την απόδοση του ταξινομητή SVM, συμπεριλαμβανομένης της ικανότητάς του να ταξινομεί σωστά τις διαφορετικές κατηγορίες και τη συνολική του ακρίβεια.

```
from sklearn.metrics import confusion_matrix, accuracy_score

# Confusion matrix
cm_train = confusion_matrix(y_train, y_pred_train)
print(cm_train)

cm_test = confusion_matrix(y_test, y_pred_test)
print(cm_test)
```

```
[ ]  # Accuracy
     accuracy_train = accuracy_score(y_train, y_pred_train)
     print(accuracy_train)
```

```
[ ]  accuracy_test = accuracy_score(y_test, y_pred_test)
     print(accuracy_test)
```

Εικόνα 38 *Confusion Matrix και accuracy score για το μοντέλο SVM*

4.3 Συμπεράσματα – Μελλοντική έρευνα

Η παρούσα εργασία ασχολήθηκε με την έννοια της πρόγνωσης του πιστωτικού κινδύνου (credit risk prediction) με τη βοήθεια αλγορίθμων μηχανικής μάθησης. Συγκεκριμένα, χρησιμοποιήσαμε τα μοντέλα Support Vector Machines (SVM), Random Forests και Artificial Neural Networks. Οι αλγόριθμοι αυτοί έδειξαν αξιοσημείωτη επιτυχία στην πρόβλεψη πιστωτικού κινδύνου. Έχουν την ικανότητα να χειριστούν μεγάλα και σύνθετα datasets, όπως στην περίπτωσή μας, να καταγράφουν μη γραμμικές σχέσεις και να εξάγουν σημαντικά μοτίβα (patterns) από τα δεδομένα.

Τα νευρωνικά δίκτυα και ειδικά οι αρχιτεκτονικές βαθιάς μάθησης (deep learning), παίζουν σημαντικό ρόλο στην πρόβλεψη του πιστωτικού κινδύνου. Κυριαρχούν στην εκμάθηση ιεραρχικών αναπαραστάσεων από τα δεδομένα και μπορούν να αποτυπώσουν περίπλοκα μοτίβα και αλληλεπιδράσεις. Τα νευρωνικά δίκτυα έχουν δείξει πολλά υποσχόμενα αποτελέσματα στην πρόβλεψη πιστωτικού κινδύνου και έχουν καλύτερη απόδοση από τα παραδοσιακά μοντέλα.

Τα Random Forests είναι ισχυρά μοντέλα που συνδυάζουν πολλαπλά δέντρα αποφάσεων. Προσφέρουν στιβαρότητα έναντι του overfitting, χειρίζονται καλά δεδομένα υψηλών διαστάσεων (high dimensional data) και μπορούν να αποτυπώσουν τόσο γραμμικές όσο και μη γραμμικές σχέσεις. Τα Random Forests παρέχουν ακριβείς και αξιόπιστες προβλέψεις πιστωτικού κινδύνου.

Τα Support Vector Machines (SVM) είναι αποτελεσματικά σε εργασίες πρόβλεψης πιστωτικού κινδύνου, καθώς στοχεύουν στην εύρεση ενός βέλτιστου υπερεπίπεδου που διαχωρίζει διαφορετικές κλάσεις. Ο αλγόριθμος SVM μπορεί να χειριστεί τόσο γραμμικές όσο και μη γραμμικές σχέσεις χρησιμοποιώντας kernel functions. Ωστόσο, τα SVM μπορεί να αντιμετωπίσουν προκλήσεις με μεγάλα σύνολα δεδομένων, όπως στην περίπτωσή μας, όπου το dataset είχε μεγάλο όγκο, οπότε απαιτούν προσεκτική ρύθμιση των παραμέτρων.

Η επιλογή του κατάλληλου μοντέλου μηχανικής μάθησης για την πρόβλεψη πιστωτικού κινδύνου εξαρτάται από διάφορους παράγοντες όπως το μέγεθος των δεδομένων, η πολυπλοκότητα των χαρακτηριστικών, η ανάγκη για κλιμάκωση των δεδομένων και οι διαθέσιμοι υπολογιστικοί πόροι, π.χ RAM. Τα Random Forests και τα νευρωνικά δίκτυα κατάφεραν υψηλές προγνωστικές επιδόσεις, αλλά μπορεί να μην έχουν καλή ερμηνευσιμότητα, ενώ τα SVM παρείχαν ένα πιο κατανοητό μοντέλο σε βάρος όμως των πιθανών προβλημάτων απόδοσης.

Όπως αναφέραμε, το υπολογιστικό κόστος παίζει κυρίαρχο ρόλο στην εφαρμογή των ανωτέρω μοντέλων μηχανικής μάθησης. Πράγματι, ο αλγόριθμος SVM κατά την εκτέλεσή του χρειάστηκε αρκετή μνήμη RAM και ο χρόνος εκτέλεσης ήταν τεράστιος. Στον αντίποδα, η χρήση Τεχνητών Νευρωνικών Δικτύων παρουσίασε πολύ καλύτερη εφαρμογή στην πρόγνωση του πιστωτικού κινδύνου λόγω του ότι είναι κατάλληλα για μεγάλο όγκο δεδομένων, όπως η περίπτωσή μας.

Συμπερασματικά, η ενασχόληση με την πρόβλεψη του πιστωτικού κινδύνου κάνοντας χρήση μεθόδων μηχανικής μάθησης παρουσίασε εξαιρετικό ενδιαφέρον στην πρόβλεψη της αποπληρωμής ενός δανείου στο τρέχον εξελισσόμενο οικονομικό σύστημα. Χρησιμοποιώντας το dataset από το Lending Club, τα μοντέλα που εφαρμόσαμε απέδωσαν ικανοποιητικά, φέρνοντας πολύ καλά αποτελέσματα. Με τα αποτελέσματα αυτά μπορούμε να αποτιμήσουμε τον πιθανό πιστωτικό κίνδυνο

από τους υποψήφιους δανειστές. Επιπλέον, τα τραπεζικά ινστιτούτα, σε παγκόσμιο επίπεδο, θα μπορούν να χρησιμοποιούν ευρέως την σύγχρονη τεχνολογία, ώστε να λειτουργούν σε ένα πιο υγιές και διαφανές πλαίσιο, το οποίο μπορεί να αποφέρει μακροπρόθεσμα σημαντικά οφέλη.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] *Crouhy, M., Galai, D., & Mark, R. (2014). The Essentials of Risk Management (2nd ed.). McGraw-Hill Education.*

[2] *Altman, E. I., Resti, A., & Sironi, A. (Eds.). (2019). Managing and Measuring Credit Risk: Emerging Global Standards and Regulations after the Financial Crisis. John Wiley & Sons.*

[3] *Bluhm, C., Overbeck, L., & Wagner, C. (2016). An Introduction to Credit Risk Modeling. CRC Press.*

[4] *Aurélien Géron (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Publisher. O'Reilly Media Publication*

[5] *Ganesan, S., & Arumugam, S. (2017). Credit Risk Prediction using Machine Learning Techniques. International Journal of Applied Engineering Research, 12(4), 609-616.*

[6] *Li, H., & Chen, Y. (2018). A Machine Learning Approach to Credit Risk Evaluation. Intelligent Automation & Soft Computing, 24(2), 415-426.*

[7] *May, R., Nguyen, N., Wills, C. E., & Kanagasundaram, A. (2019). Credit Scoring and Default Prediction using Machine Learning Techniques. Expert Systems with Applications, 121, 494-507.*

[8] *Ian Goodfellow, Yoshua Bengio, and Aaron Courville. (2016). Deep Learning. MIT Press Publication*

[9] *Michael R. Berthold, Christian Borgelt, Frank Höppner and Frank Klawonn. (2010). Guide to Intelligent Data Analysis. How to Intelligently Make Sense of Real Data. Springer*

[10] *EMC Education Services (eds.), 2015, Data science and big data analytics: Discovering, analyzing, visualizing and presenting data, John Wiley & Sons Inc.*

[11] *P. Wentworth, J. Elkner, A.B. Downey & C. Meyers (2021).* Learn Python the right way: How to think like a computer scientist.

[12] *J. Leskovec, A. Rajaraman & J.D. Ullman (2020).* Mining of Massive Datasets *(3rd edition). Cambridge University Press*

[13] *P.-N. Tan, M. Steinbach, A. Karpatne & V. Kumar (2021).* Introduction to Data Mining *(2nd Edition). Pearson.*

[14] *T. Benschop & M. Welch (2019).* Statistical Disclosure Control for Microdata: Practice*. The World Bank.*

ΔΙΚΤΥΟΓΡΑΦΙΑ

[15] *https://www.kaggle.com/datasets/wordsforthewise/lending-club*

[16] *https://www.kaggle.com/code/apriyad/logistic-and-random-forest*

[17] *https://www.kaggle.com/code/errearanhas/default-prediction-neural-network-99-acc/notebook*

[18] *https://github.com/yanhan-si/LendingClub-Loan-Default-Prediction/blob/master/Lending_Club_Loan_Default_Prediction.ipynb*

[19] *https://www.analyticsvidhya.com/blog/2021/10/loan-status-prediction-using-support-vector-machine-algorithm/*

[20] *https://www.kaggle.com/code/mariiagusarova/data-scaling-and-skewness-handling*

[21] *https://www.reneshbedre.com/blog/support-vector-machine.html*

[22] *https://colab.research.google.com/*