ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών (Π.Μ.Σ.)
Πληροφοριακά Συστήματα & Υπηρεσίες
Ειδίκευση: Προηγμένα Πληροφοριακά Συστήματα

Εξόρυξη γνώσης από δεδομένα δικηγορικού γραφείου διαχείρισης ληξιπρόθεσμων οφειλών

ΠΟΖΑΤΖΙΔΗΣ ΑΝΑΣΤΑΣΙΟΣ, Α.Μ. 2146

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΧΑΛΚΙΔΗ ΜΑΡΙΑ

ΠΕΙΡΑΙΑΣ 2023

Περίληψη

Η ανάπτυξη των παραγωγικών δυνάμεων αποτελεί στόχο κάθε κοινωνικοοικονομικού σχηματισμού, στα πλαίσια της σύγχρονης κοινωνίας και της ελεύθερης οικονομίας , η ανάπτυξη και η βελτίωση της παραγωγικότητας αποτελεί στόχο κάθε ξεχωριστής επιχείρησης τόσο σε διεθνές όσο και σε εγχώριο πεδίο δράσης. Εξάλλου η εξέλιξη της παραγωγής με νέες παραγωγικές δυνατότητες καθιστά την κάθε επιχείρηση πιο ανθεκτική έναντι του ανταγωνισμού και άρα πιθανότερο να επιβιώσει και να εξελιχθεί.

Η εισαγωγή νέων, τεχνολογικά προηγμένων, τεχνικών και μεθόδων είναι ένας από τους τρόπους εξέλιξης της παραγωγής καθώς καθιστά ευκολότερη τη διεκπεραίωση, παρακολούθηση, έλεγχο και πρόβλεψη των διαδικασιών τις οποίες πρέπει να ακολουθήσει η εκάστοτε επιχείρηση. Οδηγεί επίσης σε συγκεκριμένες αποφάσεις οι οποίες μπορούν να αλλάξουν τους επιμέρους στόχους ακόμα και τη στρατηγική της επιχείρησης.

Σε αυτά τα πλαίσια η επιστήμη των δεδομένων και πιο συγκεκριμένα η εξόρυξη γνώσης από την πληθώρα δεδομένων τα οποία ανήκουν στην εκάστοτε επιχείρηση καθίστανται πολύτιμα και η ανάλυση τους αποτελεί σημαντικό στόχο που βοηθάει στην κατανόηση της οικονομικής και παραγωγικής πραγματικότητας και άρα στην αποκάλυψη των πιθανών τάσεων που παρατηρούνται , δηλαδή στην δυνατότητα πρόβλεψης και χάραξης στρατηγικής.

Με αυτή τη σκέψη ως βασικό σημείο εκκίνησης πραγματοποιήθηκε η διπλωματική διατριβή. Τα τμήματα πληροφορικής καθίστανται όλο και πιο σημαντικά για οποιαδήποτε επιχείρηση. Το ίδιο συμβαίνει και για τις εταιρείες που δραστηριοποιούνται στον τραπεζικό τομέα κι δεν είναι τράπεζες, όπως για παράδειγμα τα δικηγορικά γραφεία μέσα στα οποία υπάρχει call center. Στη σύγχρονη Ελλάδα , τα συγκεκριμένα γραφεία έρχονται σε επικοινωνία καθημερινά με εκατοντάδες πολίτες και διεκπεραιώνουν πολλές σημαντικές ενέργειες που αφορούν οφειλές δανειοληπτών ή ληξιπρόθεσμων λογαριασμών και αντίστοιχα ενημερώνουν τα ιδρύματα που παρείχαν το δάνειο ή τους ανήκει μια παροχή.

Η λειτουργία αυτή καταλήγει να δημιουργεί βάσεις δεδομένων με πολλά διαφορετικά στοιχεία από διαφορετικές πηγές. Η άντληση, επεξεργασία και ανάλυση από τα δεδομένα μπορεί να παρέχει πολύτιμα εργαλεία και την εξόρυξη κατάλληλης γνώσης για την βελτιστοποίηση των διαδικασιών της επιχείρησης ακόμα και την αλλαγή στρατηγικής. Απασχόλησε κυρίως η εφαρμογή τεχνικών για την δυνατότητα πρόβλεψης αποπληρωμής των οφειλετών ή η συνέπεια τους.

Abstract

The development of the productive forces is the goal of every socio-economic formation, in the context of modern society and the free economy, the development of the productive forces is the goal of every individual enterprise both in the international and domestic field of action. Moreover, the development of production with new production capabilities makes each company more resistant to competition and therefore more likely to survive and develop.

The introduction of new, technologically advanced, techniques and methods is one of the ways to develop production as it makes it easier to process, monitor, control and predict the processes that each company must follow. It also leads to specific decisions that can change individual goals and even the company's strategy.

In these contexts, the science of data and more specifically the extraction of knowledge from the abundance of data that belongs to each company become valuable and their analysis is an important objective that helps to understand the economic and production reality and thus to reveal the possible trends that are observed, i.e. in the ability to predict and draw up a strategy.

With this thought as a basic starting point, the diploma thesis was carried out. IT departments are becoming increasingly important to any business. The same is the case for companies that operate in the banking sector and are not banks, such as law firms that have a call center. In modern Greece, the specific offices come into contact with hundreds of citizens every day and handle many important actions concerning the debts of borrowers or overdue accounts and accordingly inform the institutions that provided the loan or are entitled to a benefit.

This operation ends up creating large databases with many different items from different sources. Extraction, processing and analysis from data can provide valuable tools and the extraction of appropriate knowledge to optimize business processes and even change strategy.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κ. Χαλκίδη Μαρία για την πολύτιμη βοήθεια και συνεργασία της καθ’ όλη τη διάρκεια της διπλωματικής μου εργασίας και μέχρι την ολοκλήρωσή της.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Λίστα Εικόνων

1 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

Η ανάγκη καταγραφής, διαχείρισης των μεγάλων βάσεων δεδομένων αλλά και ανακάλυψης καινούργιας πληροφορίας μέσω αυτών έρχεται να ικανοποιήσει η εξόρυξη γνώσης από δεδομένα (data mining). Η εξόρυξη δεδομένων διευθετεί και επεξεργάζεται τα δεδομένα με την ανακάλυψη και αξιοποίηση προτύπων, δομών, μοντέλων, τάσεων και συσχετίσεων, με ένα αυτοματοποιημένο τρόπο. Η βέλτιστη απόκτηση γνώσης για μια επιχείρηση έχει ως αποτέλεσμα το σχεδιασμό και τη λήψη των βέλτιστων αποφάσεων, την αύξηση της παραγωγικότητας σε στρατηγικά και επιχειρησιακά επίπεδα και κατά συνέπεια την αύξηση της ανταγωνιστικότητας της έναντι άλλων επιχειρήσεων.

Η εξόρυξη δεδομένων μπορεί να θεωρηθεί ως η επιστήμη της εξερεύνησης μεγάλων συνόλων δεδομένων για εξαγωγή κρυφών, προηγουμένως άγνωστων και δυνητικά χρήσιμων πληροφοριών.

Θα παρουσιάσουμε κάποιες προσεγγίσεις λίγο πολύ παρόμοιες, οι οποίες θα σκιαγραφήσουν με σαφήνεια, τι είναι η εξόρυξη δεδομένων. Έτσι, με τον όρο εξόρυξη δεδομένων εννοούμε :

• Η αυτόματη αναζήτηση προτύπων σε τεράστιες βάσεις δεδομένων, χρησιμοποιώντας υπολογιστικές τεχνικές από στατιστικές, μηχανική μάθηση και αναγνώριση προτύπων.

• Η μη τετριμμένη εξαγωγή κρυφών, προηγουμένως άγνωστων και δυνητικά χρήσιμων πληροφοριών από δεδομένα·

• Η επιστήμη της εξαγωγής χρήσιμων πληροφοριών από μεγάλα σύνολα δεδομένων ή οι βάσεις δεδομένων.

• Η αυτόματη ή ημιαυτόματη εξερεύνηση και ανάλυση μεγάλων ποσοτήτων δεδομένων, προκειμένου να ανακαλυφθούν σημαντικά μοτίβα·

• Η διαδικασία αυτόματης ανακάλυψης πληροφοριών. Η αναγνώριση των προτύπων και των «κρυμμένων» σχέσεων στα δεδομένα.

Στην περίπτωση της εξόρυξης δεδομένων, το πρόβλημα μπορεί συνίσταται, για παράδειγμα, στον εντοπισμό παραγόντων, χωρίς να βασίζεται σε καμία a priori υπόθεση. Συμπερασματικά, οι μέθοδοι εξόρυξης δεδομένων επιδιώκουν να εντοπιστούν μοτίβα και κρυφές σχέσεις που δεν είναι πάντα προφανείς (και επομένως εύκολα αναγνωρίσιμες) υπό τις συνθήκες ορισμένων υποθέσεων.

1.1 ΤΟΜΕΙΣ ΧΡΗΣΗΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η αναγκαιότητα της «εξόρυξης» των δεδομένων εκφράζεται και με την χρήση της σε σημαντικούς τομείς της πραγματικής ζωής που χρειάζονται τέτοιες διερευνητικές τεχνικές [4]:

• Οικονομικά (επιχειρήσεις-χρηματοοικονομικά) - υπάρχει ήδη ένας τεράστιος όγκος δεδομένων σε διάφορους τομείς όπως: Δεδομένα Ιστού, ηλεκτρονικό εμπόριο, δεδομένα υπεραγορών, χρηματοοικονομικές και τραπεζικές συναλλαγές κ.λπ., έτοιμα για ανάλυση προκειμένου να ληφθούν οι βέλτιστες αποφάσεις·

• Υγειονομική περίθαλψη - υπάρχουν επί του παρόντος πολλές και διαφορετικές βάσεις δεδομένων στην υγειονομική περίθαλψη (ιατρικός και φαρμακευτικός τομέας), οι οποίοι αναλύθηκαν μόνο εν μέρει, ειδικά με συγκεκριμένα ιατρικά μέσα, που περιέχουν πολλές πληροφορίες, αλλά δεν έχουν διερευνηθεί επαρκώς.

• Επιστημονική έρευνα - υπάρχουν τεράστιες βάσεις δεδομένων που συγκεντρώθηκαν με τα χρόνια σε διάφορους τομείς (αστρονομία, μετεωρολογία, βιολογία, γλωσσολογία κ.λπ.), οι οποίες δεν μπορούν να εξερευνηθούν με παραδοσιακά μέσα.

Δεδομένου του γεγονότος ότι, αφενός, υπάρχει ένας τεράστιος όγκος ανεξερεύνητων δεδομένων και, από την άλλη πλευρά, τόσο η υπολογιστική ισχύς όσο και η επιστήμη των υπολογιστών έχουν αυξηθεί εκθετικά, η πίεση της χρήσης νέων μεθόδων για την αποκάλυψη οι πληροφορίες που «κρύβονται» στα δεδομένα αυξήθηκαν. Αξίζει να σημειωθεί ότι υπάρχουν πολλές πληροφορίες σε δεδομένα, σχεδόν αδύνατο να εντοπιστούν με παραδοσιακά μέσα και μόνο με τη χρήση της ανθρώπινης αναλυτικής ικανότητας.

1.2 ΣΤΟΧΟΙ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Ας παραθέσουμε παρακάτω δύο στόχους εξόρυξης δεδομένων για να διακρίνουμε με μεγαλύτερη σαφήνεια την περιοχή εφαρμογής της [4]:

• **Προγνωστικοί στόχοι** (π.χ. ταξινόμηση - classification, παλινδρόμηση - regression, ανίχνευση ανωμαλιών/ακραίων τιμών), που επιτυγχάνονται με τη χρήση ενός μέρους των μεταβλητών για την πρόβλεψη μιας ή περισσότερων μεταβλητών

• **Περιγραφικοί στόχοι** (π.χ. ομαδοποίηση - clustering, ανακάλυψη κανόνων συσχέτισης association rules, διαδοχική ανακάλυψη προτύπων), που επιτυγχάνονται με τον προσδιορισμό προτύπων που περιγράφουν δεδομένα και που μπορεί να γίνει εύκολα κατανοητό από τον χρήστη.

Από την άλλη πλευρά, δεν πρέπει να θεωρούμε ότι η εξόρυξη δεδομένων μπορεί να λύσει οποιοδήποτε πρόβλημα εστιασμένο στην εύρεση χρήσιμων πληροφοριών στα δεδομένα. Όπως και στην εξόρυξη πραγματικών αντικειμένων , είναι δυνατό για την εξόρυξη δεδομένων να σκάψει το «ορυχείο» των δεδομένων χωρίς τελικά να ανακαλύψει κάτι σε αυτό. Η ανακάλυψη γνώσης/χρήσιμης πληροφορίας εξαρτάται από πολλούς παράγοντες, ξεκινώντας από το «ορυχείο» των δεδομένων και τελειώνοντας με τα χρησιμοποιημένα «εργαλεία» εξόρυξης δεδομένων και την ικανότητα του «miner».

1.3 ΒΗΜΑΤΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Σχηματικά μπορούν να προσδιοριστούν τρία χαρακτηριστικά βήματα της διαδικασίας εξόρυξης δεδομένων σύμφωνα με τον S. Brown(2014) [2]:

1.3.1 ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΚΑΤΑΝΟΗΣΗ

Πριν προχωρήσουμε σε οποιοδήποτε βήμα αφορά την επεξεργασία δεδομένων ή την χρήση εργαλείων χρειάζεται να καθορίσουμε τους λόγους για τους οποίους θα προβούμε σε αυτή τη διαδικασία. Η επιχειρηματική κατανόηση περιλαμβάνει 4 σκέλη.

- Καθορισμό των επιχειρηματικών στόχων

- Αξιολόγηση της κατάστασης

- Στόχοι της εξόρυξης

- Δημιουργία πλάνου

Συχνά το αντικείμενο της εξόρυξης δεν είναι κάτι που έχει ενδιαφέρον επιχειρησιακά, για αυτό είναι σημαντικός ο καθορισμός των επιχειρησιακών στόχων ώστε να επιλεχθούν με βάση αυτό τα δεδομένα και το κατάλληλο μοντέλο.

1.3.2 ΚΑΤΑΝΟΗΣΗ ΚΑΙ ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Το δεύτερο βήμα της εξόρυξης δεδομένων περιλαμβάνει την κατανόηση των δεδομένων προς ανάλυση. Αυτό το βήμα ακολουθεί τους επιχειρησιακούς στόχους και το αντίστοιχο πλάνο που έχει δημιουργηθεί. Παράλληλα όμως αυτή η διαδικασία μπορεί να αναπροσαρμόσει το αρχικό πλάνο λόγω ζητημάτων που έχουν να κάνουν με τα δεδομένα τα οποία επιλέγονται.

Αρχικά πρέπει να γίνει η συλλογή των δεδομένων από τις πηγές δεδομένων οι οποίες έχουν επιλεχθεί. Πριν την εφαρμογή οποιασδήποτε τεχνικής εξόρυξης δεδομένων , είναι απολύτως απαραίτητο να προετοιμαστούν τα ακατέργαστα δεδομένα. Το πρόβλημα που αφορά την ποιότητα των δεδομένων. Έτσι, δουλεύοντας με πρωτογενή δεδομένα μπορούμε να βρούμε θόρυβο, ακραίες τιμές/ανωμαλίες, τιμές που λείπουν, διπλά δεδομένα, εσφαλμένα καταγεγραμμένα δεδομένα, ληγμένα δεδομένα κλπ. Αντίστοιχα, ανάλογα με τα ποιοτικά προβλήματα που εντοπίζονται στα δεδομένα, προχωράμε στην επίλυσή τους με συγκεκριμένες μεθόδους. Για παράδειγμα, στην περίπτωση ύπαρξης θορύβου ,χρησιμοποιούνται διαφορετικές τεχνικές φιλτραρίσματος για την αφαίρεση/μείωση του αποτελέσματος της παραμόρφωσης. Έτσι, σε περίπτωση επεξεργασίας σήματος μπορούμε να αναφέρουμε, εκτός από τα ηλεκτρονικά φίλτρα, τα «μαθηματικά» φίλτρα που αποτελούνται από μαθηματικούς αλγορίθμους που χρησιμοποιούνται για την αλλαγή της αρμονικής συνιστώσας του σήματος (π.χ μέσο φίλτρο, φίλτρο Fourier, κ.λπ.). Σε περίπτωση ακραίων τιμών, δηλαδή τιμών που αποκλίνουν σημαντικά από τη μέση τιμή των δεδομένων, μπορούμε να προχωρήσουμε είτε σε αφαίρεση ή στην εναλλακτική χρήση παραμέτρων (στατιστικών) που δεν είναι τόσο ευαίσθητες σε αυτές τις ακραίες τιμές (π.χ. διάμεσος αντί για μέσος όρος, το οποίο είναι πολύ ευαίσθητο σε ακραίες τιμές). Η περίπτωση των τιμών που λείπουν είναι κοινή στην πρακτική εξόρυξης δεδομένων και έχει πολλές αιτίες. Σε αυτή

την περίπτωση μπορούμε να χρησιμοποιήσουμε διαφορετικές μεθόδους, όπως: εξάλειψη αντικειμένων δεδομένων με τιμές που λείπουν, εκτίμηση τιμών που λείπουν, αντικατάστασή τους με άλλες διαθέσιμες τιμές (π.χ. μέσος/διάμεσος, πιθανώς σταθμισμένος), αγνοώντας κατά την ανάλυση, εάν είναι δυνατόν, κ.λπ. Σε περίπτωση διπλών δεδομένων (π.χ. άτομο με πολλαπλές διευθύνσεις e-mail), μπορεί να εξεταστεί το ενδεχόμενο διαγραφής των διπλότυπων. Αφού λυθεί το θέμα ποιότητας των δεδομένων, προχωράμε στη σωστή προ επεξεργασία τους.

Συνοπτικά παρουσιάζονται οι τεχνικές προ επεξεργασίας δεδομένων [4].

– Η δειγματοληψία είναι η κύρια μέθοδος επιλογής δεδομένων, αντλώντας ένα αντιπροσωπευτικό δείγμα από ολόκληρο το σύνολο δεδομένων.

– Μείωση διαστάσεων.

– Η επιλογή χαρακτηριστικών χρησιμοποιείται για την εξάλειψη άσχετων και περιττών χαρακτηριστικών.

– Η δημιουργία χαρακτηριστικών αναφέρεται στη διαδικασία δημιουργίας νέων (τεχνητών) χαρακτηριστικών, που μπορούν να συλλάβουν καλύτερα σημαντικές πληροφορίες σε δεδομένα από τα αρχικά.

– Διακριτοποίηση και δυαδοποίηση, δηλαδή, εν συντομία, η μετάβαση από συνεχή δεδομένα σε διακριτά (κατηγορικά) δεδομένα (π.χ. μετάβαση από πραγματικές τιμές σε ακέραιες τιμές) και μετατροπή πολλαπλών τιμών σε δυαδικές.

1.3.3 ΚΑΘΟΡΙΣΜΟΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Σε όλη τη διαδικασία εξόρυξης δεδομένων, το στάδιο της επεξεργασίας θα επαναληφθεί όποτε είναι απαραίτητο. Πρώτα απ’ όλα, δεδομένου ότι αντιπροσωπεύει μια διαδικασία ανάλυσης δεδομένων (εξόρυξη των δεδομένων), πρέπει να εστιάσουμε στα δεδομένα. Μόλις επιλεγούν τα δεδομένα προς εξόρυξη, θα πρέπει να αποφασίσουμε πώς να το κάνουμε δείγμα των δεδομένων, αφού συνήθως δεν δουλεύουμε με ολόκληρη τη βάση δεδομένων. Μια σημαντική πτυχή της διαδικασίας εξόρυξης δεδομένων, δηλαδή ο τρόπος επιλογής δεδομένων που θα αναλυθούν. Σημειώστε ότι ολόκληρη η έρευνα θα επηρεαστεί με την επιλεγμένη μεθοδολογία. Στο πλαίσιο αυτό, θα παρουσιαστούν με λίγα λόγια δύο τεχνικές μηχανικής μάθησης που χρησιμοποιούνται εκτενώς στην εξόρυξη δεδομένων, συγκεκριμένα η εποπτευόμενη/μη εποπτευόμενη μάθηση.

Εποπτευόμενη μάθηση σημαίνει τη διαδικασία δημιουργίας μιας αντιστοιχίας (συνάρτησης) χρησιμοποιώντας ένα σύνολο δεδομένων εκπαίδευσης, ως «προηγούμενη εμπειρία» του μοντέλου. Ο σκοπός της εποπτευόμενης μάθησης είναι να προβλέψει την τιμή (έξοδο) της συνάρτησης για οποιοδήποτε νέο αντικείμενο (εισαγωγή) μετά την ολοκλήρωση της εκπαιδευτικής διαδικασίας. Ένα κλασικό παράδειγμα της τεχνικής της εποπτευόμενης μάθησης αντιπροσωπεύεται από τη διαδικασία ταξινόμησης (μέθοδος πρόβλεψης).

Σε αντίθεση με την εποπτευόμενη μάθηση, στην μάθηση χωρίς επίβλεψη το μοντέλο προσαρμόζεται στις παρατηρήσεις, διακρίνεται από το γεγονός ότι δεν υπάρχει εκ των προτέρων έξοδος . Ένα κλασικό παράδειγμα της μάθησης χωρίς επίβλεψη η τεχνική αντιπροσωπεύεται από τη διαδικασία ομαδοποίησης (περιγραφική μέθοδος). Όταν

χρησιμοποιούνται μέθοδοι μάθησης χωρίς επίβλεψη, ο γενικός σκοπός ενός μοντέλου είναι να ομαδοποιήσει παρόμοια αντικείμενα ή να εντοπίσει εξαιρέσεις σε δεδομένα.

– Η ερμηνεία του μοντέλου αναφέρεται στη στιγμή που, μετά την εξόρυξη της βάσης δεδομένων (μελέτη/ανάλυση/ερμηνεία),το μοντέλο εξόρυξης δεδομένων δημιουργήθηκε με βάση την ανάλυση αυτών των δεδομένων, όντας έτοιμο να παράσχει χρήσιμες πληροφορίες για αυτά.

Ουσιαστικά διαβάζοντας τα δεδομένα εμείς κατανοούμε τη διαδικασία πρόσβασης σε δεδομένα (π.χ. εξαγωγή δεδομένων από ένα αρχείο κειμένου και τοποθέτησή τους σε μορφή πίνακα όπου οι γραμμές είναι περιπτώσεις και οι στήλες είναι μεταβλητές, προκειμένου να ομαδοποιηθούν (να ληφθούν παρόμοιες περιπτώσεις, π.χ., συνώνυμα). Μόλις διαβαστούν τα δεδομένα, περνάμε σε κατασκευή του μοντέλου εξόρυξης δεδομένων. Οποιοδήποτε μοντέλο θα εξάγει διάφορους δείκτες από την ποσότητα των διαθέσιμων δεδομένων, χρήσιμων για την κατανόηση των δεδομένων (π.χ., συχνότητες ορισμένων τιμών, βάρη ορισμένων χαρακτηριστικών, συσχετισμένα χαρακτηριστικά (και δεν εξετάζονται χωριστά) που εξηγούν ορισμένες συμπεριφορές κ.λπ.). Πρέπει να λάβουμε υπόψη ορισμένα σημαντικά χαρακτηριστικά:

- Η ακρίβεια του μοντέλου αναφέρεται στην ικανότητα αυτού του μοντέλου να παρέχει σωστές και αξιόπιστες πληροφορίες, όταν χρησιμοποιείται σε πραγματικές καταστάσεις. Η πραγματική ακρίβεια μετριέται σε νέα δεδομένα και όχι σε δεδομένα εκπαίδευσης, όπου το μοντέλο μπορεί να αποδώσει πολύ καλά (δείτε την περίπτωση υπερβολικής τοποθέτησης overfitting).

- Η καταληπτότητα του μοντέλου αναφέρεται στο χαρακτηριστικό του ότι είναι εύκολα κατανοητό από διαφορετικά άτομα με διαφορετικούς βαθμούς/τύπους εκπαίδευσης, ξεκινώντας από τον τρόπο σύνδεσης των εισόδων (δεδομένα που εισάγονται στα «μηχανήματα εξόρυξης») με εκροές (αντίστοιχα συμπεράσματα) και τελειώνοντας με τον τρόπο στην οποία παρουσιάζεται η ακρίβεια της πρόβλεψης.

- Η απόδοση ενός μοντέλου εξόρυξης δεδομένων καθορίζεται τόσο από το χρόνο που απαιτείται για να κατασκευαστεί αλλά και την ταχύτητα επεξεργασίας των δεδομένων προκειμένου να παρέχει μια πρόβλεψη.

- Ο θόρυβος στα δεδομένα είναι ένας «εχθρός» στη δημιουργία ενός αποτελεσματικού μοντέλου εξόρυξης δεδομένων, γιατί δεν μπορεί να αφαιρεθεί πλήρως. Κάθε μοντέλο έχει ένα κατώφλι ανοχής στο θόρυβο και αυτός είναι ένας από τους λόγους για ένα αρχικό στάδιο προ επεξεργασίας δεδομένων.

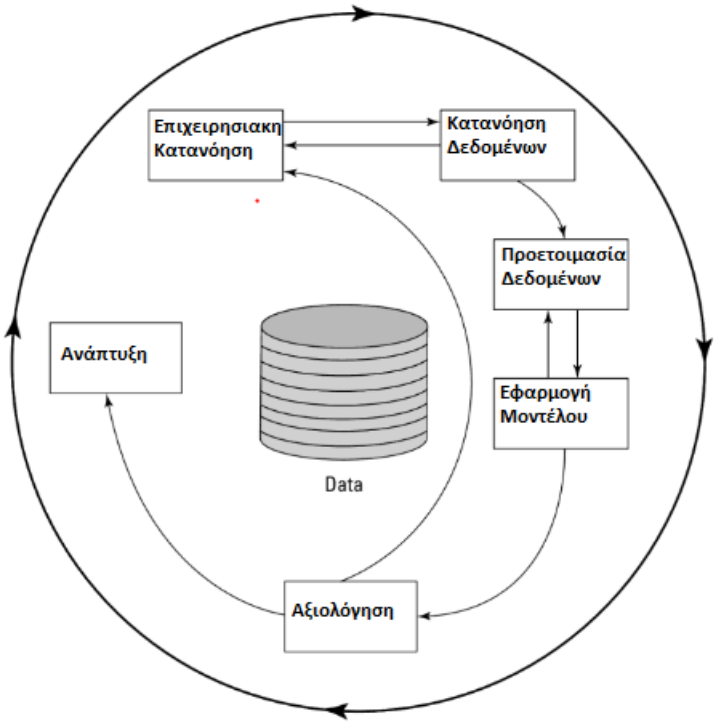
Η κατανόηση του μοντέλου αναφέρεται στη στιγμή που, μετά την εξόρυξη της βάσης δεδομένων (μελέτη/ανάλυση/ερμηνεία), το μοντέλο το οποίο εφαρμόστηκε πάνω στα συγκεκριμένα δεδομένα παρέχει χρήσιμη πληροφορία για αυτά.

1.3.4 ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΕΦΑΡΜΟΓΗ

Η αξιολόγηση δεν αφορά μόνο την αξιολόγηση του μοντέλου που εφαρμόστηκε αλλά το σύνολο της διαδικασίας. Η αξιολόγηση των αποτελεσμάτων που προέκυψαν από την εφαρμογή του μοντέλου αφορά το κατά πόσο εκπλήρωσαν τους αρχικούς στόχους που τέθηκαν στη διαδικασία της εξόρυξης δεδομένων. Είναι σημαντικό να διερευνηθούν όλες οι πλευρές και γι’ αυτό είναι σημαντική η εφαρμογή του ίδιου του μοντέλου σε πραγματικό

χρόνο , κι όχι μόνο σε δεδομένα τεστ. Με βάση τα αποτελέσματα της αξιολόγησης είναι σημαντικό να οριστούν τα επόμενα βήματα , το αν το μοντέλο είναι έτοιμο να εκτελεστεί ή αν χρειάζεται να γίνει επανεξέταση κάποιων βημάτων.

Τελικό βήμα είναι η εφαρμογή του μοντέλου , το κατά πόσο το μοντέλο συμβάλει στην βελτίωση της επιχειρησιακής λειτουργίας, ανεξάρτητα από το πόσο καλά είναι τα αποτελέσματα που δίνει. Παρακάτω (Εικόνα 1) απεικονίζεται η διαδικασία με την λογική βημάτων.



Εικόνα 1 - Βήματα Εξόρυξης Δεδομένων

1.4 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

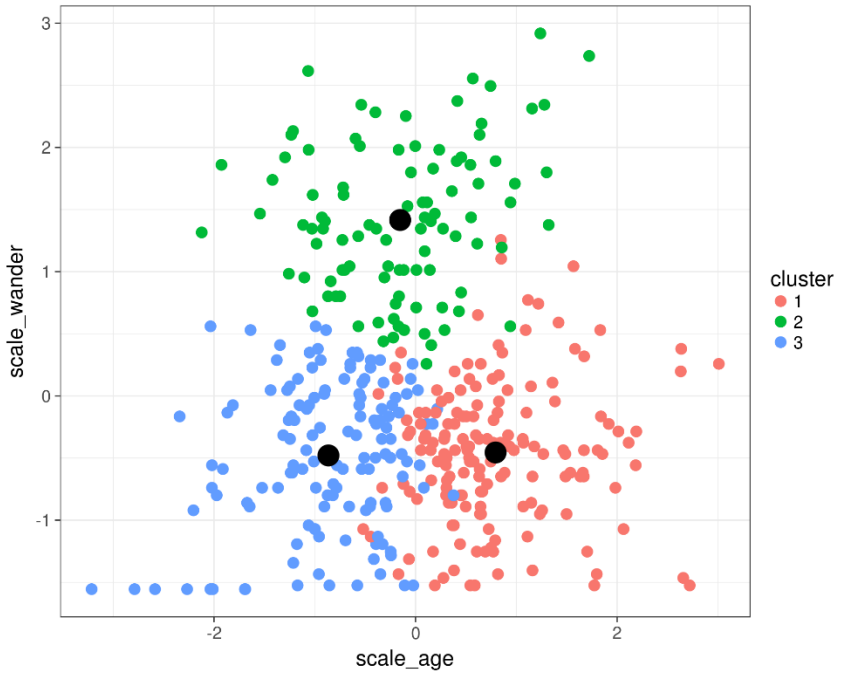
Η εξόρυξη γνώσης επιτυγχάνεται μέσα από ένα ευρύ φάσμα αλγόριθμων οι οποίοι χρησιμοποιούν τεχνικές από διαφορετικούς τομείς όπως είναι η στατιστική, η μηχανική μάθηση και η αναγνώριση προτύπων. Υπάρχει μια πληθώρα υπολογιστικών μεθόδων εξόρυξης γνώσης , οι βασικότερες από τις μεθόδους της εξόρυξης, είναι οι εξής:

- Κατηγοριοποίηση (Classification)
- Συσταδοποίηση (Clustering)
- Ανάλυση Συσχέτισης
- Παλινδρόμηση

1.4.1 ΟΜΑΔΟΠΟΙΗΣΗ – CLUSTERING

Η ομαδοποίηση δεδομένων αναφέρεται στη μέθοδο ομαδοποίησης δεδομένων σε διαφορετικές ομάδες ανάλογα με τα χαρακτηριστικά τους. Αυτή η ομαδοποίηση φέρνει μια τάξη στα δεδομένα και ως εκ τούτου η περαιτέρω επεξεργασία αυτών των δεδομένων γίνεται ευκολότερη. Γενικά, η εξόρυξη δεδομένων είναι η ανάλυση δεδομένων για σχέσεις που δεν έχουν ανακαλυφθεί προηγουμένως. Η ομαδοποίηση είναι μια δημοφιλής τεχνική ανάλυσης δεδομένων και εξόρυξης δεδομένων. Η διαδικασία ομαδοποίησης εξαρτάται από τον τύπο της ομοιότητας που επιλέγεται για την τμηματοποίηση των αντικειμένων. Κατά συνέπεια, μπορούν να χωριστούν με διάφορους τρόπους, λαμβάνοντας υπόψη το είδος της ομοιότητας μεταξύ τους.

Μια δημοφιλής τεχνική για ομαδοποίηση βασίζεται στον αλγόριθμο k-means, έτσι ώστε τα δεδομένα να χωρίζονται σε k συστάδες. Στην εικόνα 2, βλέπουμε με διαφορετικό χρώμα τις διαφορετικές συστάδες και το κέντρο τους απεικονίζεται με μαύρο χρώμα. Σε αυτή τη μέθοδο, ο αριθμός των συστάδων είναι προκαθορισμένος και η τεχνική εξαρτάται σε μεγάλο βαθμό από την αρχική αναγνώριση των στοιχείων που αντιπροσωπεύουν καλά τα συμπλέγματα.



Εικόνα 2 - k-means 1

Η διαδικασία ομαδοποίησης περιλαμβάνει βασικά τρία κύρια βήματα [4]:

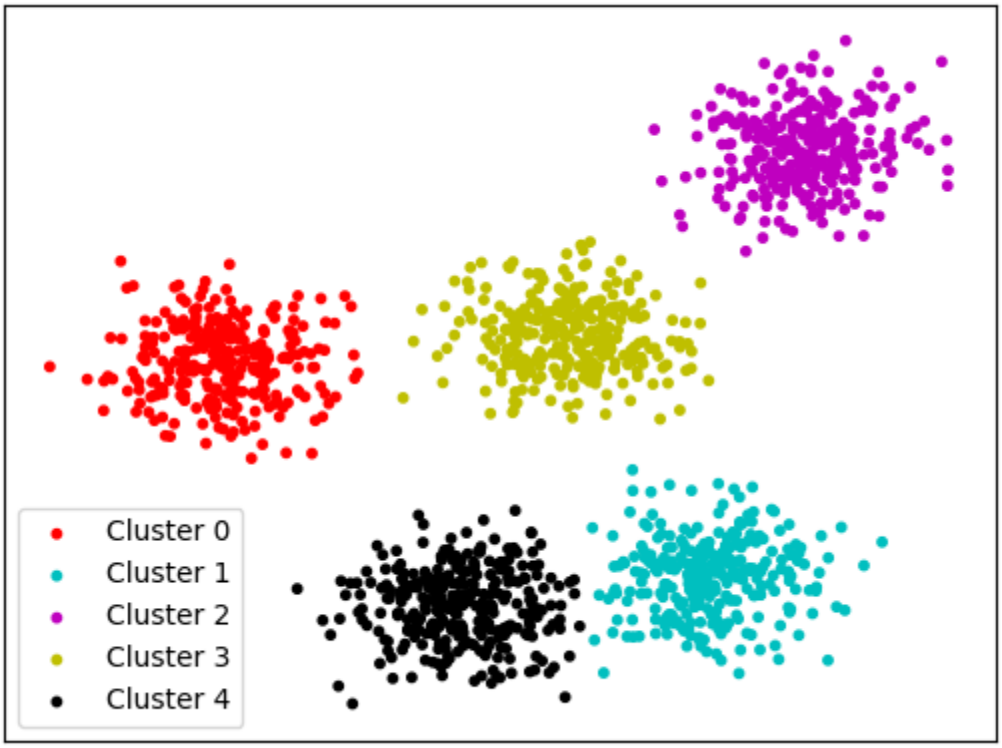
1. Καθορισμός μέτρου ομοιότητας.

2. Καθορισμός κριτηρίου για τη διαδικασία δημιουργίας clusters.

3. Δημιουργία αλγορίθμου για την κατασκευή συστάδων με βάση το επιλεγμένο κριτήριο.

Ένας αλγόριθμος ομαδοποίησης στοχεύει στον εντοπισμό φυσικών ομάδων αντικειμένων σε ένα δεδομένο σύνολο και ως εκ τούτου, πρέπει να μετρήσει τον βαθμό ομοιότητας μεταξύ των αντικειμένων, με βάση ένα συγκεκριμένο κριτήριο. Επομένως, η πρώτη ενέργεια που

πρέπει να ληφθεί είναι να εξεταστεί ένα κατάλληλο μέτρο, που αντιστοιχεί στην εγγενή φύση των δεδομένων και προορίζονται για την αξιολόγηση ορισμένων «αποστάσεων» (ανομοιότητα) μεταξύ αντικειμένων. Στην εικόνα 3 μπορούμε να διακρίνουμε τις διαφορετικές συστάδες με διαφορετικά χρώματα.



Εικόνα 3 - Συσταδοποίηση 2

1.4.2 ΠΑΛΙΝΔΡΟΜΗΣΗ

Η ανάλυση γραμμικής παλινδρόμησης χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής με βάση την τιμή μιας άλλης μεταβλητής. Η μεταβλητή που θέλετε να προβλέψετε ονομάζεται εξαρτημένη μεταβλητή. Η μεταβλητή που χρησιμοποιείτε για να προβλέψετε την τιμή της άλλης μεταβλητής ονομάζεται ανεξάρτητη μεταβλητή. Η γραμμική παλινδρόμηση ταιριάζει σε μια ευθεία γραμμή ή επιφάνεια που ελαχιστοποιεί τις αποκλίσεις μεταξύ των προβλεπόμενων και των πραγματικών τιμών εξόδου. [17]

Στην εικόνα 4 δίνεται η ευθεία της γραμμικής παλινδρόμησης.

Η ανάλυση παλινδρόμησης καθώς και η συσχέτιση έχουν την προέλευσή τους στο έργο του διάσημου γενετιστή Sir Francis Galton (1822-1911), που ξεκίνησε στο τέλος του δέκατου ένατου αιώνα την έννοια της «οπισθοδρόμησης προς το μέσο» σύμφωνα στον οποίο, δεδομένων δύο εξαρτημένων μετρήσεων, η εκτιμώμενη τιμή για τη δεύτερη μέτρηση είναι πιο κοντά στη μέση τιμή από την παρατηρούμενη τιμή της πρώτης μέτρησης (π.χ., οι ψηλότεροι μπαμπάδες έχουν μικρότερα παιδιά και, αντίθετα, πιο κοντοί πατέρες έχουν πιο ψηλά παιδιά - το ύψος των παιδιών υποχωρεί στο μέσο ύψος).

Στη Στατιστική, η ανάλυση παλινδρόμησης σημαίνει το μαθηματικό μοντέλο που καθιερώνει τη σύνδεση μεταξύ των τιμών μιας δεδομένη μεταβλητή

(απόκριση/αποτέλεσμα/εξαρτημένη μεταβλητή) και τις τιμές άλλων μεταβλητών (πρόβλεψη/ανεξάρτητες μεταβλητές). Το πιο γνωστό παράδειγμα παλινδρόμησης είναι ίσως η αναγνώριση της σχέσης μεταξύ του ύψους και του βάρους ενός ατόμου, εμφανίζεται σε πίνακες που λαμβάνονται με τη χρήση της εξίσωσης παλινδρόμησης, αξιολογώντας έτσι το ιδανικό βάρος για συγκεκριμένο ύψος. Η ανάλυση παλινδρόμησης σχετίζεται καταρχήν με [4]:

• Προσδιορισμός μιας ποσοτικής σχέσης μεταξύ πολλαπλών μεταβλητών.

• Πρόβλεψη των τιμών μιας μεταβλητής σύμφωνα με τις τιμές άλλων μεταβλητών (καθορισμός της επίδρασης των «μεταβλητών πρόβλεψης» στη «μεταβλητή απόκρισης»).

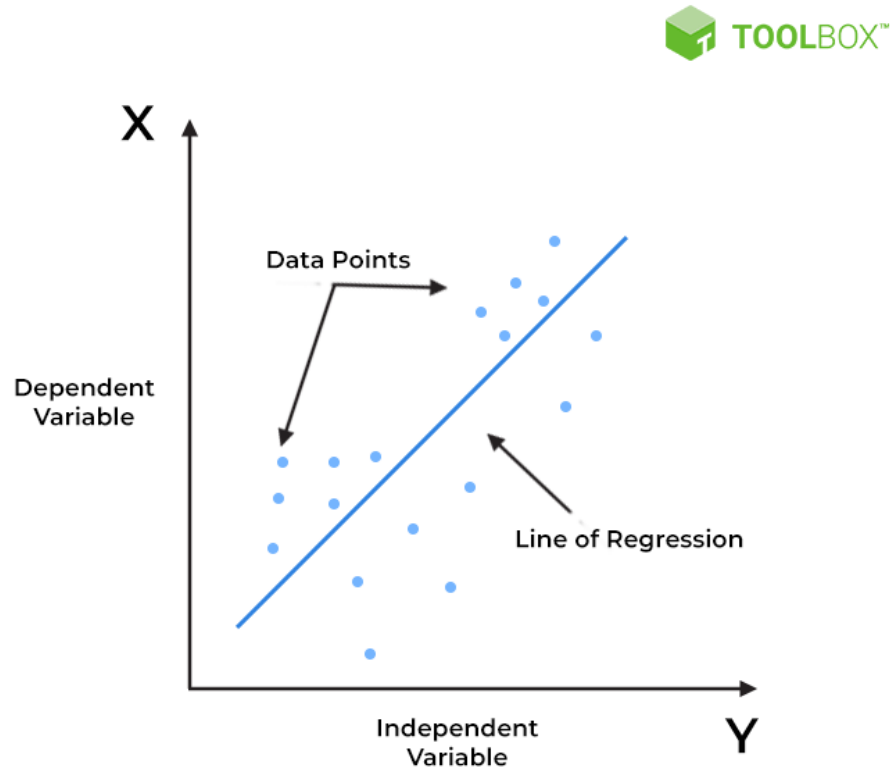
Οι εφαρμογές αυτής της στατιστικής μεθόδου στην εξόρυξη δεδομένων είναι πολλαπλές, αναφέρουμε εδώ τις ακόλουθες:

• Εμπόριο: πρόβλεψη ποσών πωλήσεων νέου προϊόντος με βάση τη διαφήμιση δαπάνη.

• Μετεωρολογία: πρόβλεψη ταχυτήτων και κατευθύνσεων του ανέμου σε συνάρτηση με τη θερμοκρασία, υγρασία, πίεση αέρα κ.λπ.

• Χρηματιστήριο: πρόβλεψη χρονοσειρών χρηματιστηριακών δεικτών (εκτίμηση τάσεων).

• Ιατρική: επίδραση του γονικού βάρους/ύψους γέννησης στο βάρος/ύψος γέννησης βρέφους, για παράδειγμα.



Εικόνα 4 - Γραμμική Παλινδρόμηση

1.4.3 ΤΑΞΙΝΟΜΗΣΗ

Η ιδέα ότι ο ανθρώπινος νους οργανώνει τη γνώση του χρησιμοποιώντας τη φυσική διαδικασία της ταξινόμηση είναι ευρέως διαδεδομένη. Η ταξινομία εμφανίστηκε πρώτα ως η επιστήμη της ταξινόμησης των ζωντανών οργανισμών (άλφα ταξινόμηση), αλλά στη συνέχεια αναπτύχθηκε ως επιστήμη της ταξινόμησης γενικά, συμπεριλαμβανομένων εδώ των αρχών της ταξινόμησης επίσης. Έτσι, η ταξινόμηση είναι η διαδικασία της τοποθέτησης ενός συγκεκριμένου αντικειμένου (έννοιας) σε ένα σύνολο κατηγοριών, με βάση τις αντίστοιχες ιδιότητες αντικειμένου (έννοιας).

Η διαδικασία ταξινόμησης βασίζεται σε τέσσερα θεμελιώδη στοιχεία [4] :

• Κλάση -η εξαρτημένη μεταβλητή του μοντέλου- που είναι μια κατηγορική μεταβλητή που αναπαριστά η «ετικέτα» που τοποθετείται στο αντικείμενο μετά την ταξινόμησή του. Παραδείγματα τέτοιων κατηγοριών είναι: παρουσία εμφράγματος του μυοκαρδίου, αφοσίωση πελατών, κατηγορία αστεριών (γαλαξίες), κατηγορία σεισμού (τυφώνας) κ.λπ.

• Πρόβλεψη -οι ανεξάρτητες μεταβλητές του μοντέλου- που αντιπροσωπεύονται από τα χαρακτηριστικά των προς ταξινόμηση δεδομένων και βάσει ποιας ταξινόμησης είναι φτιαγμένο. Παραδείγματα τέτοιων προγνωστικών είναι: κάπνισμα, κατανάλωση αλκοόλ, αίμα, πίεση, συχνότητα αγοράς, οικογενειακή κατάσταση, χαρακτηριστικά (δορυφορικών) εικόνων, συγκεκριμένα γεωλογικά αρχεία, κατεύθυνση ανέμου και ταχύτητας, εποχή, τοποθεσία εμφάνιση φαινομένου κ.λπ.

• Δεδομένα εκπαίδευσης -που είναι το σύνολο δεδομένων που περιέχει τιμές για τα δύο προηγούμενα εξαρτήματα και χρησιμοποιείται για την «εκπαίδευση» του μοντέλου ώστε να αναγνωρίζει την κατάλληλη τάξη, με βάση τους διαθέσιμους προγνωστικούς παράγοντες. Παραδείγματα τέτοιων συνόλων είναι: ομάδες ασθενών δοκιμασμένες σε εμφράγματα, ομάδες πελατών σούπερ μάρκετ, βάσεις δεδομένων που περιέχουν εικόνες για τηλεσκοπική παρακολούθηση και παρακολούθηση αστρονομικά αντικείμενα (π.χ. Παρατηρητήριο Palomar (Caltech).

• Δοκιμαστικό σύνολο δεδομένων, που περιέχει νέα δεδομένα που θα ταξινομηθούν από το (ταξινομητή) μοντέλο που κατασκευάστηκε παραπάνω και η ακρίβεια ταξινόμησης (απόδοση μοντέλου) μπορεί έτσι να αξιολογηθεί.

Ο σκοπός της εποπτευόμενης μάθησης είναι η πρόβλεψη της τιμής (έξοδος) της συνάρτησης για οποιοδήποτε νέο αντικείμενο/δείγμα (είσοδος) μετά την ολοκλήρωση της εκπαιδευτικής διαδικασίας. Η τεχνική ταξινόμησης, ως προγνωστική μέθοδος, είναι ένα τέτοιο παράδειγμα τεχνικής εποπτευόμενης μηχανικής μάθησης, με την προϋπόθεση ότι η ύπαρξη μιας ομάδας επισημασμένων παρουσιών για κάθε κατηγορία αντικειμένων.

Συνοψίζοντας, μια διαδικασία ταξινόμησης χαρακτηρίζεται από [4]:

• Εισαγωγή: ένα σύνολο δεδομένων εκπαίδευσης που περιέχει αντικείμενα με χαρακτηριστικά, ένα από τα οποία είναι η ετικέτα τάξης

• Έξοδος: ένα μοντέλο που εκχωρεί μια συγκεκριμένη ετικέτα για κάθε αντικείμενο (ταξινομεί το αντικείμενο σε μια κατηγορία), με βάση τα άλλα χαρακτηριστικά.

• Ο ταξινομητής χρησιμοποιείται για την πρόβλεψη της κλάσης νέων, άγνωστων αντικειμένων. Μια δοκιμή του συνόλου δεδομένων χρησιμοποιείται επίσης για τον προσδιορισμό της ακρίβειας του μοντέλου.

Υπάρχουν προηγμένες τεχνικές που χρησιμοποιούνται στην εξόρυξη δεδομένων, τόσο στην ταξινόμηση όσο και σε άλλους τομείς της αυτόματης εξερεύνησης δεδομένων, γνωστές μέθοδοι, όπως:

• Μπεϋζιανός ταξινομητής/Naive Bayes.

• Νευρωνικά δίκτυα.

• Μηχανές φορέα υποστήριξης.

• Εξόρυξη κανόνων σύνδεσης.

• Ταξινόμηση βάσει κανόνων.

• k-πλησιέστερος γείτονας.

• Τραχιά σετ.

• Αλγόριθμοι ομαδοποίησης.

• Γενετικοί αλγόριθμοι.

1.4.3.1 ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

Κατ 'αρχήν, τα δέντρα απόφασης χρησιμοποιούνται για την πρόβλεψη της συμμετοχής αντικειμένων σε διαφορετικές κατηγορίες (τάξεις), λαμβάνοντας υπόψη τις τιμές που αντιστοιχούν στα χαρακτηριστικά τους (προγνωστικές μεταβλητές). Η μέθοδος του δέντρου αποφάσεων είναι μία από τις κύριες τεχνικές εξόρυξης δεδομένων. Η ευελιξία αυτής της τεχνικής το καθιστά ιδιαίτερα ελκυστικό, ειδικά επειδή παρουσιάζει το πλεονέκτημα μιας πολύ υποβλητικής οπτικοποίησης (ένα «δέντρο» που συνοψίζει συνθετικά τη ταξινόμηση). Ωστόσο, πρέπει να τονιστεί ότι αυτή η τεχνική πρέπει απαραίτητα να επιβεβαιώνεται με άλλες παραδοσιακές τεχνικές, ειδικά κατά την εργασία ελέγχου υποθέσεων (π.χ. υποθέσεις σχετικά με τη διανομή δεδομένων). Παρ 'όλα αυτά, ως πειραματική διερευνητική τεχνική, ειδικά όταν οι παραδοσιακές μέθοδοι δεν είναι διαθέσιμες, τα δέντρα αποφάσεων μπορούν να χρησιμοποιηθούν με επιτυχία, προτιμώνται από άλλα μοντέλα ταξινόμησης [4].

1. Δέντρα ταξινόμησης, όρος που χρησιμοποιείται όταν το αποτέλεσμα πρόβλεψης είναι η συμμετοχή στην τάξη των δεδομένων

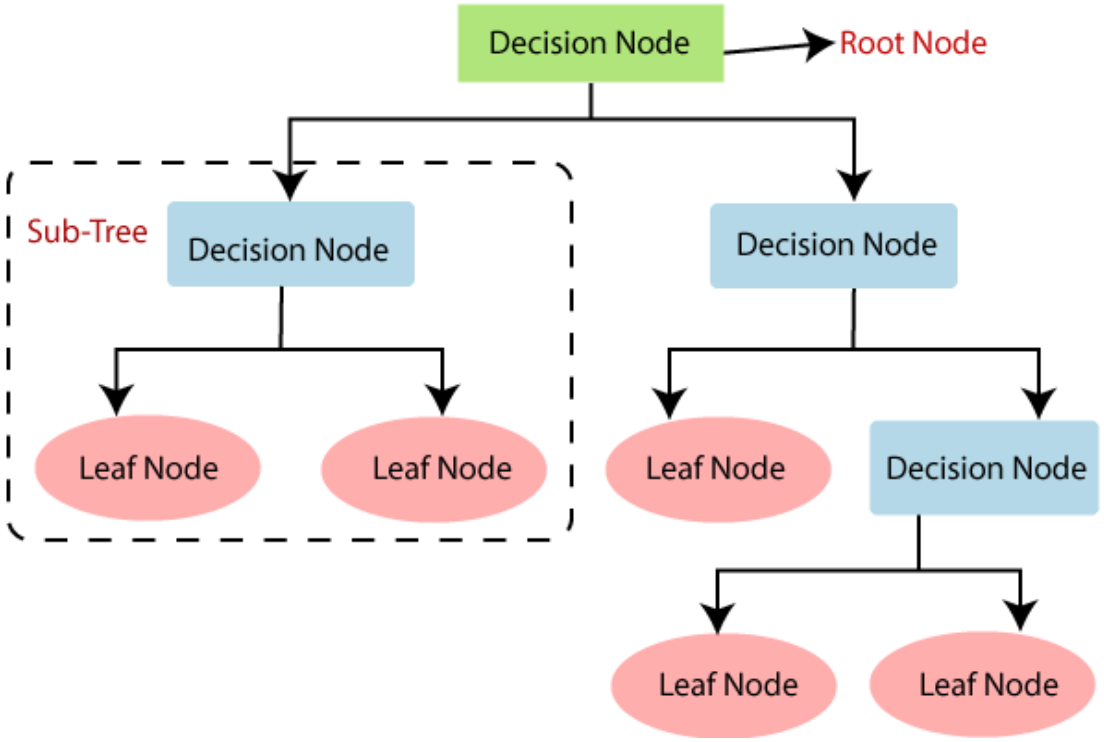
2. Δέντρα παλινδρόμησης, όταν το προβλεπόμενο αποτέλεσμα μπορεί να θεωρηθεί ως πραγματικός αριθμός (π.χ. τιμή πετρελαίου, αξία κατοικίας, τιμή μετοχής κ.λπ.)

3. CART (ή C&RT), π.χ. Classification And Regression Tree, όταν λαμβάνουμε υπόψη και τις δύο παραπάνω περιπτώσεις.

Ένα δέντρο αποφάσεων είναι μια ιεραρχική δομή που αντιπροσωπεύει ένα μοντέλο ταξινόμησης. Οι κόμβοι αντιστοιχούν σε διαχωρισμούς που εφαρμόζονται για την αποσύνθεση του τομέα σε περιοχές και τερματικούς κόμβους, οι ετικέτες κλάσης αντιστοιχούν σε περιοχές που είναι αρκετά μικρές ή αρκετά ομοιόμορφες. Για ευκολία, θα δεσμεύσουμε τον όρο κόμβος μόνο στους εσωτερικούς κόμβους και θα αναφερόμαστε σε

τερματικούς κόμβους ως φύλλα. Στην εικόνα 5 παρουσιάζεται ένα δένδρο απόφασης όπως απεικονίζεται γραφικά, ξεκινώντας από την ρίζα του δένδρου προς τους κόμβους και καταλήγοντας στα φύλλα.

Οι διαχωρισμοί καθορίζονται από ορισμένες σχεσιακές συνθήκες που βασίζονται σε επιλεγμένα χαρακτηριστικά που μπορεί να έχουν δύο ή περισσότερα αποτελέσματα. Τυπικά, μια διαίρεση μπορεί να αναπαρασταθεί από μια δοκιμαστική συνάρτηση t : X → Rt που αντιστοιχίζει περιπτώσεις σε διαχωρισμένα αποτελέσματα. Ένας ξεχωριστός εξερχόμενος κλάδος συσχετίζεται με το καθένα πιθανό αποτέλεσμα της διάσπασης ενός κόμβου. Η σχέση μεταξύ του γονικού κόμβου και του απογόνου του κόμβοι, που αντιπροσωπεύονται εννοιολογικά από τους κλάδους που συνδέουν τον πρώτο με τον δεύτερο, δεν το κάνει πρέπει πάντα να αναπαρίστανται ρητά στη δομή δεδομένων του δέντρου αποφάσεων. Ειδικότερα, όταν χρησιμοποιούνται δυαδικοί διαχωρισμοί, η σχέση μπορεί να αναπαρασταθεί σιωπηρά από έναν κατάλληλο κόμβο σχήμα αρίθμησης, π.χ., οι απόγονοι του κόμβου με αριθμό k μπορούν να αριθμηθούν 2k και 2k + 1. [4]



Εικόνα 5 - Δένδρο Απόφασης

1.4.3.2 ΤΕΧΝΗΤΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Πέρα από τις μεθόδους ταξινόμησης που βασίζονται στα δέντρα και τους κανόνες απόφασης, τα τεχνητά νευρωνικά δίκτυα είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης.

Συγκεκριμένα, είναι μια δομή που αποτελείται από ένα δίκτυο νευρώνων οι οποίοι συνδέονται μεταξύ τους και αποτελούν τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Η πιο διαδεδομένη κατηγορία

νευρωνικών δικτύων είναι τα λεγόμενα δίκτυα πρόσθιας τροφοδότησης, τα οποία επιτρέπουν την κίνηση των δεδομένων μόνο προς μια κατεύθυνση, δηλαδή από μια είσοδο προς μια έξοδο και έχουμε και τα δίκτυα που σχηματίζουν κυκλικές δομές τα οποία ονομάζονται ανατροφοδοτούμενα νευρωνικά δίκτυα.

Τα νευρωνικά δίκτυα είναι μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών. Οι μέθοδοι αυτοί είναι αποτελέσματα ακαδημαϊκών ερευνών με στόχο την μοντελοποίηση συστημάτων μάθησης. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να αντιμετωπιστεί ως ένας ειδικός για την κατηγορία της πληροφορίας που του δόθηκε να αναλύσει. Έτσι μπορεί να χρησιμοποιηθεί για να κάνει κάποιες προβλέψεις, όταν προκύψουν κάποιες νέες περιπτώσεις. Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μία καλά ορισμένη σειρά.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα όπως για παράδειγμα η σταδιακή προσέγγιση μίας συνάρτησης. Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης μιας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου, σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως παγώνουν στις κατάλληλες τιμές και έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης. Αυτό σημαίνει ότι πρέπει να δίνει ορθές εξόδους για εισόδους καινοφανείς και διαφορετικές από αυτές με τις οποίες εκπαιδεύτηκε.

Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες:

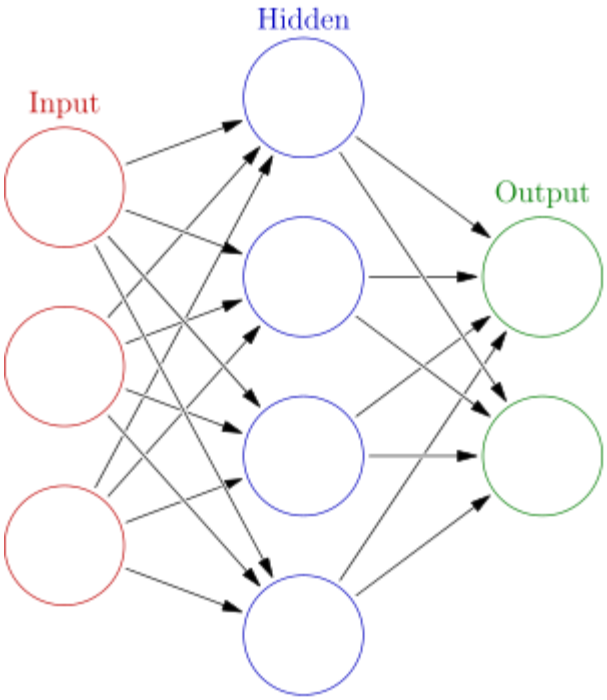
1) Τους νευρώνες εισόδου (input neurons): οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία

2) Τους νευρώνες εξόδου (output neurons): στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας

3) Τους ενδιάμεσους νευρώνες: οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και κρυφοί νευρώνες (hidden neurons).

Στην εικόνα 6 απεικονίζεται ένα νευρωνικό δίκτυο όπως χωρίζεται σε νευρώνες εισόδου, νευρώνες εξόδου και ενδιάμεσους νευρώνες.

Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου.



Εικόνα 6 - Νευρωνικό Δίκτυο

1.4.4 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

Η εκμάθηση κανόνων συσχέτισης είναι μια πολύ γνωστή μέθοδος στην εξόρυξη δεδομένων για ανακάλυψη ενδιαφερουσών σχέσεων μεταξύ μεταβλητών σε μεγάλες βάσεις δεδομένων. Ένας κανόνας συσχέτισης μπορεί να θεωρηθεί ως υπαινιγμός της μορφής X →Y, όπου X και το Y είναι διακριτά στοιχεία ή σύνολα στοιχείων (συλλογές ενός ή περισσότερων αντικειμένων), με το X να είναι ο κανόνας προϋπόθεση και Y είναι ο κανόνας συνέπεια. Είναι μια τεχνική εξόρυξης δεδομένων χωρίς επίβλεψη που αναζητά συνδέσεις μεταξύ στοιχείων/εγγραφών που ανήκουν σε ένα μεγάλο σύνολο δεδομένων. Ένα τυπικό και ευρέως χρησιμοποιούμενο παράδειγμα εξόρυξης κανόνων συσχέτισης είναι η ανάλυση καλαθιού αγοράς.

Το μοντέλο δεδομένων του καλαθιού αγοράς χρησιμοποιείται για να περιγράψει μια σχέση πολλά προς πολλά μεταξύ δύο ειδών αντικειμένων. Από τη μια έχουμε είδη (items), και από την άλλη έχουμε καλάθια (baskets), που μερικές φορές ονομάζονται "συναλλαγές". Κάθε καλάθι αποτελείται από ένα σύνολο αντικειμένων (ένα σύνολο στοιχείων) ,ο αριθμός των αντικειμένων σε ένα καλάθι είναι μικρός – πολύ μικρότερος από τον συνολικό αριθμό των ειδών. Ο αριθμός των καλαθιών συνήθως θεωρείται ότι είναι πολύ μεγάλος, μεγαλύτερος από αυτό που χωράει στην κύρια μνήμη. Τα δεδομένα θεωρείται ότι αντιπροσωπεύονται σε ένα αρχείο που αποτελείται από μια σειρά από καλάθια.[3]

Η ανάλυση καλαθιού αγοράς προσπαθεί να εντοπίσει πελάτες, αγοράζοντας ορισμένα ομαδοποιημένα είδη, παρέχοντας πληροφορίες στον συνδυασμό προϊόντων στο «καλάθι» ενός πελάτη. Για παράδειγμα, στο λιανικό εμπόριο, η ανάλυση καλαθιού αγοράς βοηθά τους λιανοπωλητές να κατανοήσουν την αγοραστική συμπεριφορά πελατών. Από την άλλη πλευρά, χρησιμοποιώντας τις πληροφορίες σαρωτή bar-code, ένα σούπερ μάρκετ, η βάση δεδομένων αποτελείται από ένα μεγάλο αριθμό εγγραφών συναλλαγών, όπου αναφέρονται

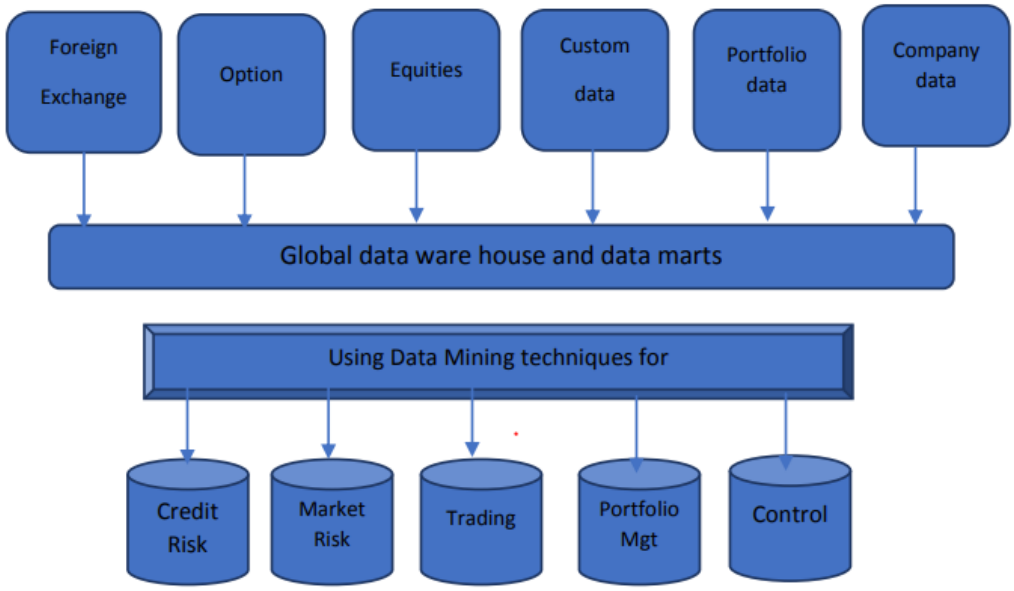
όλα τα στοιχεία που αγοράστηκαν από έναν πελάτη σε μία συναλλαγή αγοράς. Με βάση τον κανόνα συσχέτισης εξόρυξης, οι διαχειριστές θα μπορούσαν να χρησιμοποιήσουν τους κανόνες που ανακαλύφθηκαν σε μια τέτοια βάση δεδομένων για προσαρμογή σχεδιαγραμμάτων καταστημάτων, για διασταυρούμενες πωλήσεις, για προσφορές, για σχεδιασμό καταλόγου, για ταυτοποίηση τμήματα πελατών με βάση το μοτίβο αγορών τους. Ως κλασικό παράδειγμα, εξάλλου η περίφημη σχέση «μπύρας-πάνας» [4]:

{X →Y}⇔{πάνα, γάλα}→{μπύρα}

Για να επιλέξουμε ενδιαφέροντες κανόνες από το σύνολο όλων των πιθανών κανόνων, χρειαζόμαστε κάποια «μέτρα» αξιολόγησης της αποτελεσματικότητας της διαδικασίας κανόνων σύνδεσης: support, confidence και lift. Έτσι, η υποστήριξη αντιπροσωπεύει τον αριθμό των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία στα προηγούμενα και τα επακόλουθα μέρη του κανόνα (συναλλαγές που περιέχουν και τα δύο X και Y), μερικές φορές εκφράζεται ως ποσοστό. Η εμπιστοσύνη αντιπροσωπεύει την αναλογία του αριθμού των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία του Y καθώς και του X στο πλήθος των συναλλαγών που περιλαμβάνουν όλα τα στοιχεία του X. Τέλος, το lift αντιπροσωπεύει την αναλογία του κανόνα εμπιστοσύνης και του αναμενόμενου κανόνα εμπιστοσύνης.

2 ΕΦΑΡΜΟΓΗ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΤΡΑΠΕΖΙΚΟ ΤΟΜΕΑ ΚΑΙ ΤΑ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ

Ο τραπεζικός κλάδος αναγνωρίζει τη σημασία των πληροφοριών που έχει για τους πελάτες της. Αυτή η βιομηχανία απαιτεί και χρησιμοποιεί την τεχνολογία της πληροφορίας όχι μόνο για τη βελτίωση της ποιότητας των υπηρεσιών, αλλά και να αποκτήσει ανταγωνιστικό πλεονέκτημα .Ο τεράστιος όγκος δεδομένων που οι τράπεζες παράγουν όλα αυτά τα χρόνια μπορούν κι επηρεάζουν σε μεγάλο βαθμό την επιτυχία των προσπαθειών εξόρυξης δεδομένων. Χρησιμοποιώντας δεδομένα εξόρυξης για την ανάλυση προτύπων και τάσεων, στελέχη τραπεζών μπορούν να προβλέψουν με αυξημένη ακρίβεια πώς θα αντιδράσουν οι πελάτες στις προσαρμογές των τιμών, την αντίδραση των πελατών ώστε να δέχονται νέες προσφορές προϊόντων, προβλέψεις για τους πελάτες που διατρέχουν μεγαλύτερο κίνδυνο αθέτησης δανείου, και πώς να καταστήσουν κάθε πελατειακή σχέση πιο κερδοφόρα. Η εξόρυξη δεδομένων αποδεικνύεται πολύ χρήσιμο εργαλείο στον τραπεζικό κλάδο. Παρακάτω δίνονται παραδείγματα για το πώς ο τραπεζικός κλάδος αξιοποιεί αποτελεσματικά την εξόρυξη δεδομένων στους τομείς μάρκετινγκ, διαχείριση κινδύνου, εντοπισμός απάτης και απόκτηση και διατήρηση πελατών. [12]



Εικόνα 7 - Τεχνικές Εξόρυξης Δεδομένων

2.1 ΤΟΜΕΙΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΓΙΑ ΤΟΝ ΧΡΗΜΑΤΟΠΙΣΤΩΤΙΚΟ ΚΛΑΔΟ

2.1.1 MARKETING

Ένας από τους πιο ευρέως χρησιμοποιούμενους τομείς της εξόρυξης δεδομένων για τον τραπεζικό κλάδο βρίσκεται στο μάρκετινγκ. Το τμήμα μάρκετινγκ της τράπεζας μπορεί να χρησιμοποιήσει δεδομένα εξόρυξης για την ανάλυση βάσεων δεδομένων πελατών και την ανάπτυξη προφίλ ατόμων με τις προτιμήσεις των πελατών για προϊόντα και υπηρεσίες.

Προσφέροντας μόνο αυτά τα προϊόντα και τις υπηρεσίες ο πελάτης προσελκύεται, η τράπεζα αποταμιεύει χρήματα για προσφορές που διαφορετικά θα ήταν ασύμφορες. Οι τραπεζικοί υπάλληλοι μάρκετινγκ πρέπει να επικεντρωθούν στους πελάτες τους μαθαίνοντας περισσότερα για αυτούς. Με την αποκάλυψη των μοτίβων της συμπεριφοράς του πελάτη, μπορεί να προσδιοριστεί η κερδοφορία τους και η τράπεζα μπορεί επίσης να επεκτείνει τις δραστηριότητές της προσφέροντας σε κάθε μεμονωμένο πελάτη διαφορετικά προϊόντα και υπηρεσίες. Το cross selling είναι ένας τομέας μάρκετινγκ όπου τα δεδομένα μπορούν να χρησιμοποιηθούν για εξόρυξη. Cross selling είναι όταν μια υπηρεσία καθίσταται ελκυστική για έναν πελάτη προκειμένου να αγοραστούν πρόσθετα προϊόντα ή υπηρεσίες από μια επιχείρηση. Όσο περισσότερα προϊόντα και υπηρεσίες μπορεί να παρέχει μια τράπεζα σε πελάτες, τόσο πιο πιθανό είναι η τράπεζα να διατηρήσει αυτούς τους πελάτες [12].

2.1.2 ΔΙΑΧΕΙΡΙΣΗ ΚΙΝΔΥΝΟΥ

Ο πιστωτικός κίνδυνος είναι μια πιθανή ζημία που προκαλείται από την αδυναμία του οφειλέτη να ανταποκριθεί στις υποχρεώσεις αποπληρωμής του χρέους χρέος κεφαλαίου ή τόκων ή και τα δύο. Η ταξινόμηση του πιστωτικού κινδύνου στον χρηματοπιστωτικό τομέα έχει ουσιαστικό ρόλο στη χαρτογράφηση του κίνδυνου καταναλωτή. Η λανθασμένη ταξινόμηση προκαλεί αλυσιδωτά αποτελέσματα όπως η εμφάνιση κακής πίστωσης, διαταραχή της χρηματοπιστωτικής σταθερότητας, που οδηγούν σε τραπεζικές απώλειες. Ταξινόμηση σε κατηγορίες πιστωτικού κινδύνου το πελατειακό δάνειο σε δύο τύπους, καλοπληρωτές ή κακοπληρωτές. Ο στόχος αυτής της έρευνας είναι να ταξινομήσει τον κίνδυνο του καταναλωτή για να ελαχιστοποιήσει τον κίνδυνο χρεοκοπίας. Τις προηγούμενες δεκαετίες, το πιστωτικό σκορ χρησιμοποιώντας παραμετρικές τεχνικές έχει εφαρμοστεί στον χρηματοοικονομικό τομέα, δηλαδή με την Διακριτική Ανάλυση και Δυαδική Λογιστική παλινδρόμηση. Τις τελευταίες δύο δεκαετίες, οι μη παραμετρικές προσεγγίσεις μηχανικής μάθησης, όπως το Νευρωνικό Δίκτυο. Πρόσφατα, η εποχή του Deep Learning έχει μελετηθεί ευρέως στη βαθμολογία πιστώσεων, όπως το Deep Neural Network. Αυτή η μελέτη συγκρίνει την απόδοση πολλών μεθόδων μη παραμετρικής μηχανικής μάθησης και παραμετρικών στατιστικών για την ταξινόμηση των πελατών δάνεια. Η καλύτερη μέθοδος για την ταξινόμηση των δανείων πελατών είναι το DNN με τον αριθμό των νευρώνων σε δοκιμή δεδομένων.

Η διαχείριση κινδύνων καλύπτει όχι μόνο κινδύνους που αφορούν ασφάλειες, αλλά και επιχειρηματικούς κινδύνους από ανταγωνιστική απειλή, κακή ποιότητα προϊόντων και διαρροή πελατών. Η απώλεια πελατών, είναι ένα αυξανόμενο πρόβλημα και η εξόρυξη δεδομένων χρησιμοποιείται στη χρηματοδότηση, το λιανικό εμπόριο και τις βιομηχανίες τηλεπικοινωνιών για να βοηθήσουν στην πρόβλεψη και τις πιθανές απώλειες πελατών. Η απώλεια πελατών από τους ανταγωνιστές είναι σημαντικό μέλημα για τις βιομηχανίες σήμερα, με τον αυξανόμενο όγκο ανταγωνισμού που είναι αντιμέτωπες οι επιχειρήσεις. Επομένως, πρέπει να βρεθούν μέθοδοι που καθορίσει τον αριθμό των πελατών που είναι πιθανό να χαθεί από τους ανταγωνιστές, έτσι ώστε μια επιχείρηση να μπορεί να είναι καλύτερα προετοιμασμένη. Μία προσέγγιση που μπορεί να χρησιμοποιηθεί είναι να δημιουργηθεί ένα μοντέλο πελατών που είναι πιθανό να φύγουν και να πάνε σε μια ανταγωνιστική εταιρεία. Μια ανάλυση πελατών που έχουν φύγει πρόσφατα μπορεί συχνά να μην είναι διαισθητικό μοτίβα, όπως μετά την αλλαγή ενός πελάτη διεύθυνση ή πρόσφατη παρατεταμένη ανταλλαγή με έναν από τους πράκτορες της εταιρείας.

Τα τελευταία χρόνια, ο σύγχρονος τομέας έρευνας είναι η ανάπτυξη του ποσοτικού εμπορίου εφαρμόζοντας την τεχνική εξόρυξης δεδομένων που προσεγγίζεται από την ιστορική πληροφορία ως μεταβλητή εισόδου προβλέπουν τα επιτόκια, το νόμισμα, το ποσοστό πληθωρισμού, τον χρηματιστηριακό δείκτη κ.λπ. στη βραχυπρόθεσμη περίοδο. Το κίνητρο αυτής της τεχνικής είναι να καθορίσει τη χρονική περίοδο κατά την οποία οι αγορές είναι λιγότερο ασταθείς ή φθηνές, ανιχνεύοντας τους σημαντικούς παράγοντες που συμβάλλουν στις αποδόσεις της αγοράς. Όταν υπάρχει υποψία για υποτίμηση ή υπερτίμηση, η πλατφόρμα συναλλαγών αναγνωρίζει τη σχέση μεταξύ του χρηματοοικονομικού περιουσιακού στοιχείου και τις κατάλληλες πληροφορίες και τονίζει τη χρονική περίοδο αγοράς ή πώλησης. Ως περαιτέρω γνώμη, οι έμποροι μπορούν να χρησιμοποιήσουν αυτή την τεχνική συστηματικά εάν τη θεωρούν πολύ περίπλοκη ή πολύ επικίνδυνο για να εφαρμοστεί.

2.1.3 ΑΝΙΧΝΕΥΣΗ ΑΠΑΤΗΣ

Ο εντοπισμός απάτης πρέπει επίσης να αντιμετωπιστεί σε όλους τους κλάδους. Οι κλάδοι όπου γίνονται πολλές συναλλαγές είναι πιο ευάλωτοι. Οι πρωτοπόροι στη χρήση τεχνικών εξόρυξης δεδομένων για την πρόληψη της απάτης ήταν οι τηλεφωνικές εταιρείες και οι ασφαλιστικές εταιρείες, με τις τράπεζες να ακολουθούν πολύ πίσω. Η απάτη μπορεί έχει ως αποτέλεσμα μια επιχείρηση να χάσει σημαντικά ποσά των χρημάτων. Η δυνατότητα προστασίας μιας επιχείρησης από η πιθανότητα απάτης είναι μια σημαντική ανησυχία για ένας οργανισμός και η εξόρυξη δεδομένων μπορεί να βοηθήσει. Για να ανιχνεύσει δόλιες ενέργειες, μπορεί να κατασκευαστεί ένα μοντέλο χρησιμοποιώντας δόλια συμπεριφορά (ή δυνητικά δόλια συμπεριφορά) που έχει γίνει στο παρελθόν και στη συνέχεια να χρησιμοποιηθεί η εξόρυξη δεδομένων για να προσδιοριστεί μια παρόμοια συμπεριφορά.

2.1.4 ΑΠΟΚΤΗΣΗ ΚΑΙ ΔΙΑΤΗΡΗΣΗ ΠΕΛΑΤΩΝ

Η απόκτηση και η διατήρηση πελατών είναι πρωταρχικά μια λειτουργία μάρκετινγκ, αλλά συζητείται ξεχωριστά λόγω της ζωτικής σημασίας της για τις βιομηχανίες. Η απόκτηση πελατών, το καθήκον της απόκτησης νέων πελατών, είναι ένας σημαντικός στόχος για οποιαδήποτε οργάνωση γιατί χωρίς πελάτες, η επιχείρηση δεν μπορεί να ευδοκιμήσει. Η διατήρηση πελατών περιλαμβάνει τη διατήρηση αυτών των πελατών της επιχείρησης ήδη έχει. Αν και η απόκτηση πελατών είναι πολύ πιο εύκολο έργο και λιγότερο ακριβό από τη διατήρηση των πελατών. Οι Marple and Zimmerman (1999) συμπεραίνουν ότι «κοστίζει 33 έως 50 τοις εκατό λιγότερο να πουλήσει σε υπάρχοντες πελάτες από ό,τι για να πουλήσει σε καινούριους και ότι οι πιθανότητες να επαναλάβει τις συναλλαγές ενός υπάρχοντος πελάτη είναι εκθετικά υψηλότερες από τη μετατροπή κάποιου άλλου πελάτη.» Οι οργανισμοί μπορούν να χρησιμοποιήσουν εξόρυξη δεδομένων για να βοηθήσουν στην απόκτηση νέων πελατών και στη διατήρηση των υπαρχόντων πελατών.

Όσον αφορά την απόκτηση πελατών, τα προφίλ πελατών βοηθούν στον εντοπισμό των χαρακτηριστικών των καλών πελατών και θα βοηθήσει το τμήμα μάρκετινγκ να στοχεύσει σε νέους πελάτες. Η απόκτηση πελατών μπορεί να στοχεύει κατάλληλα χρησιμοποιώντας εξόρυξη δεδομένων για την εύρεση μοτίβων σε μια βάση δεδομένων πελατών. Στοχεύοντας μόνο σε άτομα που έχουν τη δυνατότητα να γίνουν πελάτες, ένας οργανισμός μειώνει τον χρόνο και τα χρήματα που δαπανώνται σε προσπάθειες απόκτησης πελατών.

Για να διατηρήσει τους υπάρχοντες πελάτες ενός οργανισμού, είναι σημαντικό για έναν οργανισμό να γνωρίζει ποιοι παράγοντες δείχνουν αν ένας πελάτης είναι πιθανό να εγκαταλείψει την επιχείρηση και να πάει σε ανταγωνιστές. Η εξόρυξη δεδομένων μπορεί να βοηθήσει στον εντοπισμό πελατών που είναι πιθανό να φύγουν και μπορεί να προσφέρει σε αυτούς τους πελάτες κίνητρα να μείνουν.

Ένας οργανισμός μπορεί να δημιουργήσει μοντέλα που προσφέρουν λεπτομερή "προφίλ" ή χαρακτηριστικά των πελατών που ενδέχεται να ανακληθούν, να είναι κερδοφόρα ή να ανταποκρίνονται σε μια συγκεκριμένη προσφορά. Όταν ένας οργανισμός γνωρίζει τους πελάτες που είναι πιο κερδοφόροι και πιο πιθανό να ανταποκριθούν ευνοϊκά σε μια προσφορά, υπάρχουν λιγότερες πιθανότητες να χαθούν αυτοί οι πελάτες από ανταγωνιστές. Χρησιμοποιώντας την εξόρυξη δεδομένων, μια επιχείρηση είναι σε καλύτερη θέση να κατανοήσει τις ανάγκες των πελατών της και, ως εκ τούτου, μπορεί να βελτιώσει και να επεκτείνει την υπάρχουσα σχέση.

Για την επιλογή των προφίλ πελατών, το σημαντικό που πρέπει να ληφθεί υπόψη είναι ότι οι πληροφορίες προφίλ που περιέχουν τις κατηγορικές μεταβλητές μπορεί να είναι μη αριθμητικές ή ονομαστικές μεταβλητές. Οι τεχνικές AI Intelligence δημιουργούν ακριβή προφίλ για τους πελάτες χρησιμοποιώντας τεχνικές αναζήτησης δέσμης και σταδιακής μάθησης έτσι ώστε να απεικονίζουν δομές ειδικών ομάδων πελατών. Οι δύο τύποι μοντέλων που χρησιμοποιούνται στην πρόβλεψη συμπεριφοράς πελατών είναι οι μέθοδοι ταξινόμησης και πρόβλεψης αξίας.

Στις μεθόδους ταξινόμησης, με βάση τις προηγούμενες πληροφορίες αθέτησης, τα επίπεδα κινδύνου συστηματοποιούνται σε δύο μέρη ως «Risky Group» και «Safe Group». Οι πελάτες που έχουν το προηγούμενο ιστορικό είναι "Risky" ενώ οι πελάτες που δεν έχουν ιστορικό είναι "Safe". Εδώ για να διαμορφώσετε τα μοντέλα που μπορούν να προβλέψουν το μέγεθος του κινδύνου με προεπιλεγμένα επίπεδα εφαρμογών νέων πελατών, εφαρμόζονται τα μοντέλα δέντρου αποφάσεων και κανόνα συσχέτισης. Οι τεχνικές επαγωγής μπορούν να εκτελεστούν χρησιμοποιώντας τις παραπάνω πληροφορίες κατηγοριοποίησης. Σε μεθόδους πρόβλεψης αξίας αντί να κατηγοριοποιεί τις νέες εφαρμογές πελατών, προσπαθεί να μαντέψει τον πιθανό όγκο αθέτησης νέων αιτήσεων πίστωσης. Οι προβλεπόμενες πληροφορίες είναι σε αριθμητική μορφή και έτσι χρειάζεται τεχνικές μοντελοποίησης για να ληφθούν αριθμητικές πληροφορίες ως μεταβλητές στόχου. Εδώ μπορεί να εφαρμοστεί ένα τεχνητό νευρωνικό δίκτυο (ANN) και η παλινδρόμηση. Οι περίφημες τεχνικές εξόρυξης δεδομένων που λαμβάνονται για τη δημιουργία των προφίλ πελατών είναι η παλινδρόμηση, ο κανόνας συσχέτισης, η ομαδοποίηση και η ταξινόμηση.

2.2 ΣΥΛΛΟΓΗ ΧΡΕΟΥΣ

Η συλλογή χρέους ήταν σημαντικό κομμάτι σε όλη την ανθρώπινη ιστορία, ακόμα και πριν την άνθηση του καπιταλισμού. Στην Ευρώπη εμφανίζεται πρώτη φορά στη ρωμαϊκή αυτοκρατορία , όταν για τη συλλογή χρέους ήταν υπεύθυνες δημόσιες εταιρείες οι οποίες ήταν υπεύθυνες να συλλέξουν φόρους της πολιτείας και να προβούν σε μη δικαστικά μέσα. Η συλλογή χρέους συνεχίστηκε και τον μεσαίωνα .

2.3 ΕΤΑΙΡΕΙΕΣ ΠΟΥ ΕΠΙΚΟΙΝΩΝΟΥΝ ΜΕ ΤΟΝ ΠΕΛΑΤΗ

Στην Ελλάδα δραστηριοποιούνται εταιρείες που ενημερώνουν τον πελάτη για το χρέος του. Συνήθως οι εταιρείες δανειστές, προσλαμβάνουν αυτές τις εταιρείες όταν το χρέος είναι μη διαχειρίσιμο, είτε για να πεισθεί ο οφειλέτης να καταβάλει πληρωμή, είτε να οριστεί κάποιος διακανονισμός. Αυτές οι εταιρείες λειτουργούν εξωδικαστικά και δεν επιτρέπεται να συλλέξουν το χρέος, ούτε το χρέος μεταφέρεται σε αυτές. Οι εταιρείες διαχείρισης χρέους δέχονται ένα ποσοστό της αποπληρωμής του χρέους (προμήθεια) για τις υπηρεσίες τους.

Οι εταιρείες είναι υποχρεωμένες να ακολουθούν τους παρακάτω κανόνες σύμφωνα με νομοθεσία του 2012 [9] :

1) Έναρξη επικοινωνίας με τον οφειλέτη μόνο δέκα μέρες αφού η οφειλή έχει γίνει εκκρεμούσα
2) Η επικοινωνία μπορεί να γίνει μόνο τις ώρες 9.00 – 20.00
3) Η επικοινωνία μπορεί να γίνει μόνο κάθε δυο μέρες

Η συλλογή χρέους μπορεί να αφορά ληξιπρόθεσμες οφειλές πολλών διαφορετικών προϊόντων. Συχνά οι οφειλές αυτές περιλαμβάνουν:

- Προσωπικό δάνειο
- Καταναλωτικό δάνειο
- Απλήρωτους λογαριασμούς σε εταιρείες τηλεφωνίας/ενέργειας

Τα γραφεία είσπραξης μπορούν να αναζητήσουν το παλιό χρέος μόλις παρέλθει μερικούς μήνες και επ' αόριστόν μετά από αυτό. Εξαρτάται από την εταιρεία που εισπράττει το χρέος, το ποσό οφειλής και το είδος του χρέους.

Εάν υπάρχει απλήρωτο χρέος σε καθυστέρηση, συνήθως η ενημέρωση πραγματοποιείται μέσω γραπτών ειδοποιήσεων και τηλεφωνικών κλήσεων μέσω του αρχικού πιστωτή. Για παράδειγμα, για ένα παλιό φοιτητικό δάνειο που δεν αποπληρώθηκε, ο δανειστής θα προσπαθήσει να επικοινωνήσει για να λάβει τον τρέχοντα λογαριασμό. Εάν δεν καταφέρει να λάβει το ποσό , τελικά θα σταματήσει. Αυτό συμβαίνει συνήθως όταν συμβαίνει η μετάβαση από τον αρχικό πιστωτή στον εισπράκτορα χρεών.

Τα γραφεία είσπραξης οφειλών και οι εισπράκτορες χρεών θα χρησιμοποιήσουν τις πληροφορίες που υπάρχουν στο αρχείο για να επικοινωνήσουν. Χρησιμοποιούνται η τρέχουσα διεύθυνσή του οφειλέτη, ο αριθμός τηλεφώνου, ακόμη και τα στοιχεία επικοινωνίας των συγγενών. Εάν μπορούν, οι εισπράκτορες χρεών θα χρησιμοποιήσουν προσωπικές τραπεζικές πληροφορίες, συμπεριλαμβανομένων λογαριασμών ταμιευτηρίου και επενδύσεων, για να καθορίσουν εάν υπάρχουν τα χρήματα για να αποπληρωθεί ένα χρέος.

3 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΛΗΞΙΠΡΟΘΕΣΜΩΝ ΟΦΕΙΛΩΝ ΔΙΚΗΓΟΡΙΚΟΥ ΓΡΑΦΕΙΟΥ

3.1 ΣΤΟΧΟΣ ΑΝΑΛΥΣΗΣ

Ο βασικός στόχος της ανάλυσης είναι να βρεθούν αυτά τα χαρακτηριστικά από τα δεδομένα που έχουν συσσωρευτεί στην βάση δεδομένων και να εξαχθούν ανάλογα συμπεράσματα σχετικά με τους πελάτες , τις οφειλές και την συμπεριφορά τους. Να αποτυπωθεί αν υπάρχει κάποια συσχέτιση μεταξύ των διαφορετικών δεδομένων και της συμπεριφοράς του οφειλέτη, αν υπάρχουν συγκεκριμένα στοιχεία τα οποία χρειάζεται να λάβουμε υπόψιν ανάλογα με το αν ο οφειλέτης αποδίδει τις πληρωμές του , απομειώνει την οφειλή ή γενικότερα τηρεί το πλάνο αποπληρωμών για το χρέος που κατέχει. Γενικότερα να μπορούμε να προβλέψουμε την συμπεριφορά ενός πελάτη και να προβούμε στις καταλληλότερες ενέργειες χρησιμοποιώντας εργαλεία μηχανικής μάθησης.

Γενικότερα τα δεδομένα τα οποία έχει στη κατοχή του το δικηγορικό γραφείο αφορούν:

1. Δημογραφικά χαρακτηριστικά
2. Στοιχεία Οφειλής
3. Ενέργειες τηλεφωνικού κέντρου
4. Ιστορικό ενεργειών

Το σύνολο των παραπάνω δεδομένων καταχωρείται σε ένα CRM, το οποίο συνδέεται με την αντίστοιχη βάση δεδομένων. Τα διαφορετικά δεδομένα έχουν αποθηκευτεί σε σχεσιακή βάση δεδομένων και η άντληση τους γίνεται με την χρήση SQL queries κι άλλων εργαλείων. Για τους σκοπούς της ανάλυσης είναι σημαντικό τα δεδομένα να χωριστούν και να καθαριστούν προκειμένου να εφαρμοστεί πάνω τους η κατάλληλη ανάλυση.

Οι στόχοι της ανάλυσης καθορίστηκαν με βάση τις επιχειρησιακές ανάγκες και τις δυνατότητες που δόθηκαν από επιλεγμένα σύνολα δεδομένων με βάση την ποιότητα και την ποσότητα των δεδομένων που είχαμε στην διάθεση μας.

Για την ανάλυση κρίθηκε χρησιμότερο να ερευνηθούν συγκεκριμένα πεδία ανάλυσης.

Το γενικό πλαίσιο των αναλύσεων είχε να κάνει με το αν:

- Οι οφειλέτες έχουν κάποιον διακανονισμό και αν τηρούν αυτόν τον διακανονισμό
- Οι οφειλέτες καταβάλλουν ή όχι πληρωμές
- Οι οφειλέτες αποπληρώνουν ή όχι την οφειλή τους
- Οι οφειλέτες καταλήγουν σε κάποια θετική ενέργεια

Τα στοιχεία του οφειλέτη εξετάστηκαν σε συνδυασμό με στοιχεία της οφειλής και των ενεργειών που διενεργήθηκε από το δικηγορικό γραφείο. Γενικότερα έγινε μια προσπάθεια να διαχωριστούν οι πελάτες ανάλογα με κοινά τους στοιχεία και αν αυτά οδηγούν σε κάποιο συμπέρασμα σχετικά με την συμπεριφορά τους όσον αφορά την οφειλή. Επίσης έγινε μια έρευνα σε επίπεδο ενεργειών και το αν μπορούμε να βγάλουμε κάποια ασφαλή

συμπεράσματα όσον αφορά τις ενέργειες οι οποίες μπορεί να ωθήσουν έναν πελάτη σε πληρωμή .Τέλος χρησιμοποιήθηκαν και τα στοιχεία της οφειλής.

Για τα παραπάνω δεδομένα χρειάστηκε να βρεθεί κατάλληλος τρόπος μηχανικής μάθησης για να εξαχθούν σωστά και αξιοποιήσιμα συμπεράσματα, τα οποία με τη σειρά τους θα δίνουν και την δυνατότητα πρόβλεψης ώστε να μπορέσει τελικά η όποια γνώση να αξιοποιηθεί σε μελλοντικά δεδομένα. Αντίστοιχες εργασίες με μοντέλα μηχανικής μάθησης έχουνε εκπονηθεί για την πρόβλεψη ρίσκου απόδοσης δανείου ή την πιθανότητα απάτης και για το profiling των πελατών [10],[11],[13] , στην περίπτωση της παρούσας εργασίας έχει να κάνει με τον βαθμό που η εφαρμογή αυτών των τεχνικών θα φέρει κάποιο αποτέλεσμα στα συγκεκριμένα δεδομένα του δικηγορικού γραφείου κι ακόμα αν δίνουν την δυνατότητα αλλαγής στρατηγικής.

3.2 ΑΝΑΛΥΣΕΙΣ ΣΕ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΔΗΜΟΓΡΑΦΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΟΦΕΙΛΕΤΩΝ ΚΑΤΑΝΑΛΩΤΙΚΩΝ ΔΑΝΕΙΩΝ ΚΑΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΔΙΑΚΑΝΟΝΙΣΜΟΥ

Για την εργασία επιλέχθηκε να γίνει ανάλυση σε στοιχεία οφειλετών τραπεζικών προϊόντων (καταναλωτικών δανείων) τα οποία δεν έχουν εξοφληθεί και προτείνεται ο διακανονισμός του χρέους του οφειλέτη.

Δίνεται ουσιαστικά η δυνατότητα στον δανειολήπτη – οφειλέτη λόγω αδυναμίας εξόφλησης του χρέους να οριστεί ένα συγκεκριμένο πόσο μικρότερο της αρχικής οφειλής το οποίο μπορεί να το αποπληρώσει σε μια ή και περισσότερες δόσεις. Το ύψος και ο αριθμός δόσεων του διακανονισμού ορίζεται με συγκεκριμένους κανόνες από την συνεργάτιδα εταιρεία – fund.

Σαν στοιχεία επιλέχθηκε να αναλυθούν μια σειρά από δημογραφικά χαρακτηριστικά του οφειλέτη τα οποία δίνονται από την συνεργάτιδα εταιρεία – fund και ενημερώνονται καθημερινά μέσω αρχείων στο σύστημα (CRM).

Για την ανάλυση επιλέχθηκε να κρατείται και η πληροφορία του αποτελέσματος του διακανονισμού ώστε να οριστεί με βάση την ιστορικότητα των δεδομένων οι οφειλέτες που αποπληρώνουν (μερικώς ή πλήρως) τον διακανονισμό και αυτοί που δεν αποπληρώνουν.

Για την ανάλυση των οφειλετών με βάση τα δημογραφικά χαρακτηριστικά και τα χαρακτηριστικά του δανείου επιλέχθηκε ένα σύνολο δεδομένων με τα κάτωθι χαρακτηριστικά.

Ονομασία Χαρακτηριστικού	**Περιγραφή**	**Εύρος Τιμών**	**Σχόλια**
VEND_CODE (int)	Μοναδικός κωδικός Fund-τράπεζας	(2 ,3 ,5 ,7 ,9)	PQH, Quant, Intrum, EOS, B2k
CITIZENSHIP (int)	Κωδικός Ελληνικής ή Άλλης υπηκοότητας	(1, 2)	1- Έλληνας 2- Αλλοδαπός
Occupation (int)	Κωδικός Επαγγέλματος	(1 – 6)	1- Ιδιωτικός Τομέας

			2- Δημόσιος Τομέας 3- Συνταξιούχος 4- Ελεύθερος Επαγγελματίας 5- Άνεργος 6- Άγνωστο
AgeIntYears (int)	Ηλικία του δανειολήπτη	(17-100)	Αποκλείστηκαν τιμές που δεν ανήκουν σε αυτό το εύρος καθώς δεν αποτελούν φυσικά πρόσωπα.
Region (int)	Περιοχή διαμονής του δανειολήπτη	(1 - 8)	1- Αττική 2- Πελοπόννησος 3- Κεντρική Ελλάδα 4- Θεσσαλία , Ήπειρος 5- Δυτική και Κεντρική Μακεδονία 6- Μακεδονία 7- Κρήτη 8- Αιγαίο
SEX (int)	Φύλο	(1, 2)	1- Γυναίκα 2- Άντρας
No_OF_CASES (int)	Αριθμός ανοιχτών υποθέσεων από όλα τα χαρτοφυλάκια	(1 – 14)	
LOV_DESC (char)	Αποτέλεσμα Διακανονισμού	(Fulfilled, Partially Fulfilled) και (Cancelled, Not Fulfilled, Out of Collection)	Διαχωρισμός των οφειλετών αν ο διακανονισμός είναι Εκπληρωμένος ή οχι
SETE_SETTLEMENT_AMT (decimal)	Ποσό του διακανονισμού		Νομισματικές τιμές
SETE_NR_INSTALLMENTS (INT)	Αριθμός δόσεων διακανονισμού		
Case Open In Years (int)	Χρόνια τα οποία οφειλή είναι ανοιχτή		
CASE_DEBT_AMT (decimal)	Ποσό Οφειλής		Νομισματικές τιμές

Πίνακας 1

3.2.1 ΑΝΤΛΗΣΗ ΚΑΙ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ

Το παραπάνω σύνολο δεδομένων (9199 εγγραφές) εξάχθηκε από σχεσιακή βάση δεδομένων (Microsoft SQL) με τη χρήση SQL Query. Για τον μετασχηματισμό των δεδομένων ακολουθήθηκαν συγκεκριμένες πρακτικές.

Τα στοιχεία citizenship, Occupation, SEX και region , δεν δίνονταν με τη μορφή που περιγράφεται παραπάνω . Κατηγοριοποιήθηκαν με βάση συγκεκριμένους κανόνες προκειμένου να δώσουν μια πιο συνεπή εικόνα και να γίνει δυνατή η ανάλυση τους.

Στην περίπτωση του region αξιοποιήθηκε ο Τ.Κ. και οι περιοχές οι οποίες αυτός συμβολίζει.

Στην περίπτωση του επαγγέλματος ο διαχωρισμός έγινε με βάση τις κατηγορίες που περιγράφονται στον πίνακα 1 με τη χρήση ενός πίνακα mapping , ο οποίος δημιουργήθηκε για να εξυπηρετήσει αυτή τη διαδικασία. Η περίπτωση το επάγγελμα να μην είναι καταχωρημένο, επιλέχθηκε να εισαχθεί σε μια πρόσθετη κατηγορία (Unknown).

Στις περιπτώσεις AgeIntYears και Case Open In Years , τα πεδία της βάσης δεδομένων περιείχαν ημερομηνίες (Ημερομηνία Γέννησης, Ημερομηνία ανοίγματος λογαριασμού) , έγινε τροποποίηση ώστε να εμφανίζονται σε έτη.

Στην περίπτωση της υπηκοότητας χρησιμοποιήθηκε η εξής μέθοδος.

Κατηγοριοποιήθηκαν τα ονόματα (FIRST NAMES) τα οποία με εμφανίζονται περισσότερες φορές σε μια υπηκοότητα (π.χ. το όνομα Γιάννης εμφανίστηκε 100 φορές για την Ελληνική , περισσότερες από οποιαδήποτε άλλη υπηκοότητα) και έτσι δόθηκε η αντίστοιχη τιμή στην ελλείπουσα περίπτωση.

Ελλείπουσες τιμές παρατηρήθηκαν και σε άλλες περιπτώσεις του συνόλου δεδομένων (Αριθμός Δόσεων Διακανονισμού, Ηλικία (111) , Περιοχή (220)), αυτές σε επόμενο στάδιο της επεξεργασίας αφαιρέθηκαν από το σύνολο δεδομένων.

3.2.2 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Στο παραπάνω dataset έγινε ανάλυση ώστε να εξαχθούν συμπεράσματα που αφορούν τους «συνεπείς» οφειλέτες και τους «ασυνεπείς» και έτσι να προκύψουν συμπεράσματα τα οποία μπορούν να αξιοποιηθούν για την διαχείριση αναθέσεων νέων οφειλετών στο σύστημα με πρόβλεψη της συμπεριφοράς τους ανάλογα με τα δημογραφικά τους χαρακτηριστικά.

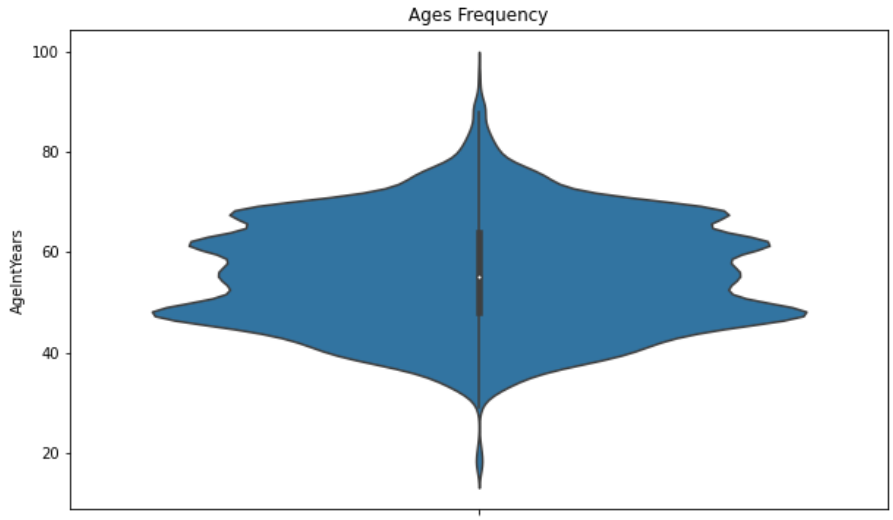
Με τον καθαρισμό τον δεδομένων (τιμές null) και διαγραφή μη εγκύρων τιμών απομένουν 7873 εγγραφές στο σύνολο δεδομένων.

Τα δεδομένα χωρίστηκαν σε δυο επιμέρους Dataset, το Dataset των οφειλετών με status διακανονισμού (Fulfilled, Partially Fulfilled) – «συνεπής» οφειλέτης και οι «ασυνεπής» οφειλέτες οι οποίοι αφορούν όλα τα υπόλοιπα statuses.

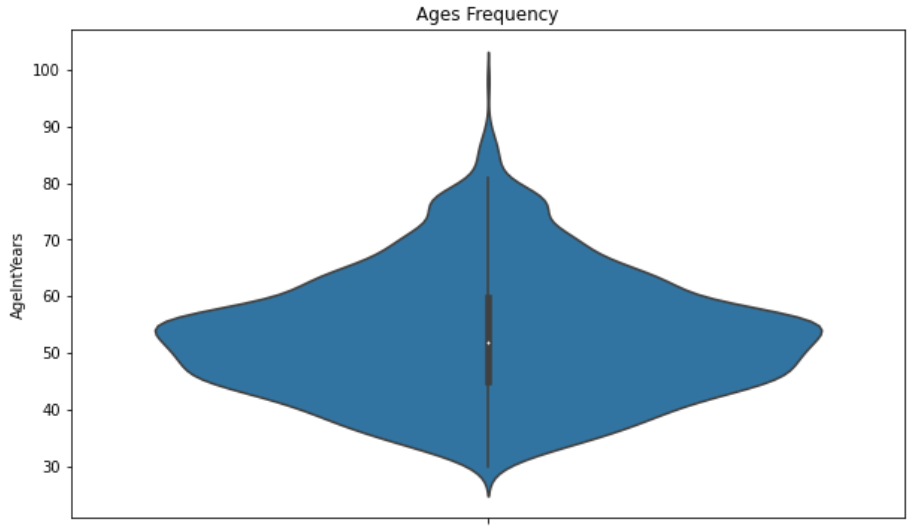
Το σύνολο δεδομένων των συνεπών οφειλετών αφορά 1410 εγγραφές, το αντίστοιχο σύνολο «μη συνεπών» οφειλετών αφορά 5486 εγγραφές.

Η πλειοψηφία των παρατηρήσεων αφορά τις ηλικίες 40-70, για τους ασυνεπείς οφειλέτες. Το συνολικό εύρος των ηλικιών που αντλήσαμε ήταν των ηλικιών 17-100. Δεδομένου ότι τα δεδομένα αφορούν καταναλωτικά Δάνεια είναι αρκετά λογικό να συγκεντρώνονται στις

ηλικίες 40-65. Αντίθετα μικρότερες (17-30) ή μεγαλύτερες (70+) ηλικίες δεν εμφανίζονται συχνά στο dataset (Εικόνες 8,9,10,11).

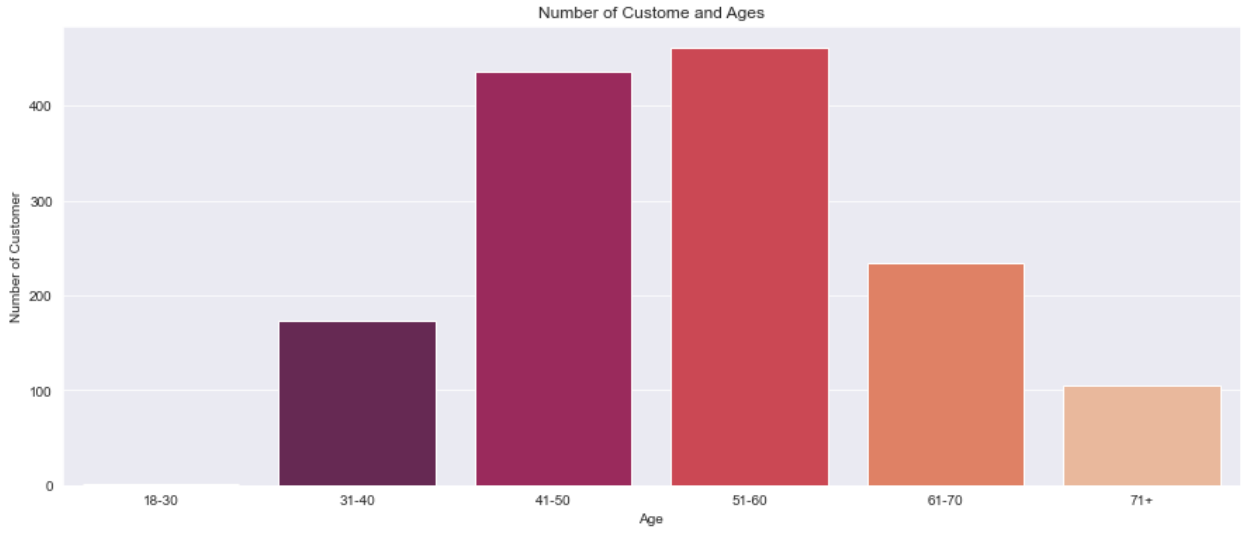


Εικόνα 8 - Ασυνεπείς Οφειλέτες Ηλικία

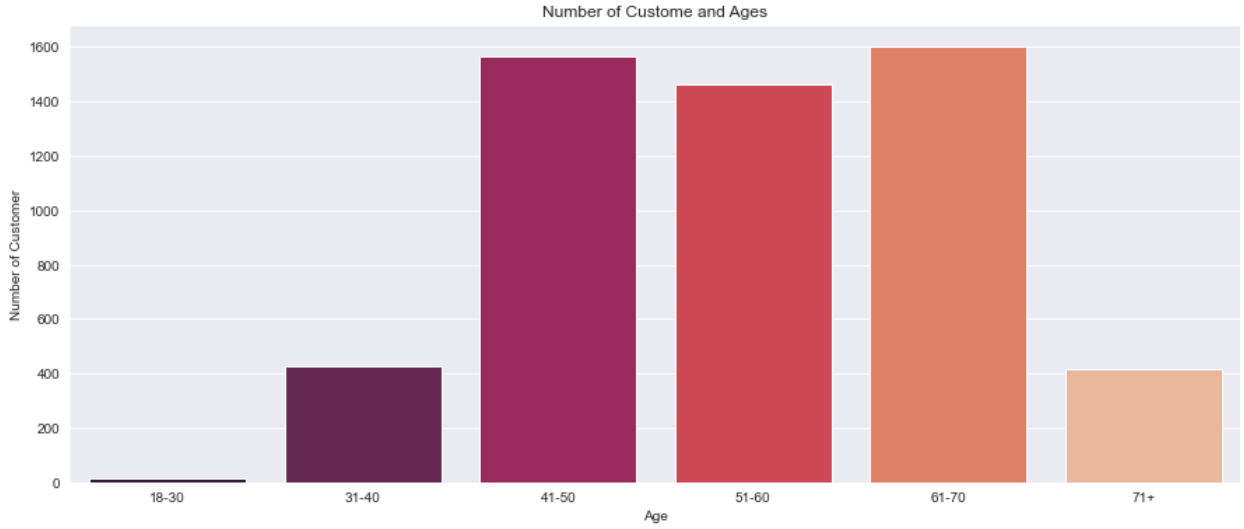


Εικόνα 9 – Συνεπείς Οφειλέτες Ηλικία

3.2.2.1 ΗΛΙΚΙΕΣ ΑΝΑ ΔΕΚΑΕΤΙΕΣ



Εικόνα 10 - Συνεπείς Οφειλέτες Ηλικία Ραβδόγραμμα

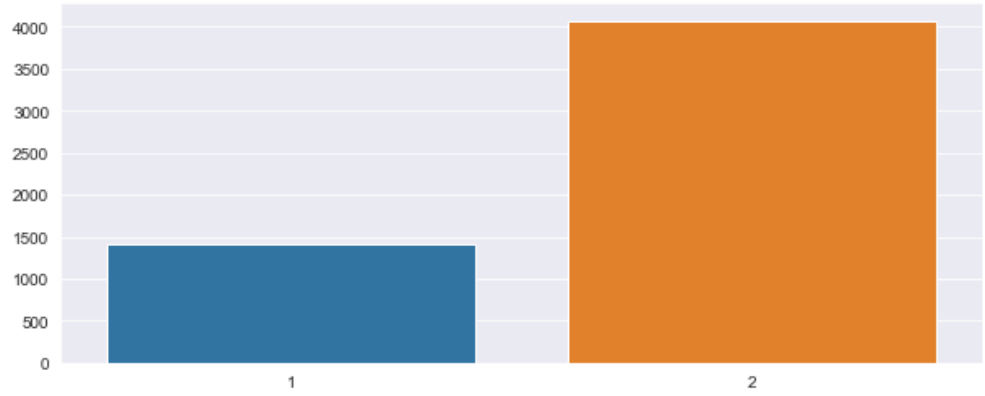


Εικόνα 11 - Ασυνεπείς Οφειλέτες

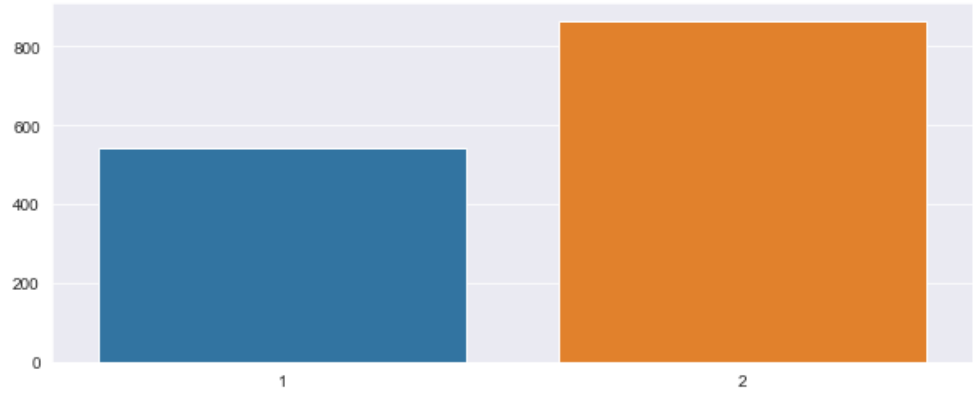
Στα δυο σύνολα δεδομένων οι Άντρες είναι περισσότεροι από τις γυναίκες , στους ασυνεπείς οφειλέτες οι γυναίκες είναι περίπου το 1/4 των ανδρών, στο σύνολο των συνεπών οφειλετών οι γυναίκες είναι περίπου μισές από τους άνδρες.

1 – Γυναίκες

2 – Άντρες

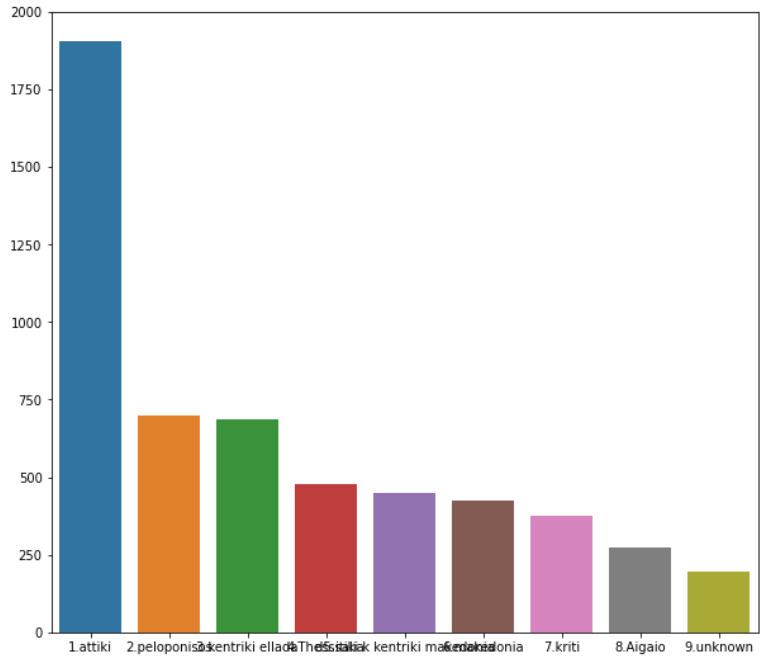


Εικόνα 12 - Ασυνεπείς Οφειλέτες Φύλο (1-Γυναίκα, 2-Αντρας)

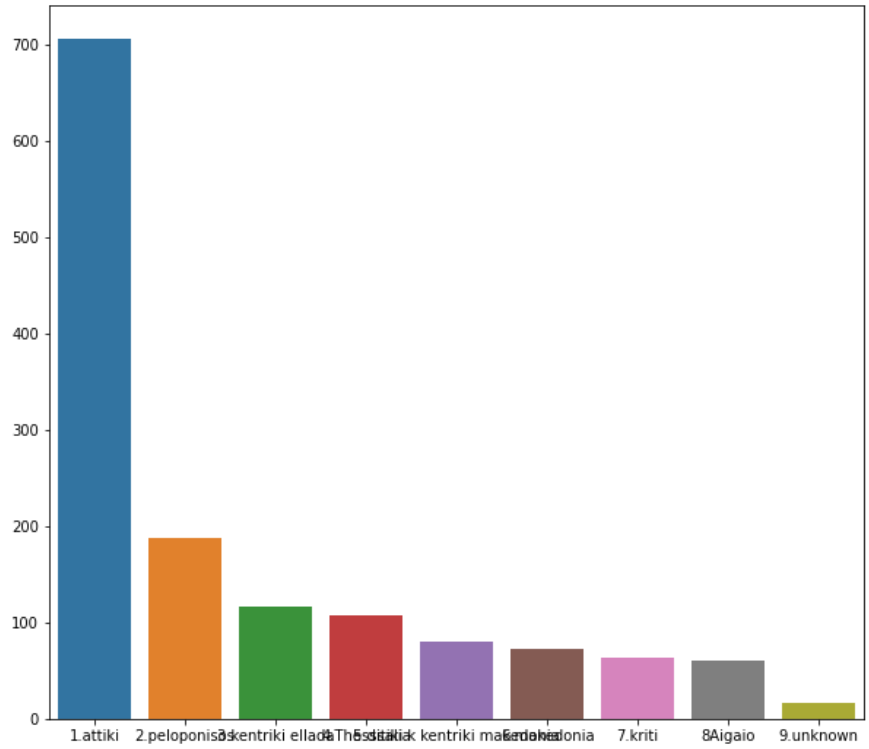


Εικόνα 13 - Συνεπείς Οφειλέτες Φύλο (1-Γυναίκα, 2-Αντρας)

Οι περιοχές στις οποίες χωρίζονται οι οφειλέτες με βάση το ΤΚ. Και στα δυο σύνολα δεδομένων οι διευθύνσεις της Αττικής είναι πλειοψηφία.



Εικόνα 14 - Ασυνεπείς Οφειλέτες Περιοχή Βασικής Διεύθυνσης



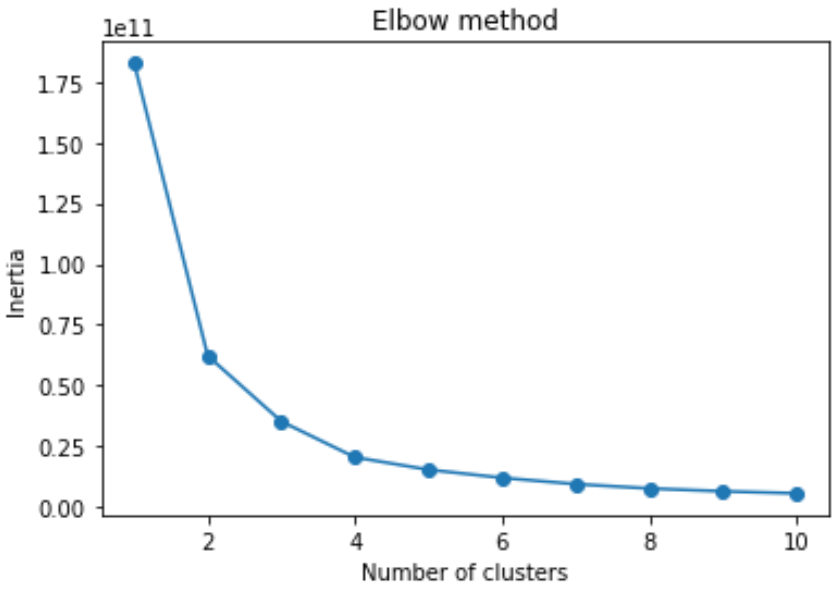
Εικόνα 15 - Συνεπείς Οφειλέτες Περιοχή Βασικής Διεύθυνσης

3.2.3 CLUSTERING

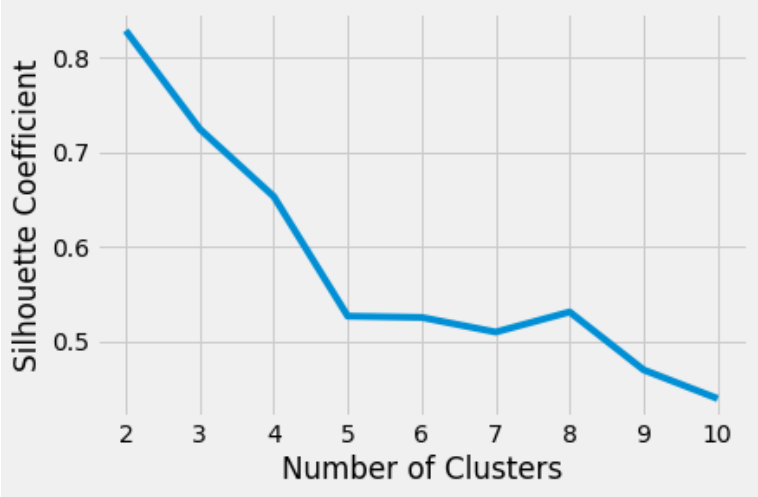
Εξετάστηκαν μοτίβα στα χαρακτηριστικά των χρηστών για τα δυο σύνολα δεδομένων που επιλέχθηκαν. Η τεχνική που επιλέχθηκε ήταν το K-means clustering και αυτό γιατί ήταν γνωστά εκ των προτέρων τα χαρακτηριστικά που εμφανίζουν τους οφειλέτες.

Για την επιλογή των clusters χρησιμοποιήθηκε αρχικά η elbow method (Εικόνες 16,18) αλλά και το sillouete score (Εικόνες 17,19) για να βρεθεί ο καλύτερος δυνατός αριθμός cluster ανά σύνολο δεδομένων.

Με την εφαρμογή του elbow method για τα δυο σύνολα δεδομένων εξήχθησαν τα παρακάτω αποτελέσματα:



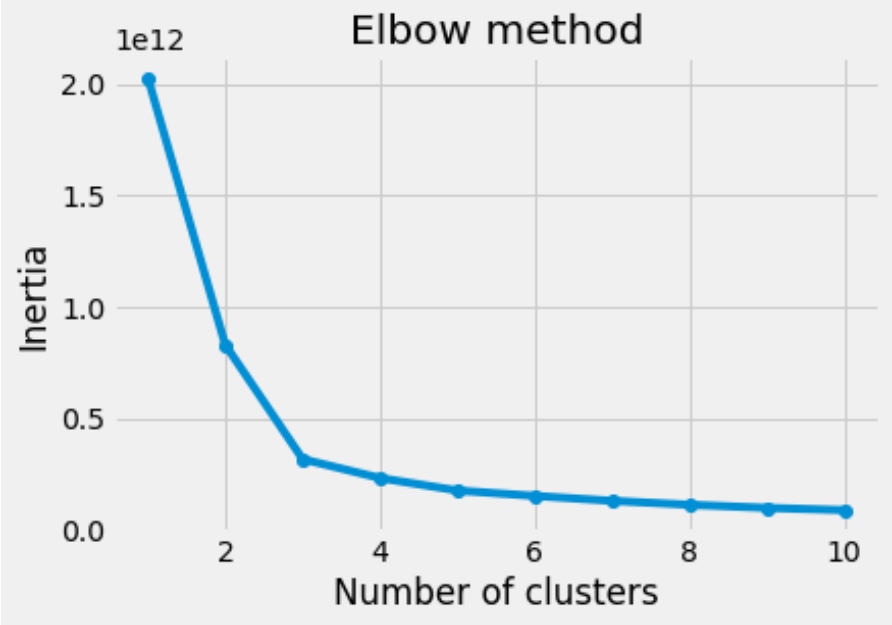
Εικόνα 16 - Συνεπείς Οφειλέτες elbow method



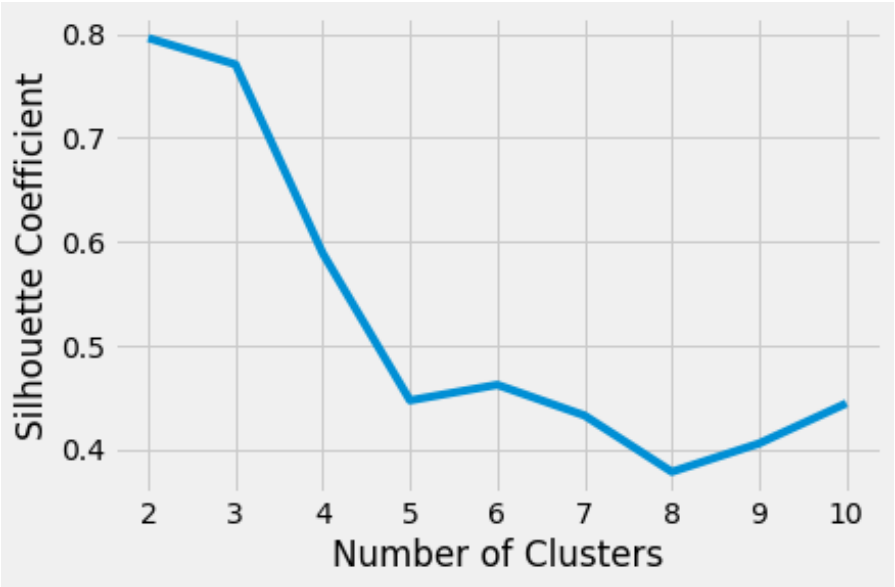
Εικόνα 17 - συνεπείς οφειλέτες – Silhouette

Φαίνεται ότι η καλύτερη επιλογή για τους συνεπείς οφειλέτες είναι τα 2 clusters (Εικόνες 16,17).

Αντίστοιχα για τους ασυνεπείς



Εικόνα 18 - Ασυνεπείς ELbow Method



Εικόνα 19 - Ασυνεπείς Silhouette

Silhouette score - 0.77

Φαίνεται ότι η καλύτερη επιλογή για τους συνεπείς οφειλέτες είναι τα 3 clusters (εικόνες 18, 19). Υλοποιούμε τον αλγόριθμο K-means στο σύνολο δεδομένων των συνεπών Οφειλετών για 2 clusters. Το ένα cluster έχει 89 υποθέσεις και το άλλο 1321

3.2.3.1 Clusters

DataSet	Clusters	Αριθμός Υποθέσεων	Μ.Ο. Ποσών διακανονισμού	Μ.Ο. Αριθμού δόσεων	Διάμεσος Ετών χρέους	Ποσό χρέους	Περιοχή
Fulfilled							
	1	1321	1.000 €	3	8 χρόνια	5.000 €	50% των εγγραφών στην Αττική
	2	89	3.673 €	2	10 χρόνια	42.000 €	50% των εγγραφών στην Αττική
Not Fulfilled							
	1	57	23.000 €	137	14 χρόνια	150.000 €	75% των εγγραφών στην Μακεδονία (όχι Κεντρική)
	2	607	9.700 €	1758	11 χρόνια	34.443 €	75% Νησιά (Αιγαίου και Σποράδες- Εύβοια) και Αττική
	3	4822	2.500 €	343	12 χρόνια	4.000 €	75% Αττική και Θεσσαλονίκη

3.2.3.2 ΣΥΜΠΕΡΑΣΜΑΤΑ

Από την ανάλυση που πραγματοποιήθηκε μπορούν να εξαχθούν τα εξής συμπεράσματα:

Η εθνικότητα δεν φαίνεται να διαδραματίζει ιδιαίτερο ρόλο στην κατηγοριοποίηση.

Η συμμετοχή των γυναικών στις κατηγορίες των συνεπών οφειλετών είναι μεγαλύτερη από ότι στην αντίστοιχη των Αντρών , οι οποίοι στις περιπτώσεις των κακοπληρωτών είναι σχεδόν απόλυτη πλειοψηφία, ενώ αντίθετα όσες γυναίκες υπάρχουν στο αρχικό σύνολο δεδομένων τόσες υπάρχουν και στις κατηγορίες των συνεπών οφειλετών.

Η ηλικία και στις δυο περιπτώσεις φαίνεται να κυμαίνεται στο εύρος των 40-60.

Όσον αφορά την περιοχή των συνεπών οφειλετών αυτή είναι κατά κύριο λόγο η Αττική, σε αντίθεση με τα clusters των ασυνεπών , στα οποία βλέπουμε μια κατηγορία να ανήκει σε Νησιά Αιγαίου, Σποράδες και Κεντρική Ελλάδα μαζί με Αττική και μια άλλη σε περιοχές της Μακεδονίας , εκτός Θεσσαλονίκης.

Ο αριθμός των ανοιχτών υποθέσεων δεν φαίνεται να διαδραματίζει σημαντικό ρόλο, ίσως παρατηρούμε μια μικρή αύξηση στους ασυνεπείς οφειλέτες.

Αντίθετα καθαρή είναι η διαφορά στον αριθμό δόσεων , όσον αφορά τους συνεπείς οφειλέτες βλέπουμε μικρό αριθμό δόσεων 1-3 και στα 2 cluster, ενώ αντίθετα στις περιπτώσεις των ασυνεπών βλέπουμε πολύ μεγάλο αριθμό δόσεων >100 με αντίστοιχα μεγάλο ποσά διακανονισμού 4.000 έως 150.000.

Τέλος ίσως μια σημασία έχει και τα έτη που είναι ανοιχτό το χρέος που στην περίπτωση των συνεπών οφειλετών κυμαίνονται από 8-10 ενώ στην περίπτωση των ασυνεπών υπερβαίνουν τα 11 χρόνια.

3.2.4 CLASSIFICATION

Η υλοποίηση του συγκεκριμένου Δένδρου απόφασης υλοποιήθηκε για να μπορέσουμε να ξεχωρίσουμε τα χαρακτηριστικά που διαδραματίζουν μεγαλύτερο ρόλο στην ανάλυση και άρα να τα επιλέξουμε έναντι όλων των χαρακτηριστικών με σκοπό την εφαρμογή λογιστικής παλινδρόμησης στα εναπομείναντα χαρακτηριστικά.

Στη συγκεκριμένη ανάλυση επιλέχθηκε ένα dataset με τα εξής χαρακτηριστικά:

```
['CITIZENSHIP','mapped_description','AgeIntYears','region','SEX',
'NO_OF_CASES','SETE_SETLEMENT_AMT','SETE_NR_INSTALLMENTS','CaseOpenInYe
ars','CASE_DEBT_AMT']
```

Σαν μεταβλητή στόχος κατηγοριοποιούνται οι εγγραφές με 1, αν οδήγησαν σε αποπληρωμή διακανονισμού και σε 0 αν δεν αποπληρώθηκε ο διακανονισμός.

Στόχος είναι να δούμε ποια από τα χαρακτηριστικά που επιλέξαμε διαδραματίζουν μεγαλύτερο ρόλο στην αποπληρωμή ενός διακανονισμού, από την εμφάνιση των πιο κρίσιμων χαρακτηριστικών στους κόμβους ενός δέντρου απόφασης.

Για την επιλογή των χαρακτηριστικών προχωρήσαμε στις παρακάτω δοκιμές:

1. Με όλα τα χαρακτηριστικά
   Το δέντρο δεν βγάζει νόημα , είναι πάρα πολύ μεγάλο , σε κάθε κόμβο ξαναεμφανίζεται χαρακτηριστικό που εμφανίζεται σε προηγούμενο επίπεδο.
2. Με χαρακτηριστικά
   ```
   ['CITIZENSHIP','region','SEX']
   ```

Χωρίς βάθος δέντρου και πάλι δεν οδηγούμαστε σε αποτέλεσμα

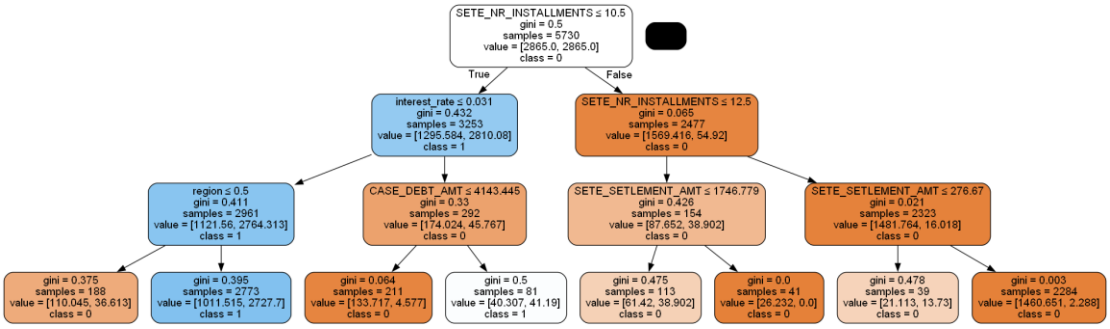
3. Με τα ίδια χαρακτηριστικά με εντροπία και μέγιστο βάθος (3) όλα τα φύλλα οδηγούν στο 0

είναι μια αρκετά ανισόρροπη αναλογία. Συμβαίνει ότι οι βέλτιστες διαχωρισμοί για το συγκεκριμένο σύνολο δεδομένων σας σε αυτό το βάθος παράγουν διαχωρισμούς όπου η κλάση πλειοψηφίας είναι η κλάση 2. Αυτή είναι φυσιολογική συμπεριφορά, και προϊόν των δεδομένων, και δεν προκαλεί έκπληξη, δεδομένου ότι η κλάση 2 υπερτερεί της κατηγορίας 1 κατά περίπου 4:1.

Αρά προσθέτουμε weight για να «διορθώσουμε» την αναλογία των imbalanced data.

Με balanced weight παίρνουμε φύλλα και από τι δύο κλάσεις , 0 και 1 αλλά το accuracy είναι πολύ χαμηλό 0,55.

4. Όλα τα χαρακτηριστικά για balanced data και βάθος 3, το accuracy είναι βελτιωμένο
   `Accuracy: 0.6991042345276873`



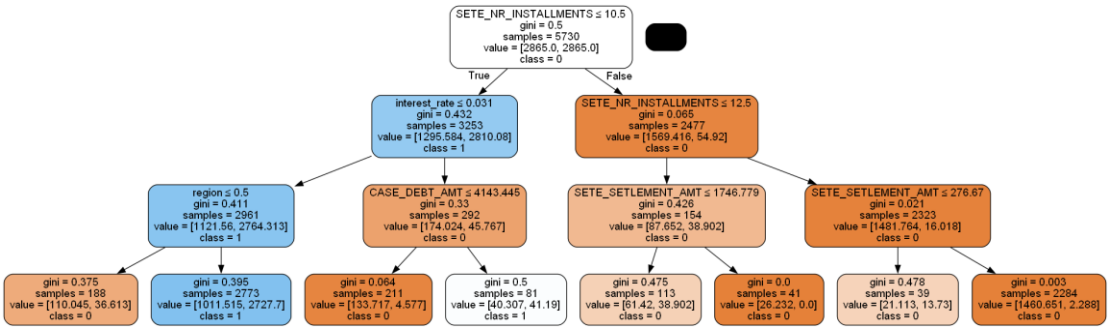
Εικόνα 20 - Δενδρο Απόφασης Διαφορα Χαραστηριστικα

Ξεχωρίζουν συγκριμένα χαρακτηριστικά που μας οδηγούν στο fulfilled or not (LOV_desc 0 ή 1)

- Number of installments
- Interest rate
- Settlement amount
- Debt amount
- Region

  Και
- Target variable : LOV_DESC (0,1)

Λόγω των προηγούμενων αποτελεσμάτων εκτελέστηκε ο αλγόριθμος με αυτά τα χαρακτηριστικά.



Εικόνα 21 - Δένδρο Απόφασης Μεμονωμένα Χαρ/κα

Ο αλγόριθμος χωρίς βάθος δίνει Accuracy: 0.8188110749185668, το δέντρο απόφασης όμως είναι όμως πολύ μεγάλο για να προκύψει κάποιο νόημα οπτικά.

3.2.4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Τελικά καταλήγουμε στο να πάρουμε τα παρακάτω στοιχεία:

- Number of installments
- Interest rate
- Settlement amount
- Debt amount
- Region

ως τις μεταβλητές οι οποίες δίνουν την καλύτερη δυνατή πρόβλεψη με μικρό βάθος δέντρου.

Φαίνεται ότι έχουν εξαιρεθεί σχεδόν όλες οι μεταβλητές οι οποίες αφορούν τα δημογραφικά χαρακτηριστικά του οφειλέτη (Ηλικία, Φύλο, Επάγγελμα), ενώ στοιχεία του δανείου και του διακανονισμού φαίνεται να διαδραματίζουν μεγαλύτερο ρόλο στην εξέλιξη αποπληρωμής ενός διακανονισμού.

3.2.4.2 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Αφού χρησιμοποιήθηκε το δένδρο απόφασης για να επιλεγούν τα χαρακτηριστικά που διαδραματίζουν τον σημαντικότερο ρόλο στην ανάλυση, επιλέχθηκαν τα ίδια χαρακτηριστικά και εφαρμόστηκε αλγόριθμος λογιστικής παλινδρόμησης με σκοπό την πρόβλεψη.

Τα δημογραφικά στοιχεία όπως το φύλλο, η ηλικία, το επάγγελμα δεν φαίνεται να διαδραματίζουν σημαντικό ρόλο στην αποπληρωμή ενός διακανονισμού. Αντίθετα τα χαρακτηριστικά ενός διακανονισμού (ποσό διακανονισμού, αριθμός δόσεων) όπως και στοιχεία (Ποσό συνολική οφειλής, interest rate) που αφορούν το χρέος (DEBT) μπορούν να ληφθούν υπόψιν για την πρόβλεψη της έκβασης ενός διακανονισμού.

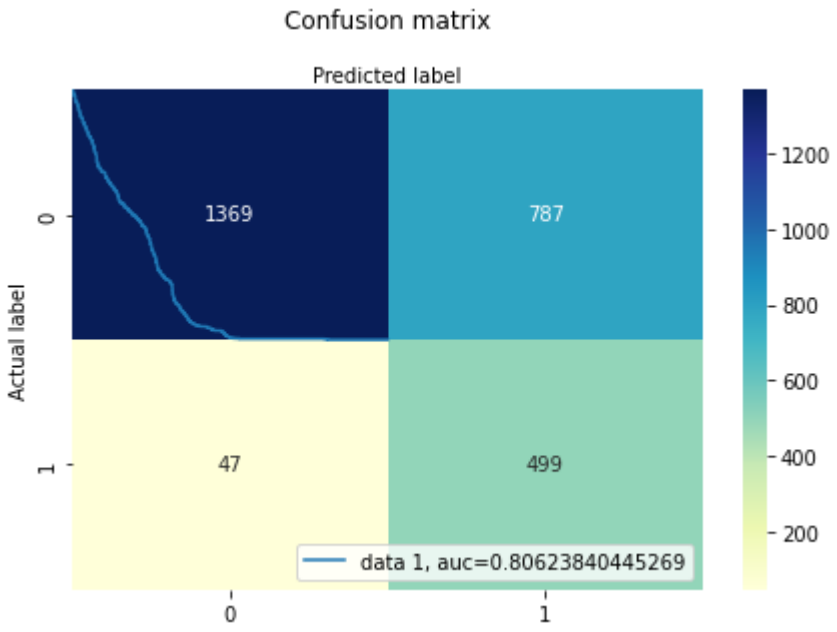
Χρησιμοποιήθηκαν οι ίδιες μεταβλητές και εφαρμόστηκε πάνω στο σύνολο δεδομένων η λογιστική παλινδρόμηση (Εικόνα 22).

Η μεταβλητή target είναι το αν ο οφειλέτης εξόφλησε τον διακανονισμό (1) ή αν δεν τον εξόφλησε (0)

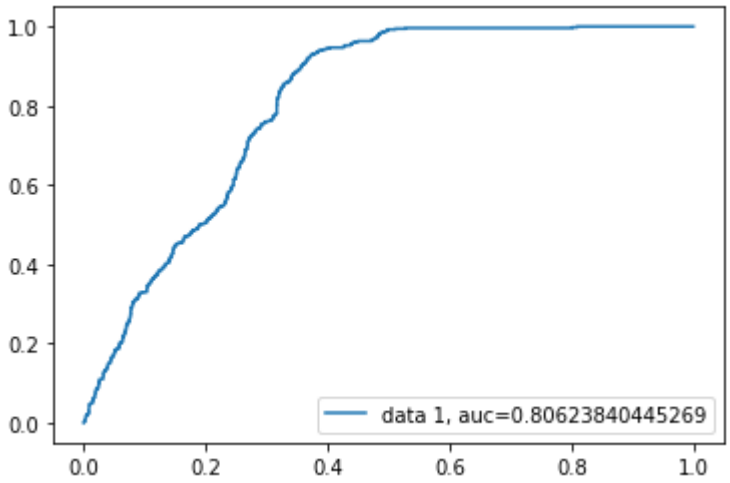
Υπάρχουν πολλές περιπτώσεις που το μοντέλο της λογιστικής παλινδρόμησης δεν λειτούργησε σωστά.

Συγκεκριμένα χαρακτήρισε πολλούς οφειλέτες που δεν εξόφλησαν μερικώς ή συνολικά ον διακανονισμό τους ως καλοπληρωτές Predicted Label: 1

Αντίθετα για την περίπτωση των κακοπληρωτών φαίνεται να λειτουργεί αρκετά καλά με τα συγκεκριμένα δεδομένα.



Εικόνα 22 - Λογιστική Παλινδρόμηση



Εικόνα 23 - Καμπύλη ROC

Accuracy: 0.691339748334567

3.3 ACTIVITIES ΤΑ ΟΠΟΙΑ ΠΡΟΗΓΟΥΝΤΑΙ ΜΙΑΣ ΠΛΗΡΩΜΗΣ ΣΕ ΔΕΔΟΜΕΝΑ ΟΦΕΙΛΩΝ ΚΑΤΑΝΑΛΩΤΙΚΟΥ ΔΑΝΕΙΟΥ

Αρχικά διαχωρίζουμε το dataset , με βάση το αν τα payments αφορούν banking ή energy συνεργασία. Στο σύνολο δεδομένων εμπεριέχονται οι πληροφορίες που προηγούνται μιας πληρωμής αλλά και οι πληρωμές ανά case οι οποίες δεν οδηγούν σε κάποια πληρωμή.

Με αυτόν τον τρόπο ομαδοποιήθηκαν τα activities ανά case τα οποία δεν οδηγούν σε κάποια πληρωμή κι αυτά που οδηγούν σε πληρωμή.

Οι στήλες του dataset είναι:

- Pay_code – ο μοναδικός κωδικός ανά πληρωμή (η κατηγοριοποίηση των activities βασίζεται σε αυτό)
- Actv_name – το όνομα activity
- Co_actv – το πλήθος των φορών που εμφανίστηκε το activity για το συγκεκριμένο pay_code

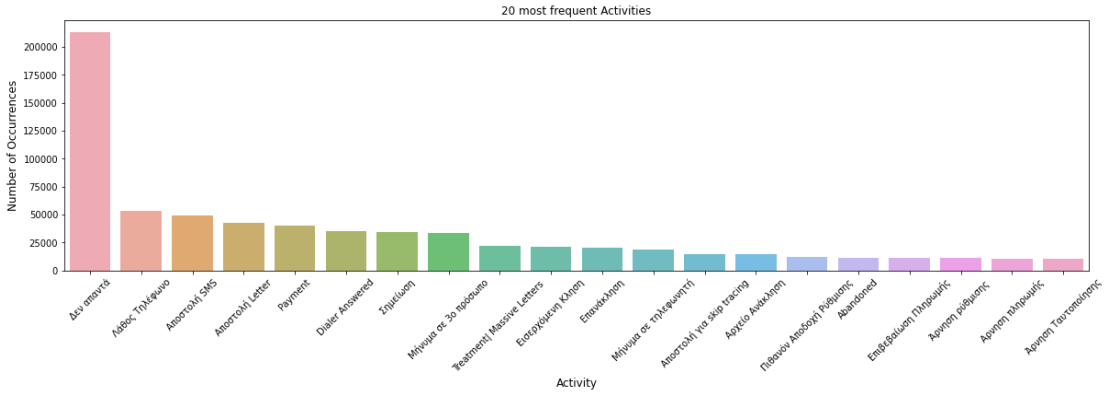
Ο πίνακας γίνεται pivot για να έχουμε μια εγγραφή για κάθε paycode- με την μεταβήτη στόχο να είναι η στήλη Payments.

Όπου στη στήλη Payments έχουμε 0 σημαίνει ότι για την συγκεκριμένη υπόθεση δεν υπάρχει πληρωμή

Όπου στη στήλη Payments έχουμε 1 σημαίνει ότι υπάρχει πληρωμή για αυτό το case και οι υπόλοιπες εγγραφές έχουν το πλήθος των φορών που εμφανίστηκε το συγκεκριμένο activity.

Αρχικά επιλέξαμε να αναλυθεί το σύνολο δεδομένων το οποίο αφορά τις banking συνεργασίες (προϊόντα καταναλωτικών δανείων). Οπτικοποιήθηκαν τα δεδομένα για να εξεταστεί η ποιότητα τους.

Στο σύνολο των δεδομένων παρουσιάζεται η παρακάτω εικόνα (Εικόνα 24) για την αντιστοιχία activities με πλήθος.



Εικόνα 24 - 20 most freq Activities

Το Activity Δεν Απαντά φαίνεται να εμφανίζεται τις περισσότερες φορές στο σύνολο δεδομένων (Εικόνα 24).

Κοιτάζοντας την μεταβλητή στόχο «Payments», τα data είναι αρκετά imbalanced όπως φαίνεται και από την παρακάτω εικόνα:

```
0.0    223368
1.0     40017
```

Εικόνα 25 - Count of cases per class Actions Lead to Payment

Όπου 0 οι εγγραφές χωρίς πληρωμή, 1 εγγραφές με πληρωμή.

Εξετάσαμε τρείς διαφορετικούς αλγόριθμους και την απόδοσή τους στα συγκεκριμένα δεδομένα.

Εκτελώντας τον Naïve Bayes αλγόριθμο λάβαμε τα παρακάτω αποτελέσματα

```
Accuracy: 0.8790751240255138
Training set score: 0.8798
Test set score: 0.8791
Precision: 0.7238532110091743
Recall: 0.32888703626511046
```

Εικόνα 26 - Naive Bayes - Actions Lead to Payment

Ενώ εκτελώντας Δένδρο απόφασης και βάζοντας την παράμετρο class_weight='balanced' για max_depth=3 λαμβάνουμε την εξής απόδοση:

```
Accuracy: 0.9031208869089805
Training set score: 0.9028
Test set score: 0.9031
Precision: 0.6483862144420132
Recall: 0.7904960400166736
```

Εικόνα 27 - Decision Tree - Actions Lead to Payment

Υλοποιούμε τον αλγόριθμο DecisionTreeClassifier.

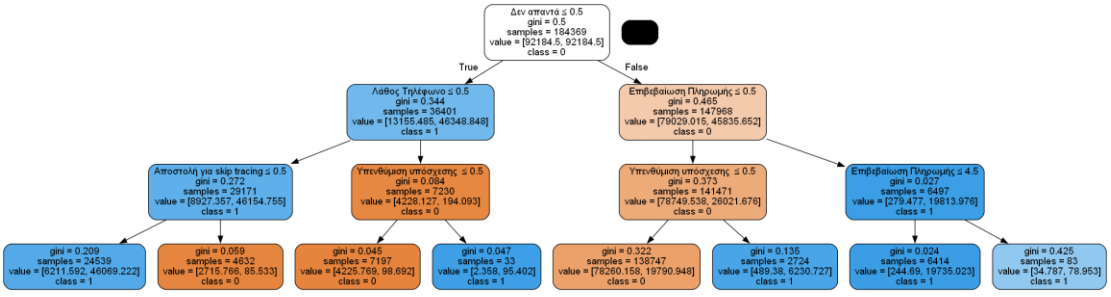
Σύμφωνα με το user guide (https://scikit-learn.org/stable/modules/tree.html#tree), στόχος είναι να δημιουργηθεί ένα μοντέλο που προβλέπει την τιμή μιας μεταβλητής στόχου μαθαίνοντας απλούς κανόνες απόφασης που συνάγονται από τα χαρακτηριστικά των δεδομένων. Ένα δέντρο μπορεί να θεωρηθεί ως μια τμηματικά σταθερή προσέγγιση.

Το DecisionTreeClassifier είναι μια κλάση ικανή να εκτελεί ταξινόμηση πολλαπλών κλάσεων σε ένα σύνολο δεδομένων.

Όπως και με άλλους ταξινομητές, το DecisionTreeClassifier λαμβάνει ως είσοδο δύο πίνακες: έναν πίνακα X, αραιό ή πυκνό, σχήματος (n_samples, n_features) που συγκρατεί τα δείγματα εκπαίδευσης και έναν πίνακα Y ακέραιων τιμών, σχήμα (n_samples,), που περιέχει τις ετικέτες κλάσης για τα δείγματα εκπαίδευσης.

Σε περίπτωση που υπάρχουν πολλές κλάσεις με την ίδια και μεγαλύτερη πιθανότητα, ο ταξινομητής θα προβλέψει την τάξη με τον χαμηλότερο δείκτη μεταξύ αυτών των κλάσεων.

Επιλέγουμε τον αλγόριθμο του Δένδρου απόφασης με τις παραμέτρους που περιγράψαμε παραπάνω. Το δένδρο απόφασης απεικονίζεται (Εικόνα 28):



Εικόνα 28 - Decision Tree Actions Lead to Payment

3.3.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Καταλήγουμε σε πληρωμή όταν:

Έχουμε τουλάχιστον ένα action Δεν Απαντά και

1. Έχουμε κάποιο action επιβεβαίωση πληρωμής αλλά όχι περισσότερα από 4

Δεν έχουμε action Δεν Απαντά και

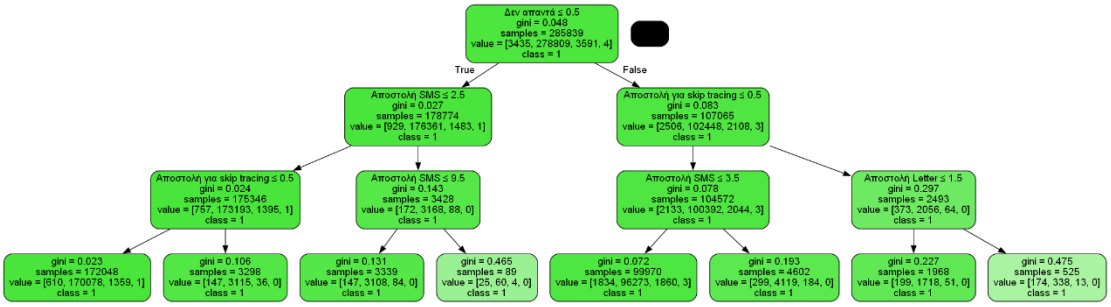
2. Έχουμε κάποιο action λάθος τηλέφωνο και έχουμε υπενθύμιση υπόσχεσης
3. Δεν έχουμε λάθος τηλέφωνο ούτε activity αποστολή για skip tracing

Συνοπτικά μπορούμε να πούμε ότι όταν έχουμε το σωστό τηλέφωνο του οφειλέτη και κανένα activity “ΔΕΝ ΑΠΑΝΤΑ” τότε ο οφειλέτης θα αποδώσει πληρωμή.

Δεν καταλήγουμε σε πληρωμή όταν:

1. Έχει δεν απαντά και δεν έχει ούτε επιβεβαίωση πληρωμής ούτε υπενθύμιση υπόσχεσης
2. Όταν δεν έχει Δεν Απαντά και έχει λάθος τηλέφωνο

Επιχειρήθηκε να γίνει η ίδια ακριβώς ανάλυση για το σύνολο δεδομένων των συνεργασιών της ενέργειας, αλλά λόγω του ότι εκεί οι πληρωμές καταβάλλονται σε πάρα πολλές περιπτώσεις χωρίς καμία ενέργεια του τηλεφωνικού κέντρου , δεν φάνηκε πως τα συγκεκριμένα χαρακτηριστικά που επιλέχθηκαν οδηγούν σε πληρωμή ή διαδραματίζουν κάποιο ρόλο στην καταβολή πληρωμής από τον οφειλέτη (Εικόνα 29) .



Εικόνα 29 - Δένδρο Απόφασης Συνεργασιών Ενέργειας

3.4 ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΔΕΝΔΡΟΥ ΑΠΟΦΑΣΗΣ ΠΑΝΩ ΣΕ ΣΥΝΔΥΑΣΤΙΚΑ ΔΕΔΟΜΕΝΑ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΟΥ RECOVERY RATE – ΣΥΝΕΡΓΑΣΙΕΣ ΕΝΕΡΓΕΙΑΣ

Το σύνολο δεδομένων περιέχει δεδομένα που αφορούν τόσο στοιχεία του προϊόντος, στοιχεία του πελάτη όσο και ενεργειών που έχουν διενεργηθεί από το τμήμα του Collection από το γραφείο μας.

Με βάση αυτά τα στοιχεία έχουμε ένα σύνολο δεδομένων το οποία επιθυμούμε να εξετάσουμε, καθώς τα δεδομένα αφορούν προϊόντα ενέργειας.

Οι στήλες οι οποίες χρησιμοποιήθηκαν είναι οι ακόλουθες και περιέχουν μόνο αριθμητικές τιμές:

```
['Bucket',
 'PlacementYears',
 'MONTHPLACEMENT',
 'Microbalance',
 'Low Tickets',
 'Medium Tickets',
 'High Tickets',
 'Asset',
 'SLow',
 'Low',
 'Medium',
 'High',
 'SHigh',
 'UHigh',
 'SEX',
 'AGE_17-25',
 'AGE_26-35',
 'AGE_36-45',
 'AGE_46-55',
 'AGE_56-65',
 'AGE_66-75',
 'AGE_76-100',
 'POS',
 'NEG',
 'PND',
 'PRM',
 'MSG',
 'LET',
 'SMS',
 'SKIP',
 'UNC',
 'Contacted',
 'PayersFlag',
 'INTEREST_RATE',
 'RECOVERY_RATE_CAT']
```

Εικόνα 30 - Attributes Recovery Rate Classification

Η τελευταία στήλη RECOVERY_RATE_CAT (Εικόνα 29), περιέχει την μεταβλητή στόχο. Το Recovery Rate είναι ο λόγος του συνολικού ποσού πληρωμών μιας υπόθεσης προς το συνολικό ποσό ανάθεσης. Δίνει πάντα έναν αριθμό μεταξύ του 0 και του 1, αν το recovery rate είναι 0 τότε ο οφειλέτης δεν έχει πληρώσει ποτέ για αυτόν τον λογαριασμό, αντίθετα αν είναι 1 τότε σημαίνει ότι εξοφλεί στο 100%. Για να δημιουργήσουμε μια μεταβλητή στόχο και προκειμένου τα δεδομένα να είναι balanced επιλέξαμε την εξής κατηγοριοποίηση για τα προϊόντα ενέργειας.

Στην περίπτωση που το recovery rate <= 0.05 , τότε ανήκει στη κλάση 0

Στην περίπτωση που το recovery 0.81>rate>0.05 , τότε ανήκει στη κλάση 1

Στην περίπτωση που το recovery rate>=0.81 , τότε ανήκει στη κλάση 2

Με αυτόν τον τρόπο τα δεδομένα είναι χωρισμένα με τον εξής τρόπο στο dataset:

```
2    46814
0    46617
1    44994
Name: RECOVERY_RATE_CAT, dtype: int64
```

Εικόνα 31 - Recovery Rate Target Classes

Έτσι ξεχωρίζουμε τους πελάτες σε αυτούς που έχουν μηδενικό Recovery Rate, με αυτούς που έχουν Recovery Rate αλλά κι από αυτούς που έχουν ιδανικό Recovery Rate , κοντά στο 1, βλέπουμε ότι τα δεδομένα είναι balanced (Εικόνα 30).

Δοκιμάστηκαν διαφορετικοί αλγόριθμοι πάνω στα δεδομένα.

Αλγόριθμος Naïve Bayes

```
Accuracy: 0.73845301738669 74
Training set score: 0.7366
Test set score: 0.7385
```

Εικόνα 32 - Recovery Rate Measures Naive Bayes

Αλλά το καλύτερο αποτέλεσμα με τον αλγόριθμο δένδρου απόφασης.

```
Accuracy: 0.7873723752648816
Training set score: 0.7859
```

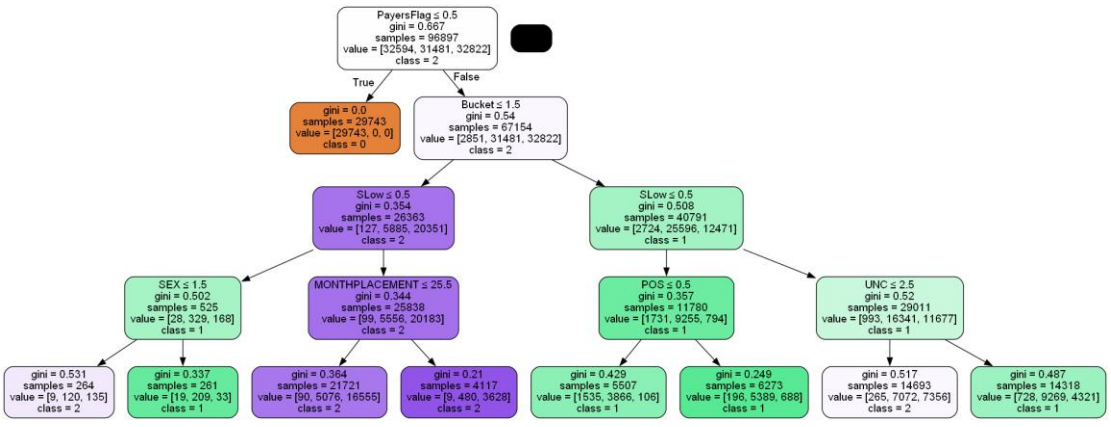
Εικόνα 33 - Recovery Rate Decision Tree

Για να έχουν νόημα και οπτικά τα αποτελέσματα αποφασίσαμε να κάνουμε pruned το δένδρο με βάθος 4. Για μέτρο διαχωρισμού των κλάσεων χρησιμοποιήθηκε ο δείκτης Gini.

Ο δείκτης Gini, υπολογίζει το ποσό της πιθανότητας ενός συγκεκριμένου χαρακτηριστικού που ταξινομείται λανθασμένα όταν επιλέγεται τυχαία. Εάν όλα τα στοιχεία συνδέονται με μία κλάση, τότε αυτή μπορεί να ονομαστεί καθαρή.

Ας αντιληφθούμε το κριτήριο του Δείκτη Gini, όπως και οι ιδιότητες της εντροπίας, ο δείκτης Gini ποικίλλει μεταξύ των τιμών 0 και 1, όπου το 0 εκφράζει την καθαρότητα της ταξινόμησης, δηλαδή όλα τα στοιχεία ανήκουν σε μια καθορισμένη τάξη ή υπάρχει μόνο μία κλάση εκεί. Και το 1 υποδεικνύει την τυχαία κατανομή στοιχείων σε διάφορες κλάσεις. Η τιμή 0,5 του δείκτη Gini δείχνει ίση κατανομή στοιχείων σε ορισμένες κλάσεις.

Έτσι προέκυψε το παρακάτω δέντρο το οποίο μπορούμε να εξηγήσουμε παρακάτω.

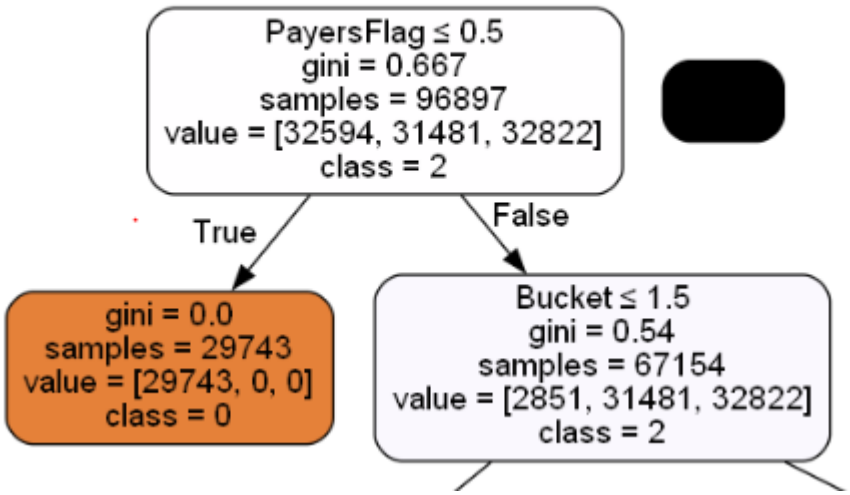


Εικόνα 34 - Recovery Rate Decision Tree Visualization

Αρχικά η ρίζα του δένδρου (Εικόνα 35) περιέχει την μεταβλητή PayersFlag, η συγκεκριμένη μεταβλητή δείχνει με τιμές ‘0’ ή ‘1’ αν ο πελάτης έχει πληρώσει οποιαδήποτε στιγμή στο παρελθόν για την συγκεκριμένη οφειλή, αν δεν έχει πληρώσει τότε το PayersFlag είναι 0, αλλιώς είναι 1.

Προφανώς αυτό κι επηρεάζει άμεσα το recovery rate καθώς αν Payers Flag είναι 1 τότε αυτομάτως και το Recovery Rate θα είναι μεγαλύτερο από το 0.

Επίσης βλέπουμε πως στη ρίζα του δένδρου ο δείκτης Gini=0.667 που δείχνει πως τα δεδομένα δεν ανήκουν μόνο σε μια κλάση.

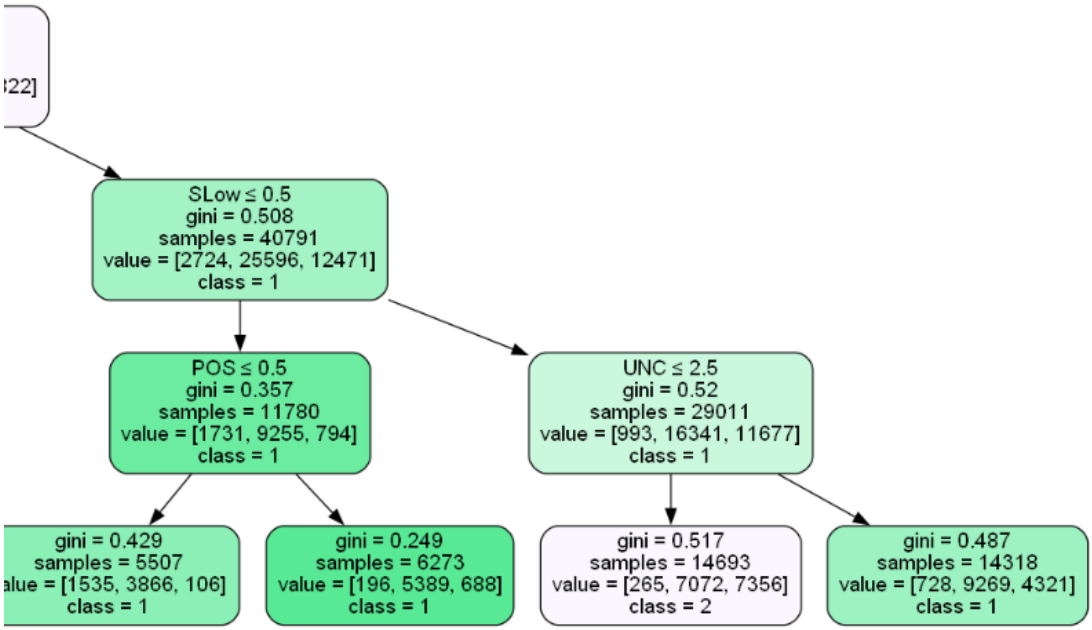


Εικόνα 35 - Recovery Rate Decision Tree Visualization Root

Από εκεί το δέντρο χωρίζεται, κι στο πρώτο επίπεδο του δέντρου βλέπουμε όσες εγγραφές έχουν payersFlag<=0.5 να ταξινομούνται στη κλάση 0, αντίθετα όσες εγγραφές έχουν payers flag=1 οδηγούνται σε έναν επόμενο διαχωρισμό ανάλογα με το Bucket στο οποίο ανήκουν.

Η μεταβλητή Bucket απεικονίζει τους μήνες καθυστέρησης, πόσους μήνες δηλαδή ο οφειλέτης έχει καθυστερήσει την πληρωμή, όσο μεγαλύτερο το Bucket, τόσο περισσότεροι οι μήνες καθυστέρησης. Οι τιμές που παίρνει η μεταβλητή Bucket ανήκουν στο εύρος [0,13] , όπου 13 αφορά αν ο οφειλέτης έχει περισσότερους από 12 μήνες να καταβάλει ποσό σε μια οφειλή.

Οπότε εδώ το δένδρο χωρίζει τις περιπτώσεις αν είναι μικρότερο ή ίσο με Bucket 1 (Ένας μήνας καθυστέρησης) ή αν είναι μεγαλύτερο.



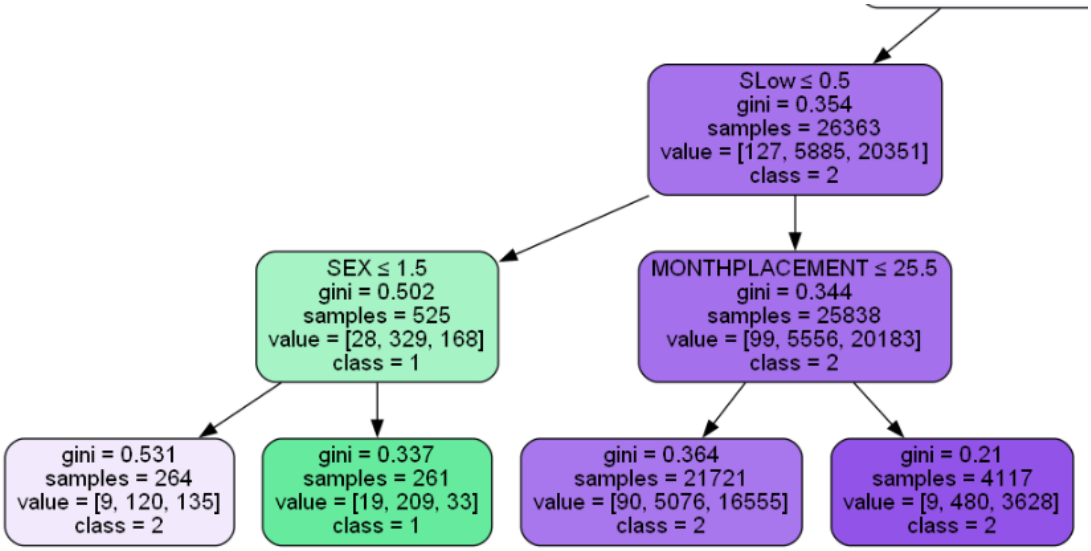
Εικόνα 36 - Recovery Rate Decision Tree Visualization Right Branch

Το δεξί κλαδί οδηγεί στο δεύτερο επίπεδο και στον έλεγχο αν η οφειλή ανήκει στην κατηγορία Slow ή όχι (Εικόνα 36). Η μεταβλητή Slow , περιέχει τις οφειλές που είναι μικρότερες από 300 ευρώ. Στο επίπεδο 3 του δέντρου ο διαχωρισμός γίνεται με βάση τις μεταβλητές POS και UNC. Η μεταβλητή POS (POSITIVE), δείχνει αν έχει γίνει κάποιο action από το τμήμα του Collection με κάποια θετικό αποτέλεσμα (Υπόσχεση Πληρωμής, Επιβεβαίωση Πληρωμής, Πιθανόν Αποδοχή Ρύθμισης), δεν χρειάζεται να αναλύσουμε περισσότερο αυτό το κλαδί καθώς καταλήγει στη κλάση 1, δηλαδή ύπαρξη Recovery Rate έως 80%.

Αντίθετα στη περίπτωση του Uncontacted βλέπουμε πως αν ο οφειλέτης δεν είναι περισσότερες από 2 φορές Uncontacted(κλήθηκε αλλά δεν αποκρίθηκε) τότε υπάρχει πιθανότητα να ανήκει στη κλάση 2 , υψηλό recovery rate.

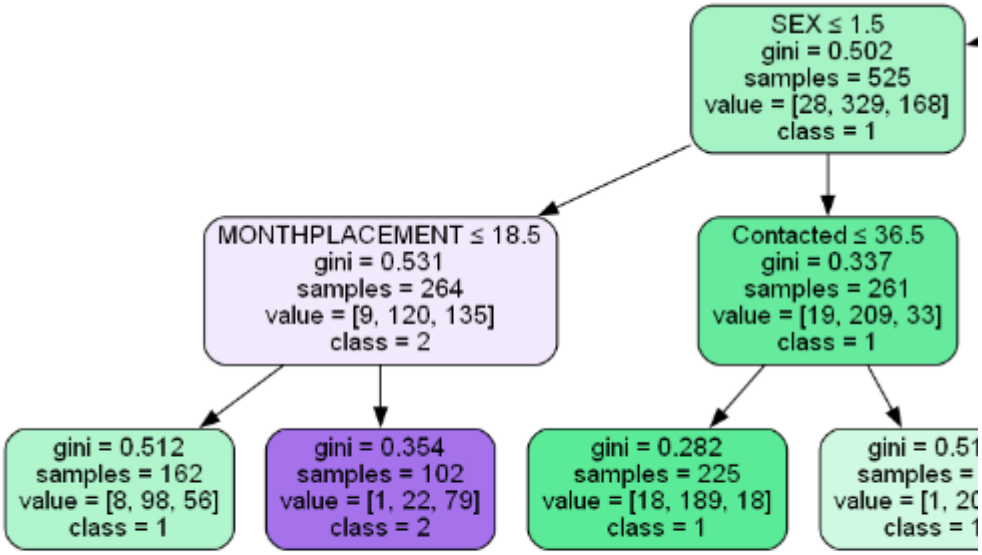
Στο αριστερό κλαδί (Εικόνα 37), αν δηλαδή το Bucket είναι μικρότερο από 2 (το πολύ ένας μήνας καθυστέρησης), τότε πάλι η μεταβλητή Slow (οφειλές<=300 ευρω), ορίζει τον επόμενο διαχωρισμό. Εδώ αν ανήκουν στην κατηγορία SLOW, τότε το RecoveryRate θα είναι υψηλό.

Αντίθετα αν ανήκουν σε άλλες κατηγορίες οφειλής (>300 ευρώ), τότε φαίνεται το φύλλο να διαδραματίζει κάποιο ρόλο. Για την μεταβλητή SEX, έχουμε τις εξής τιμές : Άνδρας=1, Γυναίκα=0. Αν ο οφειλέτης είναι γυναίκα, υπάρχουν περισσότερες πιθανότητες να ανήκει στην κλάση 2 (υψηλό recovery rate), αν και ο διαχωρισμός είναι πάρα πολύ ισορροπημένος (Gini=0.53).



Εικόνα 37 - Recovery Rate Decision Tree Visualization Left Branch

Θα μπορούσαμε να προχωρήσουμε για το συγκεκριμένο κλαδί του δέντρου (Εικόνα 38) για να δούμε αν χρειάζεται κάποια παραπάνω ανάλυση. Με βάθος 5 παρατηρούμε πως η μεταβλητή Monthplacement, πόσους μήνες είναι ανατεθειμένη η υπόθεση στο γραφείο μας, για τις περιπτώσεις που ο οφειλέτης είναι γυναίκα, αν οι μήνες ανάθεσης είναι περισσότεροι από 18, δηλαδή περισσότερο από 1,5 χρόνο, τότε υπάρχει μεγάλη πιθανότητα το Recovery Rate να είναι υψηλό.



Εικόνα 38 - Recovery Rate Decision Tree Visualization Left Branch classes 1-2

3.4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Μπορούμε να καταλήξουμε στα εξής συμπεράσματα από την ανάλυση:

- Το recovery rate επηρεάζεται από το αν ο πελάτης έχει πληρώσει οποιαδήποτε στιγμή, αν δεν έχει πληρώσει ποτέ θα είναι μηδενικό.
- Αν οι μήνες καθυστέρησης πληρωμής υπερβαίνουν τους 2, τότε σχεδόν πάντα οδηγούμαστε σε ένα χαμηλό ή μέτριο recovery rate.
- Αντίθετα αν οι μήνες καθυστέρησης είναι το πολύ 1, και το ύψος του χρέους είναι μεγάλο (περισσότερα από 300 ευρώ) και ο οφειλέτης είναι άντρας θα οδηγηθούμε σε μέτριο ή χαμηλό recovery rate.
- Αν οι μήνες καθυστέρησης είναι το πολύ 1, και το ύψος του χρέους είναι μικρό, λιγότερα από 300 ευρώ, τότε θα έχουμε υψηλό recovery rate > 80%.

3.5 ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΔΕΝΔΡΟΥ ΑΠΟΦΑΣΗΣ ΠΑΝΩ ΣΕ ΣΥΝΔΥΑΣΤΙΚΑ ΔΕΔΟΜΕΝΑ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΟΥ ΤΗΡΗΣΗΣ ΔΙΑΚΑΝΟΝΙΣΜΟΥ (ΚΕΡΤ) – ΤΡΑΠΕΖΙΚΕΣ ΣΥΝΕΡΓΑΣΙΕΣ

Το σύνολο δεδομένων περιέχει δεδομένα που αφορούν τόσο στοιχεία του προϊόντος, στοιχεία του πελάτη όσο και ενεργειών που έχουν διενεργηθεί από το τμήμα του Collection από το γραφείο μας.

Με βάση αυτά τα στοιχεία έχουμε ένα σύνολο δεδομένων το οποία επιθυμούμε να εξετάσουμε, καθώς τα δεδομένα αφορούν τραπεζικά προϊόντα.

Οι στήλες οι οποίες χρησιμοποιήθηκαν είναι οι ακόλουθες και περιέχουν μόνο αριθμητικές τιμές:

```
['Bucket',
 'MONTHPLACEMENT',
 'Microbalance',
 'Low Tickets',
 'Medium Tickets',
 'High Tickets',
 'Asset',
 'SLow',
 'Low',
 'Medium',
 'High',
 'SHigh',
 'UHigh',
 'SEX',
 'regionNA',
 'attiki',
 'peloponissos',
 'kentriki ellada,Sporades, Lefkada',
 'Thesssalia, Ipeiros Kerkyra',
 'Ditiki kai kentriki makedonia',
 'makedonia',
 'Kriti',
 'Aigaio',
```

Εικόνα 39 - Kept Settlements Attributes 1

```
'AGE_17-25',
'AGE_26-35',
'AGE_36-45',
'AGE_46-55',
'AGE_56-65',
'AGE_66-75',
'AGE_76-100',
'POS',
'NEG',
'PND',
'PRM',
'MSG',
'LET',
'SMS',
'SKIP',
'UNC',
'Contacted',
'PayersFlag',
'INTEREST_RATE',
'RECOVERY_RATE',
'SettlementKept']
```

Εικόνα 40 - Kept Settlements Attributes 2

Το σύνολο δεδομένων περιέχει μόνο στοιχεία οφειλών για τις οποίες έχει γίνει διακανονισμός (Εικόνα 40).

Η μεταβλητή στόχος για την ταξινόμηση είναι η μεταβλητή SettlementKept, η οποία έχει τις τιμές 0 ή 1 ανάλογα με το αν ο διακανονισμός τηρείται ή όχι για την κάθε υπόθεση.

Δοκιμάστηκαν διαφορετικοί αλγόριθμοι πάνω στα δεδομένα (Vector Machine, Naïve Bayes), αλλά το καλύτερο αποτέλεσμα με τον αλγόριθμο δένδρου απόφασης.

Scores για δένδρο απόφασης

```
Accuracy: 0.8502386634844868
Training set score: 0.8522
Test set score: 0.8502
Precision: 0.7850045167118338
Recall: 0.9852607709750567
```

Εικόνα 41 - Kept Settlements Decision Tree scores

Scores Naïve Bayes

```
Accuracy: 0.8031026252983293
Training set score: 0.8082
Test set score: 0.8031
Precision: 0.7857142857142857
Recall: 0.8605442176870748
```

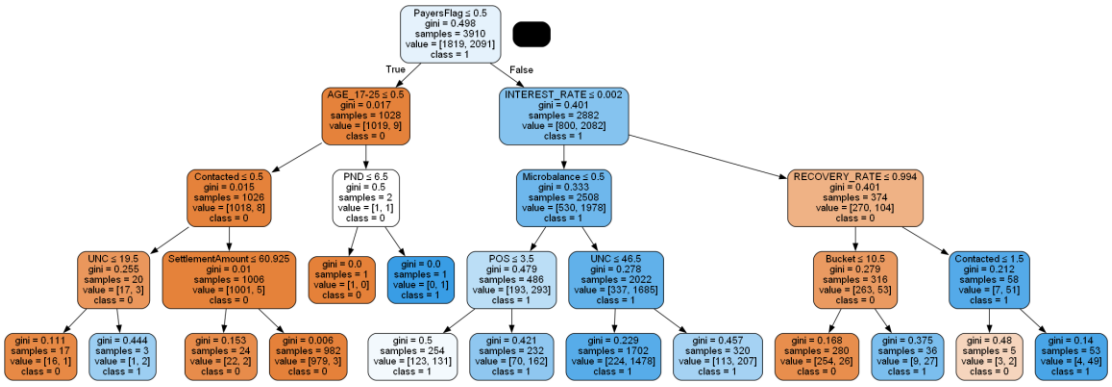
Εικόνα 42 - Kept Settlements Naive Bayes scores

Scores Vector Machine

```
Accuracy: 0.8204057279236276
Training set score: 0.8192
Test set score: 0.8204
Precision: 0.8195819581958196
Recall: 0.844671201814059
```

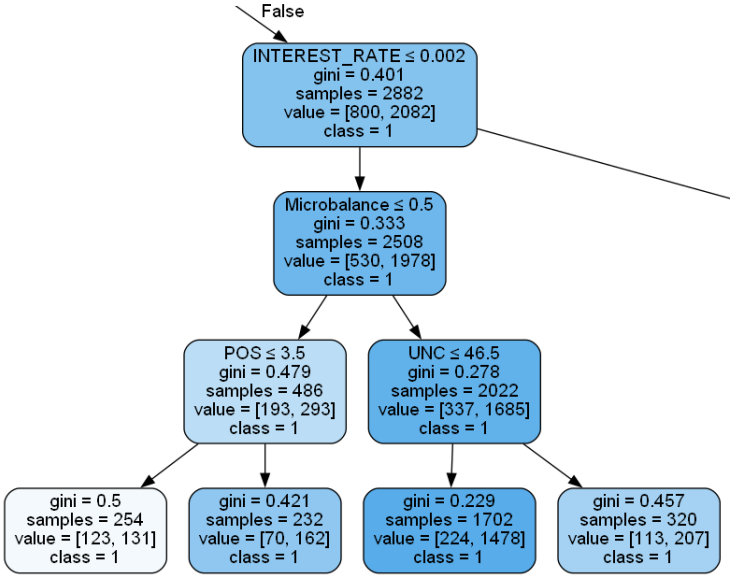
Εικόνα 43 - Kept Settlements Vector Machine scores

Τελικά προτιμήθηκε ο αλγόριθμος δένδρου απόφασης για την ανάλυση της τήρησης διακανονισμών σε banking δεδομένα.



Εικόνα 44 - Kept Settlement Decision Tree Visualization

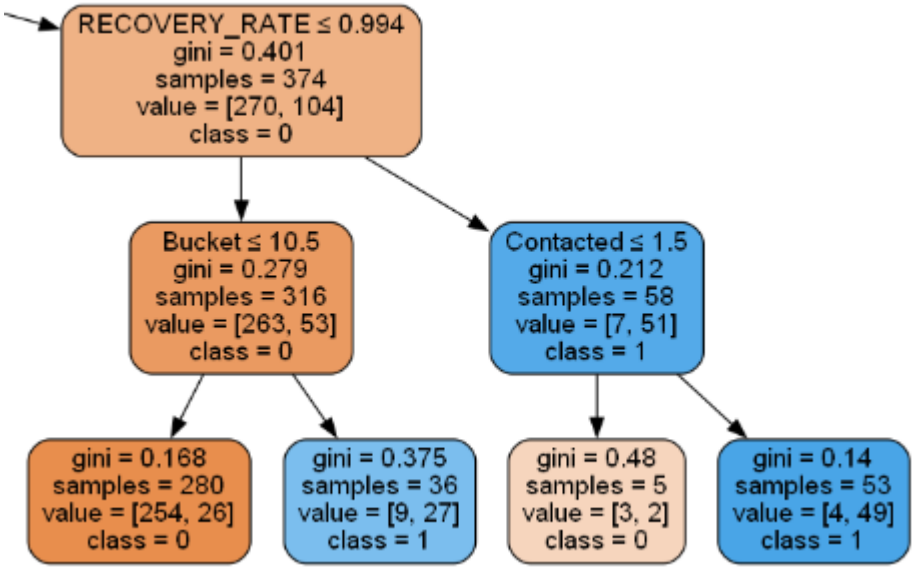
Η ρίζα του δέντρου (Εικόνα 44) αφορά την μεταβλητή PayersFlag, η οποία διαχωρίζει το δέντρο ανάλογα με το αν ο οφειλέτης έχει έστω καταβάλει οποιοδήποτε ποσό για τη συγκεκριμένη πληρωμή. Αν ο οφειλέτης έχει καταβάλει κάποιο ποσό, οδηγούμαστε στο δεξιό κλαδί του δέντρου. Εδώ ο διαχωρισμός αφορά τη μεταβλητή Interest Rate , αν το Interest Rate είναι μηδενικό, δηλαδή αν το ποσό δεν είναι τοκισμένο τότε πάντα θα οδηγηθούμε σε τήρηση διακανονισμού.



Εικόνα 45 - Kept Settlement Decision Tree Visualization Right Branch-1

Στη περίπτωση που το ποσό είναι τοκισμένο (Εικόνα 45), τα δείγματα μοιράζονται στις διαφορετικές κλάσεις. Έτσι ακολουθώντας από το επίπεδο 2, το δεξί κλαδί (τοκισμένο ποσό), το δέντρο χωρίζεται ανάλογα με το recovery rate. Σε αυτή την περίπτωση αν το recovery rate είναι μικρότερο από 1 (οι πληρωμές δεν είναι ίσες με το ποσό οφειλής), τότε ο διαχωρισμός θα γίνει με βάση τους μήνες καθυστέρησης. Αν οι μήνες καθυστέρησης είναι λιγότεροι από 10, τότε ο διακανονισμός δεν θα τηρηθεί, αντίθετα ο διακανονισμός υπάρχει πιθανότητα να τηρηθεί.

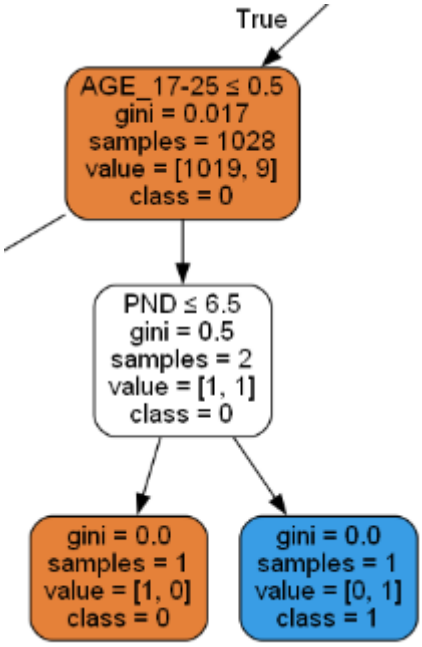
Αν το recovery rate είναι μεγάλο (Εικόνα 46) και έχουμε επικοινωνία με τον οφειλέτη , ο διακανονισμός θα τηρηθεί, αντίθετα δεν θα τηρηθεί.



Εικόνα 46 - Kept Settlement Decision Tree Visualization Right Branch-2

Πηγαίνοντας στη ρίζα του δέντρου (Εικόνα 44), ακολουθούμε την περίπτωση που ο οφειλέτης δεν έχει αποδώσει ποτέ πληρωμή.

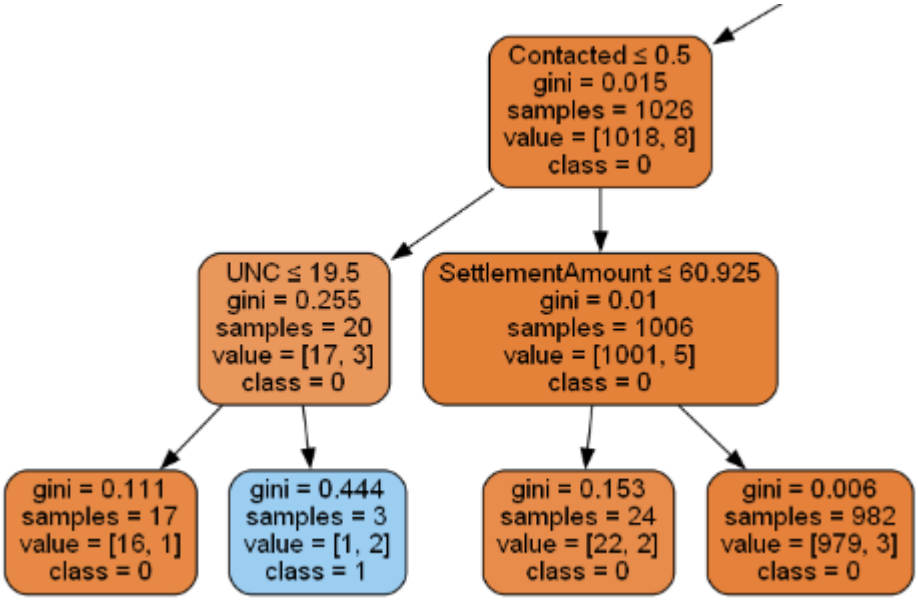
Αν ανήκει στην ηλικιακή κατηγορία 17-25 (Εικόνα 47), η πιθανότητα του να τηρήσει τον διακανονισμό είναι 50-50.



Εικόνα 47 - Kept Settlement Decision Tree Visualization Left Branch-1

Αντίθετα αν ανήκει σε άλλη ηλικιακή κατηγορία κι είναι Contacted, τότε θα ταξινομηθεί στη κλάση 0, δηλαδή δεν θα τηρήσει τον διακανονισμό.

Υπάρχει μια επιπλέον πιθανότητα ο διακανονισμός να τηρηθεί (Εικόνα 48), αν δεν έχουμε καμία επικοινωνία με τον οφειλέτη και το uncontacted είναι μεγάλο (>20 uncontacted ενέργειες)



Εικόνα 48 - Kept Settlement Decision Tree Visualization Left Branch-2

3.5.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Από την παραπάνω ανάλυση μπορούν να προκύψουν τα παρακάτω συμπεράσματα:

- Η ένδειξη πληρωμής ενός πελάτη διαχωρίζει σημαντικά τις οφειλές ανάλογα με το αν θα τηρήσουν ή όχι έναν διακανονισμό.
- Όταν ο οφειλέτης έχει πληρωμή και το ποσό που οφείλει δεν είναι τοκισμένο, τότε θα τηρήσει τον διακανονισμό
- Όταν ο οφειλέτης έχει πληρωμή και το ποσό που οφείλει είναι τοκισμένο, τότε οι πιθανότητες είναι μοιρασμένες.
- Αν ισχύει το παραπάνω ,τότε ισχύει πως μικρό recovery rate και μικρό Bucket θα οδηγήσουν σε μη τήρηση του διακανονισμού, ενώ μεγάλο recovery rate και πελάτης με ενέργειες επικοινωνίας θα τηρήσει τον διακανονισμό.
- Όταν ο πελάτης δεν έχει πληρωμή κι ανήκει στην Ηλικιακή κατηγορία 17-25, η πιθανότητα να τηρήσει τον διακανονισμό είναι μοιρασμένη.
- Όταν ο πελάτης δεν έχει πληρωμή κι δεν ανήκει στην Ηλικιακή κατηγορία 17-25, η πιθανότητα να τηρήσει τον διακανονισμό υπάρχει μόνο στη περίπτωση που δεν έχουν γίνει πολλές αποτυχημένες απόπειρες επικοινωνίας μαζί του.

3.5.2 ΕΦΑΡΜΟΓΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ

Εφαρμόσαμε τον παραπάνω αλγόριθμο ως μέθοδο πρόβλεψης σε πραγματικά δεδομένα για συγκεκριμένο χαρτοφυλάκιο. Συγκεκριμένα επιλέχθηκαν 55 οφειλέτες οι οποίοι μέχρι και τις 24/01/2023 δεν είχαν πληρώσει καμία οφειλή τους. Ο αλγόριθμος έδωσε την παρακάτω πρόβλεψη:

49 Οφειλέτες θα τηρήσουν τον διακανονισμό ως και την 1^{η} Φερουαρίου

6 Οφειλέτες θα αθετήσουν την πληρωμή τους για τον διακανονισμό.

Τρέχοντας τον αλγόριθμο στις 2/2/2023 προέκυψαν τα αποτελέσματα:

32 Οφειλέτες είχαν τηρήσει τον Διακανονισμό τους

17 Οφειλέτες αθέτησαν τον διακανονισμό

Υπολογίστηκαν τα μέτρα Precision, Recall και Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Accuracy: 0,690909 , Precision: 0,653061 και Recall: 1

Σχετικά με τις αστοχίες του αλγορίθμου έγινε η παρακάτω έρευνα από την οποία προέκυψαν τα παρακάτω αποτελέσματα.

Η πρόβλεψη του αλγορίθμου ήταν πιο «αισιόδοξη» από όσο θα έπρεπε για τους πελάτες που θα αποδίδαν πληρωμή, θεώρησε ότι 49 οφειλέτες θα τηρούσαν τον διακανονισμό, όταν στην πραγματικότητα αυτοί ήταν 32. Οι οφειλέτες αυτοί για να θεωρείται ότι τηρήσαν τον διακανονισμό θα έπρεπε να καταβάλλουν ένα ποσό μέχρι και την τελευταία μέρα του μήνα (31/1/2023). Αυτό δεν συνέβη για όλες τις θετικές προβλέψεις (32/49).

Ανάμεσα σε αυτούς τους 49, υπήρξαν και 5 οι οποίοι πλήρωσαν εκπρόθεσμα την δόση τους (τον επόμενο μήνα – Φεβρουάριος 2023). Φαίνεται δηλαδή ότι ο αλγόριθμος ήταν ακόμα πιο κοντά όσον αφορά το αν ο οφειλέτης θα πληρώσει, έστω και με κάποια καθυστέρηση (εώς και 15 μέρες σε κάποιες περιπτώσεις).

CASE_C	Label_Predicti	Actual Payment	
878262	1	0	εξοφλησε
901609	1	0	τελευταία πληρωμή 11/2022
549630	1	0	καταβάλει μια δόση κάθε δίμηνο με βάση την ιστορικότητα του
900471	1	0	μόνο μια πληρωμή 30/11/2022 - 50 ευρώ
900383	1	0	μόνο μια πληρωμή 3/11/2022 - 50 ευρω
551880	1	0	τελευταία πληρωμή 11/2022 ενώ πλήρωνε τακτικά
902104	1	0	μόνο μια πληρωμή 3/11/2022
895320	1	0	
954017	1	0	μονο μια πληρωμή 30/12/2022
552533	1	0	πολλές παρελθοντικές πληρωμές, υπάρχει επικοινωνία - εχει υποσχεθεί να καταθέση ξανα τον Μάρτιο
732082	1	0	τελευταία πληρωμή 12/2022
895206	1	0	τελευταία πληρωμή 11/2022

Εικόνα 49 - Λάθος εκτιμήσεις

Μια παράμετρος που φαίνεται να διαδραματίζει κάποιο ρόλο και δεν λήφθηκε υπόψιν είναι το πλήθος των παρελθοντικών πληρωμών. 4/12 από αυτούς που δεν πλήρωσαν είχαν μόνο μια πληρωμή στο ιστορικό τους (Εικόνα 49) , δηλαδή δεν αποτυπώνεται σε βάθος χρόνου η συνέπεια τους. Αντίθετα υπάρχουν δυο περιπτώσεις «καλοπληρωτών» που φαίνεται να μην έχουν πληρώσει. Ο ένας λανθασμένα μπήκε στην ανάλυση καθώς είχε εξοφλήσει και δεν υπήρχε επόμενη δόση, ενώ ο δεύτερος είχε επικοινωνία με το γραφείο και ενημέρωσε ότι θα πληρώσει εκ νέου τον Μάρτιο.

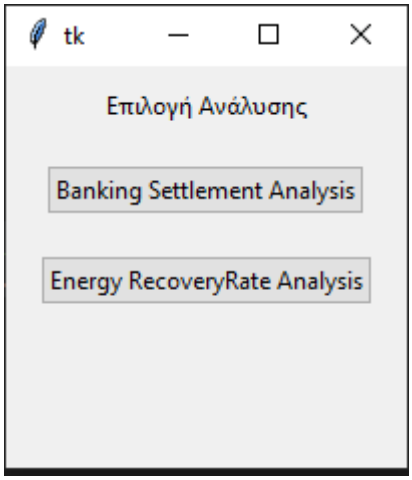
Τέλος υπάρχουν κάποιοι οι οποίοι έχουν σταματήσει περίπου 2 μήνες να πληρώνουν και να μην απαντούν στις κλήσεις παρότι το ιστορικό των καταθέσεων τους είναι πλούσιο.

3.6 ΔΗΜΙΟΥΡΓΙΑ ΕΦΑΡΜΟΓΗΣ ΣΕ ΓΛΩΣΣΑ PYTHON

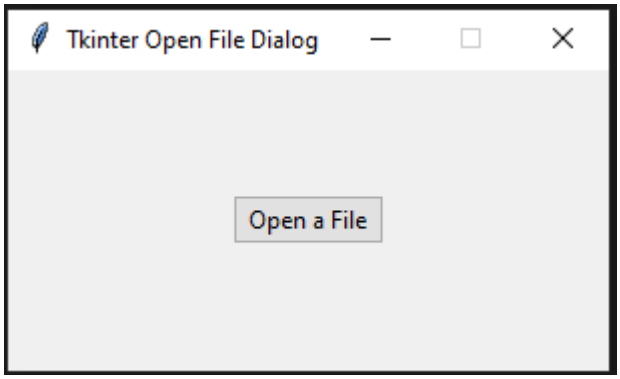
Για την αυτοματοποίηση του τρόπου πρόβλεψης του παραπάνω αλγορίθμου δημιουργήθηκε μια desktop εφαρμογή με τη γλώσσα προγραμματισμού python. Η εφαρμογή έχει σαν σκοπό την δημιουργία ενός report , το οποίο θα παρουσιάζει την πρόβλεψη των μοντέλων που υλοποιήθηκαν, σε συγκεκριμένες υποθέσεις. Με αυτόν τον τρόπο δίνεται η δυνατότητα στον διαχειριστή των υποθέσεων να ιεραρχήσει τις υποθέσεις που φαίνεται να παρουσιάζουν περισσότερες πιθανότητες για ένα θετικό αποτέλεσμα σε αντίθεση με τις υπόλοιπες υποθέσεις. Η εφαρμογή κατασκευάστηκε ώστε να τρέχουν οι αλγόριθμοι που υπολογίζουν το Recovery Rate για τις συνεργασίες της Ενέργειας κι αυτές που προβλέπουν το Kept Settlement για τις τραπεζικές συνεργασίες.

3.6.1 ΛΕΙΤΟΥΡΓΙΑ ΕΦΑΡΜΟΓΗΣ

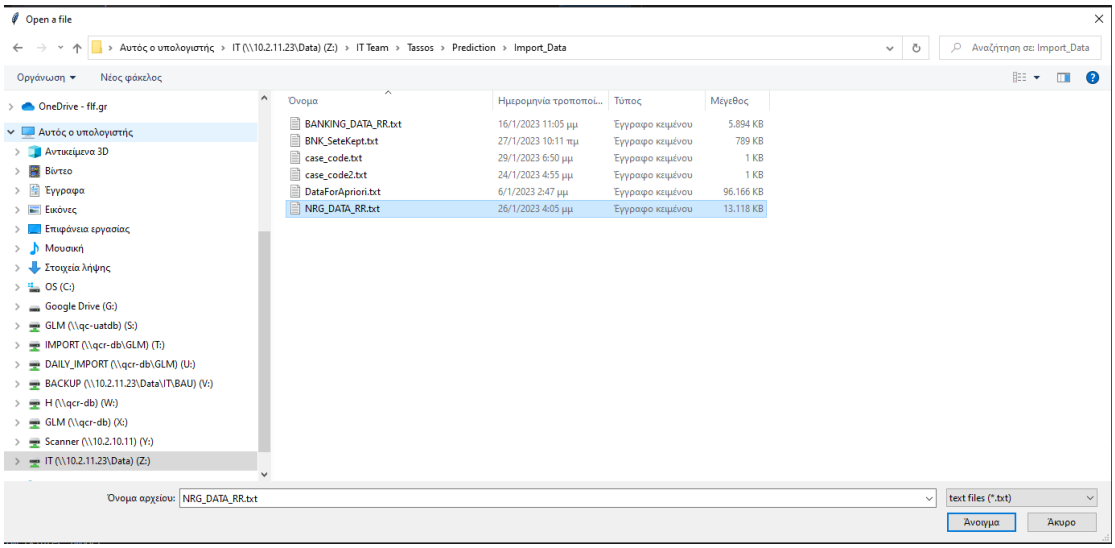
Η αρχική εικόνα της εφαρμογής μας αφήνει να επιλέξουμε μεταξύ της ανάλυσης «Banking Settlement» και “Energy Recover Rate” (Εικόνα 50). Οποιαδήποτε επιλογή και να κάνουμε η επόμενη οθόνη θα μας οδηγήσει στην επιλογή του αρχείου (σε μορφή txt), το οποίο περιλαμβάνει τις υποθέσεις για τις οποίες θα γίνει η ανάλυση.



Εικόνα 50 - Λειτουργία Εφαρμογής 1

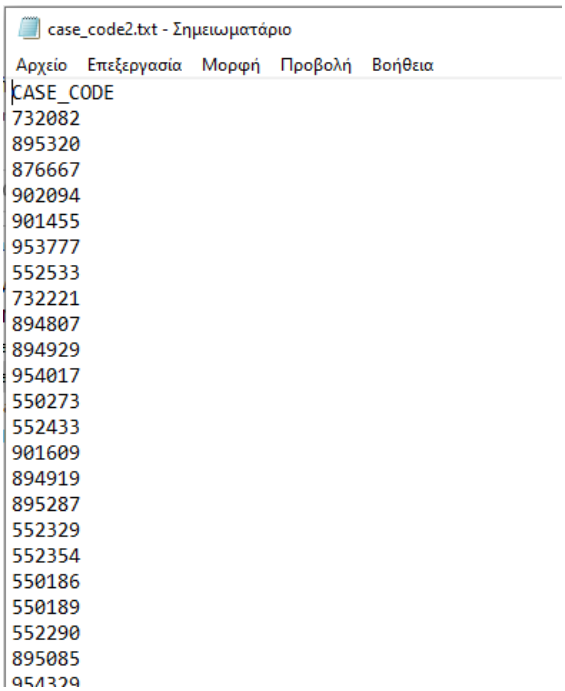


Εικόνα 51 - Άνοιγμα Φακέλου



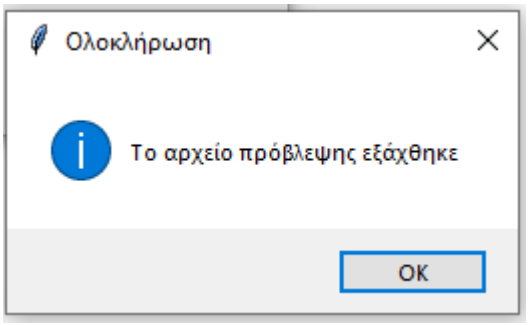
Εικόνα 52 - Επιλογή Αρχείου

Το αρχείο πρέπει να είναι της μορφής .txt, και να περιέχει μια στήλη με όλους τους κωδικούς των υποθέσεων (Εικόνα 53).



Εικόνα 53 - Μορφή Αρχείου Εισαγωγής

Μετά την επιλογή του αρχείου το πρόγραμμα εμφανίζει μήνυμα στον χρήστη για να ανοίξει το αρχείο πρόβλεψης (Εικόνα 54).



Εικόνα 54 - Μήνυμα Εξαγωγής

Το αρχείο πρόβλεψης εξάγεται σε συγκεκριμένο φάκελο που έχουν πρόσβαση όλοι οι χρήστες.

ASE_COD	el_Prediction
549629	1
878262	1
901609	1
552433	1
795982	1
549630	1
552072	1
551669	1
549900	1
900471	1
894807	1
550186	1
894919	1
900668	0
900147	1
878123	0
878269	1
878581	1
549766	1
552354	1

Εικόνα 55 - Αρχείο Εξαγωγής

Στο αρχείο Πρόβλεψης ο χρήστης βλέπει τον αριθμό υπόθεσης και δίπλα την πρόβλεψη της κλάσης στην οποία θα ανήκει η υπόθεση (Εικόνα 55).

4 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη παρούσα εργασία ασχοληθήκαμε με την εξόρυξη γνώσης σε δεδομένα ληξιπρόθεσμων οφειλών και δανείων. Κάναμε μια παρουσίαση των βασικών εννοιών καθώς και των διαφορετικών μορφών και των μοντέλων που χρησιμοποιούνται για την εξόρυξη των δεδομένων. Έγινε επίσης μια αναφορά και στον χρηματοπιστωτικό τομέα καθώς και στον τρόπο λειτουργίας των γραφείων που αναλαμβάνουν την συλλογή χρέους.

Σκοπός της εργασίας ήταν να προσφέρει σε εταιρεία συλλογής χρέους νέα εργαλεία πρόβλεψης και ανάλυσης για την ιεράρχηση της συλλογής οφειλών με υψηλή πιθανότητα αποπληρωμής ή απλής πληρωμής έναντι οφειλών με μικρότερες πιθανότητες αποπληρωμής. Εμφανίζεται στον αναγνώστη όλη η πορεία της διαδικασίας που οδήγησαν στην υλοποίηση συγκεκριμένων μοντέλων καθώς και τα αποτελέσματα τους.

Τα παραπάνω εργαλεία που αναπτύχθηκαν εφαρμόστηκαν και σε πραγματικό χρόνο για την πρόβλεψη αποπληρωμής. Έχουν την δυνατότητα να εφαρμοστούν σε εταιρείες συλλογής χρέους με παρόμοιες πηγές δεδομένων και μπορούν να ενταχθούν στην χάραξη και στον σχεδιασμό της στρατηγικής μιας εταιρείας συλλογής χρέους. Μπορούν επίσης να δοκιμαστούν και νέες τεχνικές όπως νευρωνικά δίκτυα για την επέκταση των αναλύσεων και με άλλες τεχνικές εξόρυξης γνώσης.

Μπορούν επίσης να επεκταθεί και η εφαρμογή ανάλυσης με περισσότερες επιλογές στους αλγόριθμους πρόβλεψης καθώς και με μεγαλύτερη εξειδίκευση στα σύνολα δεδομένων τα οποία θέλουμε να επεξεργαστούμε.

Εκτός από την πρόβλεψη παρέχουν στην εταιρεία μέσω της δυνατότητας οπτικοποίησης των δένδρων απόφασης, την καλύτερη γραφική κατανόηση των αποτελεσμάτων σε έναν χρήστη.

Η ενσωμάτωση των παραπάνω τεχνολογιών είναι απαραίτητη για την βελτίωση της ανταγωνιστικότητας των επιχειρήσεων καθώς και για τον ορθότερο καθορισμό των στόχων της επιχείρησης. Προϋπόθεση αποτελεί η κατάλληλη εκπαίδευση του προσωπικού αλλά και η αλλαγή διαδικασιών για την ενσωμάτωση της νέας τεχνολογίας και την αξιοποίηση της.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Tan, P.-N., Steinbach, M., Kumar, V. Introduction to Data Mining. Addison-Wesley, Reading. (2005)

[2] Meta S. Brown. Data Mining For Dummies. John Wiley & Sons. (2014)

[3] Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. Mining of Massive Datasets. (2019)

[4] Florin Gorunescu. Data Mining Concepts, Models and Techniques. Springer-Verlag Berlin Heidelberg. (2011)

[5] Dr. Sachin Kashyap, Dr. Abhishek Pandey, Dr. Sanjeev Gupta. APPLICATION OF DATA MINING IN BANKING AND FINANCE. (2019)

[6] Chen, W., Xiang, G., Liu, Y., & Wang, K. Credit risk Evaluation by hybrid data mining technique. Systems Engineering Procedia. (2012)

[7] D'Haen, J., Van den Poel, D., & Thorleuchter, D. Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. Expert systems with applications. (2013)

[8] Marple, M. and Zimmerman. Customer Retention Strategy. (1999)

[9] Regulation of Debt Collection in Europe: Understanding Informal Debt Collection Practices, Cătălin Gabriel Stănescu, Taylor & Francis. (2022)

[10] Tuğçe Ayhan, Tamer Uçar. Determining customer limits by data mining methods in credit allocation process. (2022)

[11] Sung HoHa , RamayyaKrishnan. Predicting repayment of the credit card debt. (2010)

[12] Amir M. Hormozi, Stacy Giles. DATA MINING:A COMPETITIVE WEAPON FOR BANKING AND RETAIL INDUSTRIES. (2004)

[13] Ahmad Nadali, Hamid Eslami Nosratabadi. Credit Assessment of Bank Customers by a Fuzzy Expert System Based on Rules Extracted from Association Rules. (2012)

[14] Hilala Jafarova. Applying K-Means Clustering Algorithm Using Oracle Data Mining to Banking Data. (2015)

[15] Andrii Kaminskyi, Maryna Nehrey. CLUSTERING APPROACH TO ANALYSIS OF THE CREDIT RISK AND PROFITABILITY FOR NONBANK LENDERS

[16] https://en.wikipedia.org/wiki/Debt_collection

[17] https://www.ibm.com/topics/linear-regression