ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΩΝ-
ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

«ΜΕΘΟΔΟΙ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΠΟΛΥ-ΚΡΙΤΗΡΙΑΣ ΑΝΑΛΥΣΗΣ ΑΠΟΦΑΣΕΩΝ ΓΙΑ ΤΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΩΝ ΔΙΑΔΙΚΑΣΙΩΝ ΕΙΣΠΡΑΞΗΣ ΛΗΞΙΠΡΟΘΕΣΜΩΝ ΟΦΕΙΛΩΝ»

Διπλωματική Εργασία για το Μεταπτυχιακό Πρόγραμμα Σπουδών

η παρούσα Εργασία εκπονήθηκε
ως μερική ικανοποίηση των απαιτήσεων για την απόκτηση
του αντίστοιχου τίτλου σπουδών στην
Στατιστική και Αναλογιστικά - Χρηματοοικονομικά Μαθηματικά

ΔΑΜΚΑΛΗ ΜΑΡΙΑ - ΕΛΕΝΗ

ΣΕΠΤΕΜΒΡΙΟΣ 2023

ΣΑΜΟΣ

ΔΑΜΚΑΛΗ ΜΑΡΙΑ – ΕΛΕΝΗ

«ΜΕΘΟΔΟΙ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΠΟΛΥ-ΚΡΙΤΗΡΙΑΣ ΑΝΑΛΥΣΗΣ ΑΠΟΦΑΣΕΩΝ ΓΙΑ ΤΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΩΝ ΔΙΑΔΙΚΑΣΙΩΝ ΕΙΣΠΡΑΞΗΣ ΛΗΞΙΠΡΟΘΕΣΜΩΝ ΟΦΕΙΛΩΝ»

ΣΕΠΤΕΜΒΡΙΟΣ 2023

Διπλωματική Εργασία για το Μεταπτυχιακό Πρόγραμμα Σπουδών

Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών

Συγγραφέας: ΔΑΜΚΑΛΗ ΜΑΡΙΑ – ΕΛΕΝΗ

Επιβλέπων: ΛΑΠΠΑΣ ΠΑΝΤΕΛΗΣ

Μέλος Επιτροπής: ΞΑΝΘΟΠΟΥΛΟΣ ΣΤΥΛΙΑΝΟΣ

Μέλος Επιτροπής: ΤΑΧΤΣΗΣ ΕΛΕΥΘΕΡΙΟΣ

ΣΑΜΟΣ

Στην οικογένειά μου
και στους αγαπημένους μου.

Περίληψη

Η παρούσα διπλωματική εργασία, η οποία διεξήχθη στο πλαίσιο του μεταπτυχιακού προγράμματος σπουδών «Στατιστική και Ανάλυση Δεδομένων» του Πανεπιστημίου Αιγαίου στο Τμήμα Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών, αποσκοπεί στη βελτιστοποίηση του τομέα της είσπραξης ληξιπρόθεσμων οφειλών (debt collections field). Ο συγκεκριμένος τομέας είναι ιδιαίτερα σημαντικός για τη βιωσιμότητα των οργανισμών, ενώ ταυτόχρονα ο τρόπος εφαρμογής στρατηγικών είσπραξης ληξιπρόθεσμων οφειλών απαιτεί ιδιαίτερη προσοχή καθώς έχει άμεση σχέση με τον άνθρωπο.

Για την επίτευξη αυτού του στόχου χρησιμοποιείται μία προσέγγιση πολλαπλών μεθόδων. Η έρευνα ξεκινάει εξετάζοντας το εννοιολογικό πλαίσιο του όρου «είσπραξη οφειλών», τον ρόλο, τις δυσκολίες αλλά και τη σημαντικότητα ύπαρξής του εν λόγω τομέα. Στη συνέχεια, αναλύονται βασικές θεωρητικές έννοιες της υπολογιστικής νοημοσύνης και της πολυκριτήριας ανάλυσης αποφάσεων. Τέλος, προτείνεται μία ολοκληρωμένη μεθοδολογία διαχείρισης ληξιπρόθεσμων οφειλών και διερευνάται η δυνατότητα εφαρμογής της σε πραγματικές συνθήκες. Πιο συγκεκριμένα, εξετάζεται μία μελέτη περίπτωσής όπου αξιοποιώντας πραγματικά δεδομένα επιχειρείται η στοχαστική κατηγοριοποίηση των πελατών (φυσικά πρόσωπα και επιχειρήσεις) και η εύρεση βέλτιστων στρατηγικών προσέγγισης και είσπραξης ληξιπρόθεσμων οφειλών. Η στοχαστική κατηγοριοποίηση των πελατών βασίζεται στον υπολογισμό πιθανοτήτων αναφορικά με (i) την λήψη θετικής υπόσχεσης πληρωμής (promise to pay) κατόπιν επιτυχούς τηλεφωνικής επικοινωνίας με τον πελάτη (φυσικό πρόσωπο ή επιχείρηση), (ii) τη μερική ή ολική αθέτηση των υποχρεώσεων πληρωμής (default payment) από τον πελάτη (φυσικό πρόσωπο ή επιχείρηση), αλλά και (iii) άλλες διαθέσιμες μεταβλητές ενδιαφέροντος στην βάση δεδομένων (π.χ., πόσες ημέρες ένας πελάτης είναι ενεργός στο σύστημα, υπόλοιπο οφειλής, κλπ.). Αξίζει να σημειωθεί ότι ο υπολογισμός των σχετικών πιθανοτήτων πραγματοποιείται με τη βοήθεια μοντέλων Λογιστικής Παλινδρόμησης (Logistic Regression Models), ενώ το σύνολο των μεταβλητών (i), (ii) και (iii) συνδυάζονται κατάλληλα με τη βοήθεια ενός Ασαφούς Συστήματος τύπου Mamdami (Mamdami Fuzzy Inference System) προκειμένου να πραγματοποιηθεί η τελική κατηγοριοποίηση των πελατών σε τρεις κλάσεις κινδύνου (κακοί, καλοί και πολύ καλοί πελάτες). Για κάθε κατηγορία πελατών, ένα σύστημα πολυκριτήριας ανάλυσης αποφάσεων βασισμένο στην Αναλυτική Ιεραρχική Διαδικασία (Analytic Hierarchy Process, AHP) εκτελείται συνδυάζοντας την κλάση κινδύνου που προκύπτει από το Ασαφές Σύστημα τύπου Mamdami με τις προτιμήσεις ενός ειδικού προκειμένου να σκοραριστούν και στη συνέχεια να ιεραρχηθούν/προταθούν οι πλέον κατάλληλες στρατηγικές είσπραξης οφειλών, ανά περίπτωση.

Τα αποτελέσματα της παρούσας εργασίας υπογράμμισαν τη σημασία της ενσωμάτωσης των προτεινόμενων μεθοδολογιών στις διαδικασίες είσπραξης οφειλών

που εφαρμόζει ένας οργανισμός προκειμένου προωθηθεί η επιστήμη στον τομέα της είσπραξης οφειλών και να καθιερωθεί ένα πλαίσιο για ακριβέστερες και αποτελεσματικότερες στρατηγικές επικοινωνίας με τους πελάτες.

Λέξεις Κλειδιά: Debt Collections, Λογιστική Παλινδρόμηση, Ασαφής Λογική, Πολυκριτήρια Ανάλυση Αποφάσεων

Abstract

This thesis, which was carried out in the context of the Master's degree programme "Statistics and Data Analysis" of the University of the Aegean at the Department of Statistics and Actuarial - Financial Mathematics, aims to optimize the field of debt collections. This area is particularly important for the sustainability of organizations, while at the same time the way of implementing strategies for the debt collection requires special attention as it is directly related to people.

A multi-method approach is used to achieve this goal. The research starts by examining the conceptual framework of the term 'debt collection', its role, the difficulties and the importance of its existence. Then, basic theoretical concepts of computational intelligence and multi-criteria decision analysis are analyzed. Finally, a comprehensive methodology for managing overdue debts is proposed and its applicability in real-life situations is explored. More specifically, a case study is considered where, using real data, a stochastic categorization of customers (individuals and companies) is attempted and optimal strategies for approaching and collecting overdue debts are found. The stochastic categorization of customers is based on the calculation of probabilities regarding (i) the receipt of a positive promise to pay after a successful telephone contact with the customer (individual or business), (ii) the partial or total default of payment obligations (default payment) by the customer (individual or business), and (iii) other available variables of interest in the database (e.g., how many days a customer is active in the system, outstanding debt, etc.). It is worth noting that the calculation of the relative probabilities is carried out with the help of Logistic Regression Models, while the set of variables (i), (ii) and (iii) are appropriately combined with the help of a Mamdami Fuzzy Inference System in order to carry out the final categorization of customers into three risk classes (bad, good and very good customers). For each customer class, a multi-criteria decision analysis system based on the Analytic Hierarchy Process (AHP) is performed by combining the risk class resulting from the Mamdami Fuzzy System with the preferences of an expert in order to score and then prioritize/propose the most appropriate debt collection strategies, case by case.

The results of this study highlighted the importance of incorporating the proposed methodologies into the debt collection processes implemented by an organization in order to advance the science in the field of debt collection and establish a framework for more accurate and effective customer communication strategies.

Key words: Debt Collections, Logistic Regression, Fuzzy Logic, Multicriteria Decision Analysis

Περιεχόμενα

Περίληψη 3
Abstract 5
Συντομογραφίες και Ακρωνύμια 8
Κεφάλαιο 1ο: Εισαγωγή 10
1.1 Αντικείμενο και Σκοπός της εργασίας 10
1.2 Δομή και Περιεχόμενο της εργασίας 11
Κεφάλαιο 2ο: Debt Collections 12
2.1 Εισαγωγή 12
2.2 Εισπράκτορας Οφειλών 14
2.3 Γιατί είναι σημαντικό; 16
2.4 Παλιότερες Σχετικές Μελέτες 17
Κεφάλαιο 3ο: Μέθοδοι Υπολογιστικής Νοημοσύνης 21
3.1 Υπολογιστική Νοημοσύνη 21
3.2 Θεωρία Μάθησης 24
3.2.1 Μηχανική Μάθηση 24
3.2.2 Επιβλεπόμενη Μάθηση 26
3.2.3 Παλινδρόμηση 27
3.2.4 Λογιστική Παλινδρόμηση 28
3.2.5 Μέτρα απόδοσης 30
3.3 Ασαφής Λογική 32
3.3.1 Ιστορική Ανασκόπηση 32
3.3.2 Ασαφής Λογική και Θεωρία Ασαφών Συνόλων 33
3.3.3 Πράξεις Ασαφών Συνόλων 35
3.3.4 Ασαφής Συλλογιστική 36
3.3.5 Αποασαφοποίηση 37
3.3.6 Σύστημα ασαφούς λογικής 37
3.4 Διαχείριση Δεδομένων 38
3.4.1 Καθαρισμός Δεδομένων 38
3.4.2 Υπερδειγματοληψία - Υποδειγματοληψία 39
3.4.3 Pearson Correlation 40
3.4.4 Μέθοδοι Επιλογής Μεταβλητών 40

Κεφάλαιο 4ο: Πολυκριτήρια Ανάλυση Αποφάσεων 43
4.1 Εισαγωγή 43
4.2 Ιστορική Αναδρομή 43
4.3 Βασικές έννοιες και μεθοδολογίες 44
4.4 Κύρια θεωρητικά ρεύματα 46
4.5 Πολυκριτήρια θεωρία χρησιμότητας 47
4.6 Αναλυτική Ιεραρχική Διαδικασία 49
Κεφάλαιο 5ο: Προτεινόμενη Μεθοδολογία 52
5.1 Προεργασία και Προετοιμασία Δεδομένων 52
5.2 Λογιστική Παλινδρόμηση για Πρόβλεψη Υπόσχεσης Πληρωμής (P2P) 53
5.3 Λογιστική Παλινδρόμηση για Πρόβλεψη Αθέτησης Πληρωμής (DPD) 55
5.4 Ασαφής Λογική 57
5.5 Πολυκριτήρια Ανάλυση Αποφάσεων 58
Κεφάλαιο 6ο: Μελέτη Περίπτωσης και Υπολογιστικά Αποτελέσματα 61
6.1 Περιβάλλον Ανάπτυξης και Σετ Δεδομένων 61
6.2 Εφαρμογή της Μεθόδου σε Επιχειρήσεις 62
6.2.1 Μοντέλο Λογιστικής Παλινδρόμησης για P2P 62
6.2.2 Μοντέλο Λογιστικής Παλινδρόμησης για DPD 65
6.2.3 Ασαφές Σύστημα Τύπου Mamdani 67
6.2.4 Πολυκριτήρια Ανάλυση Αποφάσεων 69
6.3 Εφαρμογή της Μεθόδου σε Φυσικά Πρόσωπα 70
6.3.1 Μοντέλο Λογιστικής Παλινδρόμησης για P2P 71
6.3.2 Μοντέλο Λογιστικής Παλινδρόμησης για DPD 73
6.3.3 Ασαφές Σύστημα Τύπου Mamdani 75
6.3.4 Πολυκριτήρια Ανάλυση Αποφάσεων 76
Κεφάλαιο 7ο: Συμπεράσματα και Μελλοντικές Επεκτάσεις 78
Παράρτημα 80
Βιβλιογραφία 91

Συντομογραφίες και Ακρωνύμια

Ελληνικά

ΑΛ	Ασαφής Λογική
ΓΑ	Γενετικός Αλγόριθμος
ΛΠ	Λογιστική Παλινδρόμηση
ΜΜ	Μηχανική Μάθηση
ΤΝ	Τεχνική Νοημοσύνη
ΤΝΔ	Τεχνητά Νευρωνικά Δίκτυα
ΥΝ	Υπολογιστική Νοημοσύνη

Αγγλικά

AI	Artificial Intelligence
AHP	Analytic Hierarchy Process
ANN	Artificial Neural Network
CI	Computational Intelligence
FL	Fuzzy Logic
GA	Genetic Algorithm
LR	Logistic Regression
ML	Machine Learning
MCDM	Multicriteria Decision Making

Συντομογραφίες Σχετικά με την Μελέτη Περίπτωσης

B	Bad Customer
BD1	Business Dataset 1
BD2	Business Dataset 2
FN	False Negative
FP	False Positive
G	Good Customer
RD1	Retail Dataset 1
RD2	Retail Dataset 2
P2P	Promise to Pay
DA	DF days active

DP	Days Past Due
DPD	Default Payment
Str-B	Strategy for Bad Customer
Str-G	Strategy for Good Customer
Str-VG	Strategy for Very Good Customer
TB	Total Balance
TN	True Negative
TP	True Positive
VG	Very Good Customer

Κεφάλαιο 1ο:
Εισαγωγή

1.1 Αντικείμενο και Σκοπός της εργασίας

Στις μέρες μας όλο και πιο συχνό είναι το φαινόμενο μία εταιρεία να προσφέρει παροχές ή υπηρεσίες αλλά να μένει σε αναμονή για την πληρωμή. Αυτού του είδους οι καταστάσεις μπορούν να επηρεάσουν σοβαρά τις ταμειακές ροές, την πιστοληπτική ικανότητα, ακόμη και τη φήμη της επιχείρησης. Σε τέτοιες περιπτώσεις αναλαμβάνει ο τομέας είσπραξης οφειλών να έλθει σε επαφή με τους πελάτες με στόχο την είσπραξη του μέγιστου δυνατού ποσού με το ελάχιστο κόστος. Το γεγονός της άμεσης επικοινωνίας με τον άνθρωπο καθιστά ακόμη πιο ευαίσθητο το θέμα και θα πρέπει οι εισπράκτορες, λαμβάνοντας υπόψη πολλαπλούς παράγοντες, να είναι πολύ προσεκτικοί όσον αφορά τη συμπεριφορά τους προς τους πελάτες και λόγω των νομοθεσιών αλλά και λόγω ηθικών βάσεων. Συμπεραίνουμε λοιπόν ότι, η είσπραξη οφειλών είναι πολύ σημαντική για την βιωσιμότητα των οργανισμών αλλά ταυτόχρονα απαιτεί ιδιαίτερη προσοχή ως προς τον τρόπο εφαρμογής της. Αν και υπάρχει πληθώρα προγνωστικών μοντέλων που έχουν σχεδιαστεί για τη διαχείριση κινδύνων, η είσπραξη οφειλών έχει λάβει ελάχιστη προσοχή στην ακαδημαϊκή κοινότητα. Επομένως, ένα έξυπνο σύστημα έγκαιρης είσπραξης θα είναι επωφελές τόσο για τους οργανισμούς όσο και για τους ακαδημαϊκούς.

Στην παρούσα μελέτη στόχος είναι η βελτιστοποίηση του τομέα είσπραξης οφειλών. Προτείνεται μία μεθοδολογία η οποία συμβάλλει σε αποτελεσματικότερες στρατηγικές επικοινωνίας με τον πελάτη καταλήγοντας στην μέγιστη είσπραξη οφειλών. Εμβαθύνει στις θεωρητικές βάσεις του τομέα και εξετάζει διάφορες τεχνικές κατηγοριοποίησης και αξιολόγησης πελατών, όπως τη Λογιστική Παλινδρόμηση, την Ασαφή Λογική και τη Πολυκριτήρια Ανάλυση Αποφάσεων. Επιπλέον, η προτεινόμενη μεθοδολογία εφαρμόζεται σε πραγματικά στοιχεία πελατών, ώστε να καταδείξει την αποτελεσματικότητα της. Η μέθοδος αυτή θα μπορούσε να χρησιμοποιηθεί ως εναλλακτική λύση στατιστικών μεθόδων για την εξεύρεση προσεγγιστικών λύσεων σε προβλήματα του πραγματικού κόσμου που περιλαμβάνουν ποικίλα σφάλματα και αβεβαιότητες.

Τα αποτελέσματα της παρούσας εργασίας τόνισαν τη σημαντικότητα των μεθοδολογιών της στατιστικής και των αλγορίθμων βελτιστοποίησης στην είσπραξη οφειλών, στην βελτιστοποίηση των διαδικασιών λήψης αποφάσεων και στη μείωση

πιθανών κινδύνων. Συνεπώς, η παρούσα μελέτη αποσκοπεί στην παροχή κινήτρων, καινοτομίας και βελτιστοποίησης του τομέα. Τα ευρήματα και οι ιδέες αυτής της μελέτης αποτελούν σημαντικό βήμα για την ενίσχυση των διαδικασιών λήψης αποφάσεων και της είσπραξης οφειλών.

1.2 Δομή και Περιεχόμενο της εργασίας

Η παρούσα διπλωματική αποτελείται από 7 κεφάλαια. Στο πρώτο που συνιστά και την εισαγωγή, αναλύονται το αντικείμενο και ο σκοπός της έρευνας, η δομή και το περιεχόμενο αυτής και γίνεται αναφορά στην περίπτωση μελέτης που εφαρμόστηκε.

Στο δεύτερο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο του τομέα είσπραξης ληξιπρόθεσμων οφειλών (debt collections field). Αρχικά, γίνεται αναφορά στην κατάσταση που επικρατεί σε παγκόσμιο επίπεδο με τις ληξιπρόθεσμες οφειλές δίνοντας μάλιστα τεκμηριωμένα ποσοστά επίσημων οργανισμών από Αμερική, Ευρώπη αλλά και Ελλάδα. Έπειτα, αναλύεται η έννοια του «εισπράκτορα οφειλών», των τρόπων επικοινωνίας που χρησιμοποιεί αλλά και τις δυσκολίες που μπορεί να αντιμετωπίσει τονίζοντας στη συνέχεια τη σημαντικότητα ύπαρξης τέτοιων φορέων για την επιβίωση μιας επιχείρησης. Το κεφάλαιο κλείνει με παλιότερες έρευνες και μελέτες σχετικές επί του θέματος.

Κατά το τρίτο κεφάλαιο, αναλύονται βασικές θεωρητικές έννοιες της Υπολογιστικής Νοημοσύνης, όπως η Θεωρία Μάθησης, η Παλινδρόμηση και η Ασαφής Λογική. Οι έννοιες αυτές χρειάζονται για την πορεία της εργασίας στη μελέτη περίπτωσης.

Σημαντικές έννοιες για την πορεία της εργασίας αναλύονται και στο τέταρτο κεφάλαιο, όπου αποκρυσταλλώνεται η Πολυκριτήρια Ανάλυση Αποφάσεων, η οποία αποβλέπει στην επιλογή της καταλληλότερης λύσης μέσα από ένα σύνολο εναλλακτικών επιλογών.

Στο πέμπτο κεφάλαιο, παρουσιάζεται η μεθοδολογία η οποία προτείνεται στην παρούσα διπλωματική και η οποία χωρίζεται σε 5 μέρη με τελικό στόχο την επιλογή της καλύτερης στρατηγικές επικοινωνίας για την είσπραξη ληξιπρόθεσμων λογαριασμών ανάλογα με τον εκάστοτε πελάτη.

Στο έκτο κεφάλαιο, γίνεται η εφαρμογή της προαναφερθείσας μεθοδολογίας με χρήση της γλώσσας προγραμματισμού Python. Τα δεδομένα που χρησιμοποιήθηκαν είναι στοιχεία πελατών του πραγματικού κόσμου και μελετήθηκαν σε δύο διαφορετικά πλαίσια δεδομένων, ένα για εταιρείες και ένα για φυσικά πρόσωπα.

Τέλος, στο έβδομο περιέχονται τα συμπεράσματα και καθώς και προτάσεις για μελλοντικές έρευνες.

Κεφάλαιο 2ο:
Debt Collections

2.1 Εισαγωγή

Στις μέρες μας, όλο και πιο σύνηθες είναι το φαινόμενο των απλήρωτων τιμολογίων και πιστώσεων. Πολλές από τις εταιρείες μπορεί να έχουν πουλήσει αγαθά ή υπηρεσίες και να τα έχουν τιμολογήσει στους πελάτες τους, αλλά να έχουν μείνει σε αναμονή για την πληρωμή, ή οι τράπεζες μπορεί να έχουν δώσει πιστώσεις αλλά δεν μπόρεσαν να πάρουν τις εκταμιεύσεις. Αυτού του είδους οι καταστάσεις μπορούν να επηρεάσουν σοβαρά τις ταμειακές ροές, τον κύκλο εργασιών, την πιστοληπτική ικανότητα, ακόμη και τη φήμη της επιχείρησης (Sezi Cevik Onar et al., 2015).

Η παγκόσμια οικονομία αντιμετωπίζει πρόβλημα καθυστερημένων πληρωμών. Το 2013, περίπου 30 εκατομμύρια άτομα, ή το 14% των Αμερικανών ενηλίκων, είχαν χρέη που βρίσκονταν σε διαδικασία είσπραξης ή υπόκειντο σε αυτή, με μέσο όρο περίπου 1.400 δολάρια (CFPB, 2014). Σύμφωνα με το FDCPA (2022), κατά τη διάρκεια του 2021, το χρέος των Αμερικανών καταναλωτών αυξήθηκε, από 14,33 τρισεκατομμύρια δολάρια το πρώτο τρίμηνο σε 15,58 τρισεκατομμύρια δολάρια το τελευταίο τρίμηνο του 2021. Το μη στεγαστικό χρέος αυξήθηκε επίσης κατά 4,33 τρισεκατομμύρια δολάρια το τελευταίο τρίμηνο του 2021, με σημαντικούς παράγοντες στην αύξηση του να αποτελούν η αύξηση χρεών από πιστωτικές κάρτες κατά 90 δισεκατομμύρια δολάρια και από δάνεια αυτοκινήτων κατά 80 δισεκατομμύρια δολάρια από το πρώτο έως το τελευταίο τρίμηνο του έτους. Το χρέος των φοιτητικών δανείων παρέμεινε περίπου αμετάβλητο και άλλα είδη χρέους εκτός κατοικίας είδαν σχετικά μέτρια αυξήσεις κατά τη διάρκεια του 2021[1].

Ανάλογα αποτελέσματα με τα παραπάνω παρατηρούνται και στην Ευρώπη. Σύμφωνα με την έκθεση του European Payment Report (EPR, 2023), το ποσοστό των επιχειρήσεων από τις οποίες έχει ζητηθεί να αποδεχθούν μεγαλύτερες προθεσμίες πληρωμής αυξάνεται χρόνο με το χρόνο. Το 2021, το 54% των ερωτηθέντων της EPR δήλωσε ότι είχε λάβει ένα αίτημα παράτασης από μια μεγάλη εταιρεία. Το 2022, το ποσοστό αυτό αυξήθηκε στο 61% και σήμερα βρίσκεται στο 66%. Ο αριθμός των ερωτηθέντων που λαμβάνουν το ίδιο αίτημα από μικρότερες επιχειρήσεις έχει αυξηθεί

[1] Αυτά τα στοιχεία του καταναλωτικού χρέους είναι σε ονομαστικά δολάρια και είναι μη προσαρμοσμένα για τον πληθωρισμό και την αύξηση του πληθυσμού. Δεν περιλαμβάνουν τα αναδυόμενα προϊόντα καταναλωτικών δανείων, όπως οι συμφωνίες εισοδηματικών μεριδίων (ISA) και το buy-now-paylater (BNPL) (FDCPA, 2022).

από 49% σε 55%. Αυτό είναι ένα πρόβλημα σε ολόκληρη την οικονομία, με τις επιχειρήσεις να αναφέρουν ότι οι πελάτες χρειάζονται σταθερά περισσότερο χρόνο για να ρυθμίσουν τις οφειλές τους. Ο τυπικός εταιρικός πελάτης των εταιριών χρειάζεται πλέον 56 ημέρες για να εξοφλήσει ένα τιμολόγιο, από 53 ημέρες το 2022. Στις επιχειρήσεις ενέργειας και κοινής ωφέλειας, οι πελάτες πληρώνουν τους λογαριασμούς, κατά μέσο όρο, μετά από 63 ημέρες. Οι κυβερνητικοί οργανισμοί και οι οργανισμοί δημόσιων υπηρεσιών πληρώνονται συνήθως μετά από 69 ημέρες. Τέτοιες καθυστερήσεις προκαλούν στις επιχειρήσεις σημαντικά προβλήματα στις ταμειακές ροές, αυξάνοντας ταυτόχρονα την ανησυχία για τον κίνδυνο αθέτησης πληρωμών. Περισσότερα από τα δύο τρίτα των επιχειρήσεων (70%), δηλώνουν ότι οι πιστωτικές απώλειες αποτελούν πρόβλημα για αυτές, από το 60% το 2022. Όλο και περισσότερο, οι επιχειρήσεις φοβούνται ότι μια καθυστερημένη πληρωμή θα μετατραπεί σε μη ανακτήσιμη απώλεια. Οι πτωχεύσεις στην Ευρώπη αποτελούν επίσης ένα αυξανόμενο πρόβλημα. Σύμφωνα με τη Eurostat, ο αριθμός των δηλώσεων πτώχευσης το τέταρτο τρίμηνο του 2022 έφθασε στο υψηλότερο επίπεδό του μεταξύ των επιχειρήσεων της ΕΕ από τότε που άρχισε η συλλογή δεδομένων το 2015.Οι επιχειρήσεις παραδέχονται όλο και περισσότερο ότι δυσκολεύονται να πείσουν τους πελάτες τους να πληρώσουν, με αποτέλεσμα να καθυστερούν τις πληρωμές των προμηθευτές τους, οδηγώντας δυνητικά σε διαταραχή της αλυσίδας εφοδιασμού και σε μια λιγότερο ανθεκτική ευρωπαϊκή οικονομία. Μόλις πριν από δύο χρόνια, το 29% των ερωτηθέντων παραδέχονταν ότι πλήρωναν τους προμηθευτές αργότερα από ό,τι θα δέχονταν από τους δικούς τους πελάτες. Το 2023, αυτό ποσοστό έχει αυξηθεί στο 37%.

Με τους πελάτες να πληρώνουν καθυστερημένα και να απαιτούν πιο επιεικείς όρους πληρωμής, σχεδόν οι μισές επιχειρήσεις στην Ευρώπη (44%) αναγνωρίζουν την ανάγκη να επενδύσουν στην διαχείριση καθυστερημένων πληρωμών με στόχο την αντιμετώπιση αυτού του προβλήματος (EPR, 2023). Το ποσοστό των επιχειρήσεων στην Ελλάδα που δήλωσαν ότι λαμβάνουν μέτρα για τη μείωση του πιστωτικού κινδύνου και τη βελτίωση των καθυστερήσεων πληρωμών ανέρχεται στο 64%, εκ των οποίων το 75% εστιάζει στις πρόωρες καθυστερήσεις (EPR Greece, 2023). Επιπλέον, κατά μέσο όρο, οι ευρωπαϊκές επιχειρήσεις δαπανούν 10,4 ώρες την εβδομάδα για να κυνηγήσουν καθυστερημένες πληρωμές, που ισοδυναμεί με 74 ημέρες ετησίως. Οι ελληνικές επιχειρήσεις ξοδεύουν ακόμη περισσότερο χρόνο - 78 ημέρες το χρόνο για κατά μέσο όρο (EPR Greece, 2023).

Η ανάγκη συνεπώς για εστίαση στον τομέα είσπραξης οφειλών γίνεται όλο και μεγαλύτερη για την βιωσιμότητα των επιχειρήσεων με αρκετές από αυτές να συνεργάζονται με φορείς είσπραξης οφειλών για την επένδυση σε νέα τεχνολογία και διαμόρφωση στρατηγικών συνεργασίας για την αντιμετώπιση του ζητήματος (EPR Greece, 2023). Ιδιαιτέρως, έπειτα από την έλευση του FDCPA το 1977, ο κλάδος είσπραξης οφειλών γνώρισε δραματική ανάπτυξη και μαζί σημειώθηκε και σημαντική εξέλιξη στις επιχειρηματικές πρακτικές (CFPB, 2014). Η εμφάνιση και η ανάπτυξη της αγοράς χρέους είναι μία από τις σημαντικότερες αλλαγές στην αγορά είσπραξης χρεών (CFPB, 2014). Η βιομηχανία είσπραξης οφειλών επηρεάζει εκατομμύρια πολίτες.

Σύμφωνα με τα τελευταία διαθέσιμα εκτιμήσεις, η αγορά είσπραξης οφειλών από τρίτους είναι μια βιομηχανία 18,6 δισεκατομμύρια δολαρίων που απασχολεί περίπου 138.000 άτομα σε περισσότερα από 7.000 γραφεία είσπραξης στις Ηνωμένες Πολιτείες (FDCPA, 2022).

2.2 Εισπράκτορας Οφειλών

Debt Collector

Ο εισπράκτορας οφειλών (debt collector) είναι ένα άτομο ή οργανισμός που ασχολείται με την ανάκτηση χρημάτων που οφείλονται σε ληξιπρόθεσμους λογαριασμούς. Πολλοί εισπράκτορες χρεών προσλαμβάνονται από εταιρείες στις οποίες οφείλονται χρήματα από ιδιώτες, και λειτουργούν έναντι πάγιας αμοιβής ή για ένα ποσοστό του ποσού που μπορούν να εισπράξουν. Ορισμένοι, μπορεί να είναι αγοραστές χρέους, αγοράζοντας το χρέος σε ένα κλάσμα της ονομαστικής τους αξίας και προσπαθώντας στη συνέχεια να ανακτήσουν το πλήρες ποσό του χρέους ή όσο περισσότερο μπορούν. Ένας εισπράκτορας οφειλών μπορεί επίσης να είναι γνωστός ως οργανισμός είσπραξης (collection agency) (IBISworld, 2023).

Οι οργανισμοί είσπραξης οφειλών διαθέτουν μεγάλο όγκο δεδομένων από τράπεζες, φορείς εκμετάλλευσης GSM και παρόχους ηλεκτρικού ρεύματος, φυσικού αερίου ή υπηρεσιών διαδικτύου. Συνεπώς, μεγάλος όγκος δεδομένων σχετικά με αυτές τις οφειλές συλλέγεται και αποθηκεύεται από τους οργανισμούς αυτούς (Sezi Cevik Onar et al., 2015). Επιπλέον, όσο αναφορά τον λογαριασμός του οφειλέτη επειδή τα στοιχεία είναι συνήθως ξεπερασμένα, οι περισσότερες επιχειρήσεις είσπραξης, πληρώνουν μια μικρή αμοιβή σε ένα ιδιωτικό πρακτορείο προκειμένου να αποκτήσουν έναν κατάλογο με πρόσθετες πιθανές διευθύνσεις και αριθμούς τηλεφώνων. Αφού λάβουν τη λίστα με τις πιθανές διευθύνσεις, οι επιχείρησεις πρέπει να προσδιορίσουν ποιες διευθύνσεις αποτελούν βιώσιμα στοιχεία (Martin Del Vecchio et al., 2006).

Με την χρήση όλων των παραπάνω δεδομένων οι οργανισμοί είσπραξης έχουν σαν στόχο τον εντοπισμό των καθυστερημένων οφειλών και τον καθορισμό των δυνατοτήτων των πελατών να αποπληρώσουν το χρέος ώστε να λάβουν το μέγιστο δυνατό ποσό στο ελάχιστο κόστος, πριν την έναρξη κάθε είδους νομικής διαδικασίας. Για την επίτευξη αυτού, οι εισπράκτορες χρησιμοποιούν πολύ καλά μοντέλα μηχανικής μάθησης (EPR, 2023) ώστε να χρησιμοποιούν αποδοτικά τους ανθρώπινους πόρους για την επικοινωνία τους με τους πελάτες (Sezi Cevik Onar et al., 2015). Συνεπώς, είναι απαραίτητη η σωστή αξιολόγηση και άμεση εξαγωγή συμπερασμάτων λαμβάνοντας υπόψη πολλαπλούς παράγοντες όπως το μέγεθος του εν λόγω χρέους και ο χρόνος που απομένει για την παραγραφή του χρέους (Martin Del Vecchio et al., 2006).

Στη συνέχεια, λαμβάνονται διάφορες ενέργειες με βάση τη δυνατότητα αποπληρωμής του χρέους. Τα μέσα επικοινωνίας που χρησιμοποιούνται για να έρθουν σε επαφή με τους οφειλέτες είναι κατά κύριο λόγο το διαδίκτυο, τα τηλέφωνα, τα ηλεκτρονικά μηνύματα

και τα φωνητικά μηνύματα (Sezi Cevik Onar et al., 2015). Κάθε μία από αυτές τις τεχνολογίες έχει διαφορετικές επίπεδα αυτοματοποίησης και κόστους. Για παράδειγμα, τα μηνύματα ηλεκτρονικού ταχυδρομείου μπορούν να σταλούν αυτόματα και δωρεάν, τα μηνύματα στο κινητό μπορούν να αποστέλλονται αυτόματα αλλά αυτή τη φορά για κάθε μήνυμα καταβάλλεται ένα τέλος στον πάροχο επικοινωνίας. Αξιοσημείωτο είναι ότι για να πραγματοποιηθεί μια τηλεφωνική κλήση, ο εισπράκτορας πρέπει να πληρώσει ένα τέλος στον πάροχο και επίσης χρειάζεται ένας χειριστής για την κλήση. Η επιλογή του καταλληλότερου τύπου επικοινωνίας επηρεάζει άμεσα το κόστος του συλλέκτη και τη σχέση του με τον πελάτη. Σημαντικοί παράγοντες που πρέπει να λαμβάνονται υπόψιν αποτελούν ο τύπος επικοινωνίας που προτιμά ο οφειλέτης, ο τύπος του μηνύματος που χρησιμοποιείται για τις γραπτές επικοινωνίες, ο τόνος ομιλίας κατά τις τηλεφωνικές κλήσεις, και η ώρα της ημέρας που καλείται ο οφειλέτης *(Sezi Cevik Onar et al.,2015)*.

Σημαντικό κομμάτι του τομέα αυτού αποτελούν οι σχέσεις των εταιρειών με τους πελάτες τους. Η είσπραξη των πληρωμών από τους πελάτες με το ελάχιστο δυνατό κόστος πριν από τη νομική διαδικασία είναι ζωτικής σημασίας για τους οργανισμούς, δεδομένου ότι η νομική διαδικασία δεν είναι μόνο δαπανηρή, αλλά βλάπτει και τις σχέσεις με τους μαζί τους. Ως αποτέλεσμα, οι εταιρείες προσπαθούν να εισπράξουν τις οφειλές πριν από την έναρξη της νομικής διαδικασίας η οποία αρχίζει μετά από 90 ημέρες καθυστέρησης. Τα συστήματα είσπραξης που προσπαθούν να εισπράξουν οφειλές εντός 0-90 ημερών ονομάζονται συστήματα έγκαιρης είσπραξης (Sezi Cevik Onar et al., 2015). Επιπλέον, οι εισπράκτορες πρέπει να είναι πολύ προσεκτικοί όσον αφορά τη συμπεριφορά τους προς τους πελάτες λόγω των νομοθεσιών σχετικά με την είσπραξη οφειλών, όπως ο νόμος περί δίκαιων πρακτικών είσπραξης οφειλών (FDSPA, 2014) στις ΗΠΑ, και τις πολιτικές που έχουν σχεδιαστεί από τα κράτη μέλη της Ευρωπαϊκής Ένωσης (Huls N., 1992). Σύμφωνα με το FDCPA (2014) η είσπραξη χρεών αποτελεί ένα από τα σημαντικότερα προβλήματα των καταναλωτών και κορυφαία πηγή καταγγελιών καθώς το 2013. Οι ομοσπονδιακές υπηρεσίες έλαβαν περισσότερες από 200.000 καταγγελίες καταναλωτών σχετικά με τη συμπεριφορά των εταιρειών είσπραξης. Επιπλέον, από τη σκοπιά των εισπρακτόρων, εάν ο οφειλέτης έχει πρόβλημα ενδέχεται να οφείλετε σε βραχυπρόθεσμο πρόβλημα ρευστότητας και πιέζοντας πολύ σκληρά για την είσπραξη οφειλών θα μπορούσε η εταιρεία να τον χάσει ως πελάτη ή να τον οδηγήσει σε αφερεγγυότητα, ενώ η κατανόηση των προβλημάτων του και η λήψη κατάλληλων μέτρων μπορεί να θέσει τον εισπράκτορα σε πολύ καλύτερη θέση. Είναι λογικό λοιπόν ότι, οι ακατάλληλες διαδικασίες είσπραξης οφειλών μπορεί να επηρεάσουν τη σχέση με τους πελάτες και να οδηγήσουν σε αύξηση των παραπόνων των πελατών και απώλεια εσόδων (Hsiu-Yu Wang et al., 2013).

Οι πράκτορες είσπραξης καθορίζουν τη δυνατότητα των πελατών να αποπληρώσουν το χρέος με διάφορα κριτήρια. Η διαδικασία αυτή περιέχει υψηλό βαθμό υποκειμενικότητας, ασάφειας και ανακρίβειας, δεδομένου ότι ο ορισμός του υψηλού δανείου μπορεί να είναι διαφορετικός για κάθε εισπράκτορα (Sezi Cevik Onar et al., 2015). Παρ' όλα αυτά, μια λανθασμένου στυλ επικοινωνίας μπορεί επίσης να βλάψει τη σχέση μεταξύ των εταιρειών και των πελατών. Ο Lund (2010) προσδιορίζει τον όρο Soft Debt Collection (SDC) ως μια διαδικασία κατανόησης του πελάτη, των λόγων για τους

οποίους δεν πληρώνει, και την επιλογή της κατάλληλης πορεία δράσης. Τα τελευταία χρόνια η διαχείριση των σχέσεων με τους πελάτες υιοθετείται όλο και περισσότερο ως βασική επιχειρηματική στρατηγική και επενδύεται σε μεγάλο βαθμό από τις εταιρείες. Οι υπηρεσίες προστιθέμενης αξίας καθώς και η κατανόηση των αναγκών των πελατών αναγνωρίζονται ως σημαντικοί συντελεστές οι οποίοι προσδιορίζουν την επιτυχία ή την αποτυχία μιας εταιρείας (Hsiu-Yu Wang et al., 2013). Ως εκ τούτου, είναι σημαντικό για τους εισπράκτορες σήμερα να διαθέτουν κατάλληλες στρατηγικές για την είσπραξη οφειλών από μεμονωμένους πελάτες, διατηρώντας παράλληλα θετική σχέση μαζί τους. Ταυτόχρονα, μια κατάλληλη τακτική υπενθύμισης στην διαδικασίας είσπραξης μπορεί επίσης να μειώσει τη δυσφορία και τα παράπονα των πελατών, και στη συνέχεια να βελτιώσει τις σχέσεις με τους πελάτες (Hsiu-Yu Wang et al., 2013).

2.3 Γιατί είναι σημαντικό;

Τα συστήματα είσπραξης οφειλών είναι πολύ σημαντικά για την επιβίωση των οργανισμών. Αναλαμβάνουν σε περίπτωση απλήρωτων τιμολογίων ή πιστωτικών εκταμιεύσεων να συγκεντρώσουν το ανεξόφλητο ποσό, ιδίως στο σημερινό παγκόσμιο κλίμα ύφεσης όπου τα απλήρωτα τιμολόγια και οι απλήρωτες πιστώσεις γίνονται όλο και πιο συχνές. Επιπλέον, αποτελούν μία πολύ σημαντική επιχειρηματική εφαρμογή της προβλεπτικής ανάλυσης. Το έργο αυτό συνίσταται στην πρόβλεψη των πιθανοτήτων αποπληρωμής των καθυστερημένων πληρωτών. Υπό αυτή την έννοια, τα κέντρα επαφής έχουν κεντρικό ρόλο στην είσπραξη οφειλών, καθώς βελτιώνουν την κερδοφορία μετατρέποντας τις χρηματικές απώλειες σε άμεσο όφελος για τις τράπεζες και άλλα χρηματοπιστωτικά ιδρύματα (Catalina Sanchez et al., 2022).

Σύμφωνα με το EPR (2023), οι εταιρείες είσπραξης, χρησιμοποιούν πολύ καλά μοντέλα μηχανικής μάθησης τα οποία βοηθούν σε κάθε στάδιο του κύκλου ζωής της πίστωσης, από το να προσφέρουν στους πελάτες τα καλύτερα προϊόντα για τις ανάγκες τους, έως την πρόβλεψη της πιθανής ζημίας στο πιστωτικό χαρτοφυλάκιο και την εύρεση της καλύτερης στρατηγικής κατά τη διάρκεια των διαδικασιών είσπραξης και ανάκτησης για συγκεκριμένες ομάδες πελατών. Σύμφωνα με το IBISworld 2014 στην αγορά των ΗΠΑ υπάρχουν περίπου 10.000 γραφεία που πραγματοποιούν έσοδα 13 δισεκατομμυρίων δολαρίων. Επιπλέον, τα έσοδα από την είσπραξη οφειλών αυξήθηκαν με CAGR 1,6% σε 20,2 δισ. δολάρια μέχρι το τέλος του 2023 (IBISworld, 2023). Όσο πιο αποτελεσματικές γίνονται οι επιχειρήσεις στην είσπραξη ληξιπρόθεσμων πληρωμών, τόσο πιο γρήγορα θα μπορούν να ξεκλειδώσουν οφέλη. Σε ένα δύσκολο οικονομικό περιβάλλον, αξίζει να είναι κανείς καλός στις εισπράξεις (EPR Greece, 2023).

Συμπεραίνουμε λοιπόν ότι, η είσπραξη οφειλών είναι πολύ σημαντική για την επιβίωση των οργανισμών και την προβλεπτική ανάλυση. Τα μοντέλα είσπραξης οφειλών όχι μόνο επιτρέπουν τη διενέργεια προβλέψεων των μελλοντικών ταμειακών ροών και απαιτήσεων, αλλά βελτιώνουν επίσης την επιτυχία των εργασιών είσπραξης. Για παράδειγμα, το εργατικό δυναμικό μπορεί να ανακατευθυνθεί προς εκείνους τους

πελάτες που είναι πιο πιθανό να μην αποπληρώσουν τις υποχρεώσεις πληρωμής, ή μπορούν να σχεδιαστούν νέες πολιτικές είσπραξης με τη διάκριση μεταξύ οφειλετών που είναι πιθανό να αποπληρώσουν ή όχι. Υπό αυτή την έννοια, οι οργανισμοί είσπραξης έχουν κεντρικό ρόλο στην είσπραξη οφειλών, καθώς βελτιώνουν την κερδοφορία μετατρέποντας τις χρηματικές απώλειες σε άμεσο όφελος για τις τράπεζες και άλλα χρηματοπιστωτικά ιδρύματα.

2.4 Παλιότερες Σχετικές Μελέτες

Παρά την πληθώρα προγνωστικών μοντέλων που έχουν σχεδιαστεί για τη διαχείριση κινδύνων και, ειδικότερα, για την πιστωτική βαθμολόγηση, η είσπραξη οφειλών έχει λάβει ελάχιστη προσοχή στην ακαδημαϊκή κοινότητα (Catalina Sanchez et al., 2022). Τα μοντέλα είσπραξης οφειλών όχι μόνο επιτρέπουν τη διενέργεια ακριβών προβλέψεων των μελλοντικών ταμειακών ροών και απαιτήσεων, αλλά βελτιώνουν επίσης την επιτυχία των εργασιών είσπραξης. Επομένως ένα έξυπνο σύστημα έγκαιρης είσπραξης θα είναι επωφελές τόσο για τους οργανισμούς όσο και για τους ακαδημαϊκούς. Στη βιβλιογραφία υπάρχει ένας αριθμός μελετών που χρησιμοποιούν αναλυτικές προσεγγίσεις για την ενδυνάμωση των συστημάτων είσπραξης οφειλών, οι οποίες παρουσιάζονται με φθίνουσα χρονολογική διάταξη παρακάτω.

Κατά την πιο πρόσφατη μελέτη, των P. Z. Lappas et al. (2023) εισάγεται ένας εξελικτικός αλγόριθμος ομαδοποίησης με σκοπό την τμηματοποίηση πελατών βάση των χαρακτηριστικών τους. Για την έρευνα, συλλέχθηκαν δύο σύνολα δεδομένων του πραγματικού κόσμου από χρηματοπιστωτικά ιδρύματα στην Ελλάδα και διερευνήθηκαν με σκοπό την τμηματοποίηση πελάτων σε ομάδες και την απεικόνιση των σημαντικών χαρακτηριστικών κάθε ομάδας. Η έρευνα διαχωρίζεται σε τρία στάδια, κατά το πρώτο στάδιο χρησιμοποιείται γενετικός αλγόριθμος (ΓΑ) που συνδέεται με την επιλογή χαρακτηριστικών, στο δεύτερο στάδιο χρησιμοποιείται ένας αλγόριθμος μηχανικής μάθησης (ML) χωρίς επίβλεψη (Kmeans) που σχετίζεται με το πρόβλημα της ομαδοποίησης και κατά το τρίτο στάδιο εφαρμόζεται ένας αγωγός εκπαίδευσης στον οποίο το αποτέλεσμα της συσταδοποίησης χρησιμοποιείται για την εκπαίδευση αλγορίθμων ML με επίβλεψη όπως η βαθιά μάθηση και τα τυχαία δάση.

Στην μελέτη των Catalina Sanchez et al. (2022) διερευνάται η πιθανότητα επιτυχούς επικοινωνίας με έναν οφειλέτη, η πιθανότητα επίτευξης μιας επαφής που οδηγεί σε μια υπόσχεση πληρωμής και η πιθανότητα ο πελάτης να εξοφλήσει τις ληξιπρόθεσμες οφειλές του. Επιπλέον, προτείνεται ένα πλαίσιο ενοποίησης δεδομένων που έχει σχεδιαστεί για την διαδικασία είσπραξης οφειλών ενός χρηματοπιστωτικού ιδρύματος. Η προσέγγισή βασίζεται σε πρόσφατες επεξηγήσιμες μεθόδους, όπως η TreeSHAP, με σκοπό την εξαγωγή γνώσης από τις μεθόδους πρόβλεψης και παροχή μιας νέας εφαρμογής στους ακαδημαϊκούς στους τομείς της συγχώνευσης και ολοκλήρωσης δεδομένων.

Οι Bellotti et al. (2021) υιοθέτησαν μια προσέγγιση προγνωστικής ανάλυσης, προβλέποντας τα ποσοστά ανάκτησης δανείων. Εφάρμοσαν αυτή τη στρατηγική στον κλάδο της τραπεζικής, συγκρίνοντας 20 διαφορετικές μεθόδους στατιστικής και μηχανικής μάθησης. Κατέληξαν στο συμπέρασμα ότι σύνολα βασισμένα σε κανόνες, όπως το gradient boosting, το random forest και οι cubist πέτυχαν την καλύτερη προβλεπτική απόδοση. Παρόλο που οι συγγραφείς δεν προβλέπουν την καθυστέρηση πληρωμής, περιλαμβάνουν μηνιαίες πληροφορίες σχετικά με τις κλήσεις, τις επισκέψεις και τις επαφές μεταξύ του χρηματοπιστωτικού ιδρύματος και των δανειοληπτών.

Οι Bahrami et al. (2020) εξέτασαν την βαθμολόγηση συμπεριφοράς για την πρόβλεψη πληρωμής τιμολογίων λαμβάνοντας υπόψη τον χρόνο λήξης τους. Συνέκριναν τη λογιστική παλινδρόμηση και τα SVM, καταλήγοντας ότι το πρώτο είναι σε θέση να επιτύχει εξαιρετικά αποτελέσματα σε ένα μεγάλο σύνολο δεδομένων 1,6 εκατομμυρίων πελατών. Μεταξύ των μεταβλητών που περιλαμβάνονται στο μοντέλο, λαμβάνουν υπόψη τους ιστορικά αρχεία με τις ενέργειες επικοινωνίας των εταιρειών προς τους πελάτες για να είσπραξη οφειλών, συμπεριλαμβανομένων των ηλεκτρονικών μηνυμάτων (email), την υπηρεσία μηνυμάτων (sms) και τις τηλεφωνικές κλήσεις.

Στην ίδια κατεύθυνση, οι Chehrazi et al. (2019) πρότειναν μια στοχαστική προσέγγιση προγραμματισμού για τη δυναμική βελτιστοποίηση των πιστωτικών εισπράξεων. Παρόλο που η προσέγγιση αυτή δεν εξετάζει τη μοντελοποίηση βάσει δεδομένων, δείχνει σε παραδείγματα προσομοίωσης, ότι το στοχαστικό μοντέλο που προτείνεται είναι σε θέση να μειώσει τις απώλειες, καθορίζοντας τον καλύτερο δυνατό χρόνο για τις πολιτικές είσπραξη.

Οι Van De Geer et al. (2018) πρότειναν μια μεθοδολογία που συνδυάζει μηχανική μάθηση και δυναμικό προγραμματισμό, ακολουθώντας το σκεπτικό της εργασίας των Abe et al. (2010). Η μεθοδολογία τους βελτιστοποιεί τις τηλεφωνικές κλήσεις με τέτοιο τρόπο ώστε να δίνεται προτεραιότητα στους πελάτες που είναι πιθανότερο να πληρώσουν τα χρέη τους. Θεωρούν την ενίσχυση κλίσης (gradient boosting) ως μια προγνωστική προσέγγιση, επιτυγχάνοντας εξαιρετικά αποτελέσματα όσον αφορά την είσπραξη οφειλών σε μια πραγματική εφαρμογή. Η μελέτη αυτή εξετάζει βασικές πληροφορίες σχετικά με προηγούμενες επαφές με τους πελάτες και σχετίζεται με το ότι οι πληροφορίες των εταιρειών είσπραξης αποτελούν πολύτιμη πηγή δεδομένων για την πρόβλεψη καθυστερήσεων πληρωμών.

Η εργασία των Kim και Kang (2016) πρότεινε ένα σύστημα για την ανάθεση οφειλετών σε εισπράκτορες με αποτελεσματικό και δίκαιο τρόπο. Ανέπτυξαν ένα σύστημα βασισμένο στη μηχανική μάθηση με προσέγγιση πρόβλεψης πληρωμής, συγκρίνοντας δέντρα αποφάσεων, ANNs, Support Vector Machines (SVM) και τυχαίο δάσος. Τα μοντέλα πρόβλεψης ενεργούν ως είσοδος για την εξαγωγή κανόνων βαθμολόγησης για δίκαιη κατανομή των πελατών σε εισπράκτορες τηλεφωνικού κέντρου. Έδειξαν ότι οι προγνωστικές αναλύσεις παράγουν καλύτερα κανόνες βαθμολόγησης σε σχέση με τις υπάρχουσες μεθόδους.

Στην μελέτη των Sezi Cevik Onar et al. (2015), χρησιμοποιούνται ευφυή συστήματα βασισμένα στον ασαφές συμπερασμό τύπου Mamdani με σκοπό την ενίσχυση των διαδικασιών έγκαιρης είσπραξης οφειλών. Βασίζονται στις αρχές της επιστήμης των δεδομένων με χρήση διαδικασιών μάθησης και γενικευμένης εξαγωγής συμπερασμάτων με σκοπό την απόκτηση χρήσιμων προβλέψεων και γνώσεων. Πιο ειδικά, αναπτύσσεται ένα σύστημα έγκαιρης είσπραξης οφειλών που αποτελείται από τρία διαφορετικά συστήματα ασαφούς εξαγωγής συμπερασμάτων (FIS), ένα για πιστωτικά χρέη, ένα για χρέη από πιστωτικές κάρτες και ένα για τιμολόγια. Τα συστήματα αυτά χρησιμοποιούν διαφορετικές εισροές, όπως το ποσό του δανείου, τον πλούτο του οφειλέτη, ιστορικό μέρους του οφειλέτη, ποσό άλλων χρεών, ενεργός πελάτης από τότε, πιστωτικό όριο και κρισιμότητα για να καθορίσουν την πιθανότητα εξόδου για την αποπληρωμή του χρέους. Αυτή η έξοδος χρησιμοποιείται αργότερα για τον προσδιορισμό του καταλληλότερου προφίλ επικοινωνίας με τον οφειλέτη.

Οι Takahashi και Tsuda (2013) επικεντρώνονται στα χαρακτηριστικά των κακών πελατών στον κλάδο των ταχυδρομικών παραγγελιών. Οι συγγραφείς δημιουργούν ένα σύστημα για τον εντοπισμό δυνητικών οφειλετών χρησιμοποιώντας μια προσέγγιση τυχαίου δάσους. Τα αποτελέσματα της μελέτης δείχνουν ότι η τοποθεσία και η χρηματική αξία της συναλλαγής είναι δύο σημαντικοί παράγοντες για τον εντοπισμό δυνητικών οφειλετών.

Οι Wang et al. (2013) εστιάζουν στην αγορά τηλεπικοινωνιών στην Ταϊβάν και προτείνουν ένα μοντέλο για την αποτροπή της αύξησης των επισφαλών απαιτήσεων και της απομάκρυνσης πελατών από τις εταιρείες. Οι συγγραφείς δημιουργούν ένα μοντέλο βαθμολόγησης πελατών με βάση τη συμπεριφορά χρησιμοποιώντας μια απόφαση δέντρου. Τα αποτελέσματα της πρακτικής εφαρμογής δείχνουν ότι το συνολικό κόστος είσπραξης μειώνεται κατά 0,4% των ετήσιων εσόδων και σημαντική μείωση της απομάκρυνσης πελατών που προκαλείται από χαοτικές στρατηγικές είσπραξης.

Οι Chen και Huang (2011) παρουσιάζουν μια προσέγγιση εξόρυξης δεδομένων για την πρόβλεψη της συμπεριφοράς των πελατών πιστωτικών καρτών. Οι συγγραφείς χρησιμοποιούν τεχνητά νευρωνικά δίκτυα και δέντρα αποφάσεων για την πρόβλεψη τακτικών μοτίβων κατανάλωσης, πληρωμής/αθέτησης πληρωμών και επισφαλών χρεών.

Οι Abe et al. (2010) χρησιμοποιούν μοντελοποίηση δεδομένων και τεχνικές βελτιστοποίησης για τη διαχείριση των διαδικασιών στο πλαίσιο της περιορισμένης διαδικασίας απόφασης Markov (MDP) και είσπραξης χρεών σε χρηματοπιστωτικά ιδρύματα. Βασικός στόχος είναι η εύρεση των οφειλετών που πρέπει να προσεγγιστούν, οι πιθανές ενέργειες είσπραξης που πρέπει να εφαρμοστούν σε αυτούς, ποιος πρέπει να λάβει αυτές τις ενέργειες και πότε πρέπει να ληφθούν. Το σύστημα τέθηκε σε λειτουργία στο φορολογικό και οικονομικό τμήμα της Νέας Υόρκης τον Δεκέμβριο του 2009 και υπήρξε σημαντική κάλυψη και προσοχή από τον Τύπο (π.χ. CNN Money, NY Channel 1).

Ο Fei (2010) ανέπτυξε ένα μοντέλο βαθμολόγησης της συμπεριφοράς αποπληρωμής των κατόχων πιστωτικών καρτών, σε δύο στάδια. Στο πρώτο στάδιο, γίνεται η αρχική

ταξινόμηση των πελατών με χρήση του αυτόματο ανιχνευτή αλληλεπίδρασης Chi-square (CHAID) και τεχνητά νευρωνικά δίκτυα. Με βάση τα αποτελέσματα αυτού, καθορίζονται οι σημαντικοί παράγοντες της ταξινόμησης. Οι παράγοντες αυτοί χρησιμοποιούνται για τη δημιουργία περιβάλλουσας ανάλυσης δεδομένων (DEA) στο δεύτερο στάδιο. Στόχος της μελέτης, είναι η αξιολόγηση σε ατομικό επίπεδο, προκειμένου οι τράπεζες να μειώσουν το κόστος της πιθανής λανθασμένης ταξινόμησης των πελατών.

Οι Georgopoulos και Giannaropoulos (2007) επικεντρώνονται στη βελτιστοποίηση των πόρων των κέντρων επαφής που αλληλεπιδρούν με τους οφειλέτες για την είσπραξη οφειλών. Οι συγγραφείς χρησιμοποιούν τεχνητά νευρωνικά δίκτυα (ANN) προκειμένου να αναπτύξουν ένα σύστημα βαθμολόγησης για την είσπραξη οφειλών. Χρησιμοποιώντας αυτές τις βαθμολογίες, τα κέντρα είσπραξης μπορούν να αξιοποιήσουν καλύτερα τους πράκτορές τους και να βελτιστοποιήσουν το συνολικό κόστος είσπραξης.

Σε μία άλλη μελέτη, οι Vecchio et. al (2006) διερευνούν τον εντοπισμό ενός κακοπληρωτή ο οποίος επιχειρεί να διαφύγει από ένα χρέος. Σε τέτοιες περιπτώσεις, οι φορείς είσπραξης χρέους αναζητούν σε μια βάση δεδομένων πιθανές διευθύνσεις του οφειλέτη. Στόχος της μελέτης είναι να προσδιοριστεί η ακριβής διεύθυνση όταν υπάρχουν πολλαπλές διευθύνσεις που επιστρέφονται από τη βάση δεδομένων. Οι συγγραφείς αναπτύσσουν ένα MS Excel Macro που βασίζεται στην απόσταση Levenshtein και την ιεραρχική ομαδοποίηση και αναφέρεται η ακρίβεια του αλγορίθμου 70,46%.

Τέλος, το πρόβλημα βελτίωσης της είσπραξης οφειλών μέσω της προγνωστικής ανάλυσης και της επιχειρησιακής έρευνας εξετάστηκε για πρώτη φορά από τους Mitchner και Peterson (1957), οι οποίοι βελτιστοποίησαν τον χρόνο στον οποίο ένας πράκτορας πρέπει να επικοινωνήσει με έναν καθυστερημένο πληρωτή για να εισπράξει το χρέος του με βάση το κόστος που δημιουργεί. Ανέπτυξαν στατιστικές τεχνικές λήψης αποφάσεων για τη διαλογή δανείων και παρείχαν έναν οδηγό ως προς το κατάλληλο χρονικό διάστημα που θα πρέπει να επιδιώκονται κινήσεις επικοινωνίας από τους εισπράκτορες για τα μη πληρωτέα δάνεια προτού χαρακτηριστούν ως μη εισπράξιμα. Η εφαρμογή αυτών των κανόνων με προσομοίωση σε τυχαίο δείγμα δανείων υποδεικνύει δυνητική αύξηση του καθαρού κέρδους περίπου κατά 33%.

Κεφάλαιο 3ο:
Μέθοδοι Υπολογιστικής Νοημοσύνης
Machine Learning, Fuzzy Logic

3.1 Υπολογιστική Νοημοσύνη

Computational Intelligence

Η Τεχνητή Νοημοσύνη, προσπαθεί να μιμηθεί τον άνθρωπο συγκεντρώνοντας μια τεράστια γνώση που αποκτάται με τη χρήση της συλλογιστικής, του σχεδιασμού, της αναζήτησης και της πρόβλεψης. Δυστυχώς αποτυγχάνει σε ορισμένους τομείς που απαιτούν την κατασκευή ενός μεγάλου συνόλου κανόνων και αντιμετωπίζει επίσης προκλήσεις λόγω των αυξανόμενων απαιτήσεων στη μάθηση και τη βελτιστοποίηση της αναζήτησης. Αυτές οι αποτυχίες της τεχνητής νοημοσύνης άνοιξαν το δρόμο για την ανάπτυξη των υπολογιστικών εργαλείων που οδήγησαν στην άνοδο του νέου σχήματος που είναι η Υπολογιστική Νοημοσύνη (Computational Intelligence) (Jennifer S. Raj, 2019).

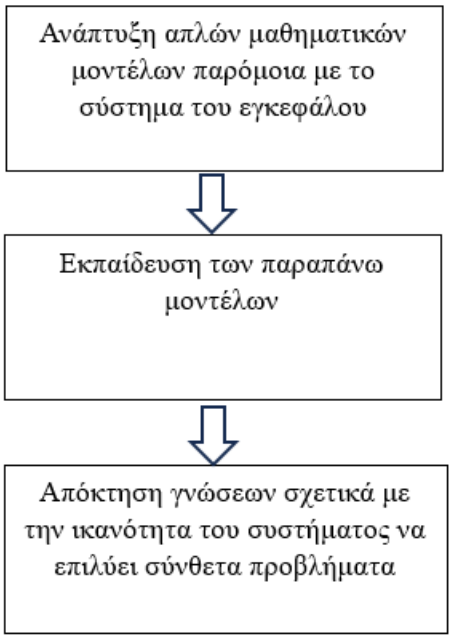
Αρκετοί επιστήμονες έχουν ορίσει με διαφορετικούς τρόπους τον όρο «Υπολογιστική Νοημοσύνη», ανάλογα με τις τότε εξελίξεις αυτού του νέου επιστημονικού κλάδου. Μια τομή αυτών των ορισμών, επικεντρώνεται στην Ασαφή Λογική (Fuzzy Logic), στα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Network-ANN) και το Γενετικό Αλγόριθμο (Genetic Algorithm-GA), όμως γενίκευση όλων αυτών των ορισμών περιλαμβάνει πολλά άλλα θέματα όπως η θεωρία των ακατέργαστων συνόλων (rough set theory), η θεωρία του χάους (chaos theory) και η θεωρία της υπολογιστικής μάθησης (computational learning theory). Περαιτέρω, η Υπολογιστική Νοημοσύνη ως αναδυόμενος κλάδος δεν θα πρέπει να περιορίζεται μόνο σε έναν περιορισμένο αριθμό θεμάτων, αλλά θα πρέπει μάλλον να έχει το περιθώριο να επεκτείνεται προς διάφορες κατευθύνσεις και να συγχωνεύεται με άλλους υπάρχοντες κλάδους (Konar, 2006).

Η Υπολογιστική Νοημοσύνη (ΥΝ) σύμφωνα με τον Konar (2006), έχει οριστεί ως ο συνδυασμός των ευφυών εργαλείων και των υπολογιστικών μοντέλων που είναι ικανά να δέχονται ακατέργαστα δεδομένα και να τα επεξεργάζονται άμεσα, χρησιμοποιώντας τον παραλληλισμό αναπαράστασης και τη διοχέτευση του προβλήματος, δημιουργώντας αξιόπιστες και έγκαιρες απαντήσεις με υψηλή ανοχή σε σφάλματα.

Παρακάτω παρουσιάζονται συνοπτικά οι βασικές τεχνικές της υπολογιστικής νοημοσύνης σύμφωνα με την J. S. Raj (2019):

- **Νευρωνικά Δίκτυα (Neural Networks)**

Τα Νευρωνικά Δίκτυα βασίζονται σε έννοιες υπολογισμού που είναι παρόμοιες με τον εγκέφαλο, εξασφαλίζοντας έναν υπολογισμό με παράλληλο τρόπο παρέχοντας μια γρήγορη επεξεργασία σε αντίθεση με τους παραδοσιακούς υπολογιστές που λειτουργούν με σειριακό τρόπο και χρειάζονται πολύ χρόνο για τους υπολογισμούς. Η βασική ιδέα των Νευρωνικών Δικτύων είναι ο σχεδιασμός ενός απλού μαθηματικού πλαισίου που μοιάζει με το σύστημα του εγκεφάλου και στην εκμάθησή του ώστε η συσκευή να έχει την ικανότητα επίλυσης διαφόρων πολύπλοκων προβλημάτων. Το Σχήμα 3.1 δείχνει την λογική των Νευρωνικών Δικτύων και τα βήματα που εμπλέκονται σε αυτά.



Σχήμα 3.1

- **Εξελικτικός Αλγόριθμος (Evolutionary Algorithm)**

Η θεμελιώδης έννοια του εξελικτικού αλγορίθμου βασίζεται στις διαδικασίες της φυσικής επιλογής. Προτιμάται κατά κύριο λόγο στους τομείς όπου οι συμβατικές μαθηματικές μέθοδοι είναι ασύμβατες για ένα ευρύτερο φάσμα προβλημάτων και συνήθως απασχολούνται σε εφαρμογές όπως η ανάλυση DNA και τα προβλήματα χρονοδιαγράμματος. Ένας από τους εξέχοντες αλγορίθμους εξέλιξης που βασίζεται στη διαδικασία της φυσικής επιλογής είναι ο Γενετικός Αλγόριθμος που ακολουθεί τα βήματα της αρχικοποίησης του πληθυσμού και της αξιολόγησης του βέλτιστου με τη βοήθεια της μετάλλαξης και της διασταύρωσης. Γενικά, οι Εξελικτικοί Αλγόριθμοι στοχεύουν στην

ανάδειξη νέων εξελικτικών τεχνικών που εκμεταλλεύονται τη δύναμη της φυσικής εξέλιξης και πιθανότατα ασχολούνται με τα προβλήματα βελτιστοποίησης.

- **Πιθανολογικές Μέθοδοι (Probabilistic Methods)**

Αποτελούν σημαντικό στοιχείο της ασαφούς λογικής. Αναφέρεται ως μη εποικοδομητική μέθοδος και ασχολείται κυρίως με την καταμέτρηση. Απαριθμεί τις εξόδους των ευφυών συστημάτων που βασίζονται στην υπολογιστική νοημοσύνη με τυχαίο τρόπο, προκειμένου να προσδιοριστούν τα πιθανά αποτελέσματα για το πρόβλημα που βασίζονται στη γνώση που αποκτήθηκε προηγουμένως.

- **Θεωρία Μάθησης (Learning Theory)**

Είναι μια από τις σημαντικές προσεγγίσεις της Υπολογιστικής Νοημοσύνης. Η Θεωρία της Μάθησης επιτρέπει την κατανόηση του αποτελέσματος και των εμπειριών ενός γεγονότος και την απόκτηση βαθιών γνώσεων από αυτά και την αξιοποίησή τους για την πρόβλεψη των μελλοντικών αποτελεσμάτων. Η πιο δημοφιλής μάθηση στις μέρες μας είναι η Μηχανική Μάθηση (Machine Learning), η οποία μπορεί να κατηγοριοποιηθεί ως η επιβλεπόμενη, η μη επιβλεπόμενη και η ενισχυτική μάθηση. Οι μηχανές συλλέγουν τις πληροφορίες από τα προηγούμενα γεγονότα που έχουν συμβεί, τις εκπαιδεύουν και τις χρησιμοποιούν για την πρόβλεψη της ακριβούς εξόδου του προβλήματος. Κάθε φορά που αναλύουν το πρόβλημα, το αντιστοιχίζουν με τα προηγούμενα συμβάντα που έχουν μάθει για να παράγουν ακριβή αποτελέσματα. Αυτοί οι τύποι Υπολογιστικής Νοημοσύνης απασχολούνται συνήθως στην εφαρμογή που απαιτεί διαγνώσεις, ανίχνευση και πρόβλεψη.

- **Ασαφής Λογική (Fuzzy Logic)**

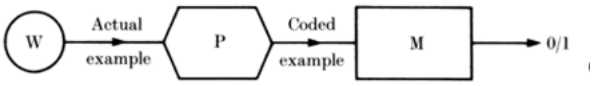
Το ασαφές σύνολο θεωρείται το υπερσύνολο της λογικής Boole. Στη λογική Boole όλα τα δεδομένα μπορούν να αναπαρασταθούν μόνο με τη χρήση δύο τιμών, είτε του "μηδέν" είτε του "ένα". Αλλά αυτό δεν ισχύει στην περίπτωση της ασαφούς λογικής η οποία ορίζει τις πολλαπλές καταστάσεις που βρίσκονται μεταξύ της υψηλής και της χαμηλής στάθμης. Για ένα δεδομένο εύρος εισόδων που κυμαίνεται από 0 έως 1, το ασαφές σύνολο μπορεί να οριστεί ως το σύνολο των ενδιάμεσων αριθμών που βρίσκονται μεταξύ του μηδενός και του ενός και όλες οι τιμές που περιγράφουν το βαθμό αλήθειας μπορούν να συμβολιστούν ως η συνάρτηση συμμετοχής M, όπου το M εκφράζεται μαθηματικά ως $0 < M < 1$.

Στην συνέχεια θα γίνει μία πιο λεπτομερής ανάλυσης των δυο τελευταίων τεχνικών που αναφέρθηκαν παραπάνω, λόγω του ότι χρησιμοποιήθηκαν στην μελέτη και την εφαρμογή της παρούσας εργασίας.

3.2 Θεωρία Μάθησης

Learning Theory

Θεωρία Μάθησης ονομάζεται ένα συνεπές εννοιολογικό πλαίσιο το οποίο έχει ως στόχο να περιγράψει και να εξηγήσει τον τρόπο λειτουργείας της ανθρώπινης μάθησης, δηλαδή να παρουσιάσει μια λεπτομερή περιγραφή του πώς μαθαίνει ο άνθρωπος. Συνεπώς, μια θεωρία μάθησης, είναι κατά κύριο λόγο ένα θεωρητικό μοντέλο, το οποίο πρωτίστως, περιγράφει τις βασικές πτυχές της μάθησης, διερευνά πειραματικά τη σχέση μεταξύ των βασικών παραμέτρων της και διατυπώνει ερευνητικά θεμελιωμένα συμπεράσματα που αποκαλύπτουν το φαινόμενο της μάθησης. Το Σχήμα 3.2 αναπαριστά το γενικό πλαίσιο της έννοιας της «μάθησης» για την παρούσα εργασία. Ο κύκλος W αναπαριστά ένα σύνολο αντικειμένων του πραγματικού κόσμου τα οποία θα ονομάζουμε παρατηρήσεις. Το εξάγωνο P είναι ένας επεξεργαστής ο οποίος λαμβάνει μια παρατήρηση και τη μετατρέπει σε ένα κωδικοποιημένο μήνυμα, όπως έναν χαρακτήρα σε bits. Η κωδικοποιημένη έκδοση των παρατηρήσεων αποτυπώνεται στο M η οποία είναι μια μηχανή που έχει στόχο να αναγνωρίσει ορισμένες μελλοντικές παρατηρήσεις (Anthony M. et al., 1997).



Σχήμα 3.2 (Anthony M. et al., 1997)

Παρά το γεγονός ότι, επί χρόνια επιστήμονες της Γνωστικής Ψυχολογίας και φιλόσοφοι μελετούν την έννοια της μάθησης, ακόμη δεν έχει γίνει πλήρως κατανοητή. Συνεπώς, το έργο των επιστημόνων του χώρου της Τεχνητής Νοημοσύνης να δημιουργήσουν υπολογιστικά συστήματα ικανά να μάθουν και να επιτύχουν είναι ιδιαιτέρως πολύπλοκο και δύσκολο (Γεωργούλη Α., 2015).

3.2.1 Μηχανική Μάθηση

Machine Learning

Η πιο γνωστή Θεωρία Μάθησης είναι η Μηχανική Μάθηση, η οποία αποτελεί ένα τεράστιο διεπιστημονικό πεδίο που βασίζεται σε έννοιες της Πληροφορικής, της Στατιστικής, της Γνωστικής Επιστήμης, της Μηχανικής και πολλούς άλλους κλάδους (Soofi & Awan, 2017). Υπάρχουν πολυάριθμες εφαρμογές για τη μηχανική μάθηση, αλλά η εξόρυξη δεδομένων είναι σημαντικότερη από όλες. Υπάρχουν πολλοί διαφορετικοί τρόποι κατηγοριοποίησης των αλγορίθμων της Μηχανικής Μάθησης, όμως συνηθίζεται να διακρίνονται σε κατηγορίες με βάση τον τρόπο με τον οποίο λαμβάνεται η μάθηση ή

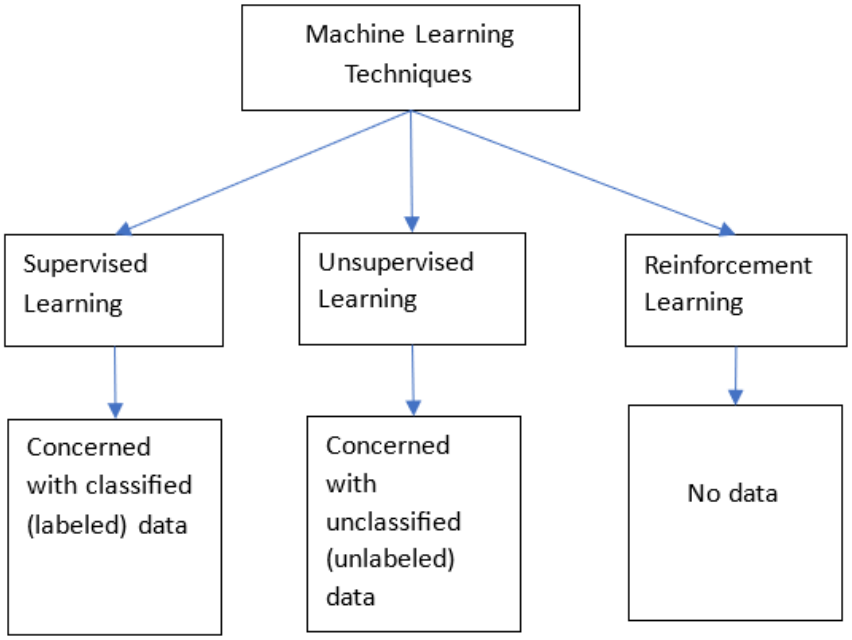
τον τρόπο με τον οποίο γίνεται ανάδραση στην εκμάθηση στο ανεπτυγμένο σύστημα (Κουδέρη Χ., 2020).

Παρακάτω αναφέρονται οι κύριοι τύποι αλγορίθμων μηχανικής μάθησης (Σχήμα 3.3):

- **Επιβλεπόμενη Μάθηση (Supervised learning)**
  Η διαδικασία μάθησης μιας συνάρτησης που αντιστοιχίζει μια είσοδο σε μια έξοδο, με βάση παραδειγμάτων από ζεύγη εισόδου – εξόδου από τα δεδομένα εκπαίδευσης, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για εισόδους με άγνωστη έξοδο.

- **Μη-επιβλεπόμενη Μάθηση (Unsupervised learning)**
  Κατά την εκπαίδευση, χρησιμοποιούνται πληροφορίες που δεν είναι ούτε ταξινομημένες, ούτε επισημασμένες και έτσι ο αλγόριθμος λειτουργεί χωρίς καθοδήγηση. Σκοπός είναι να ομαδοποιούνται ασαφείς πληροφορίες χωρίς προηγούμενη εκπαίδευση σε δεδομένα.

- **Ενισχυτική Μάθηση (Reinforcement learning)**
  Ο αλγόριθμος μαθαίνει μια στρατηγική για το πώς θα ενεργήσει δεδομένης μιας παρατήρηση του κόσμου. Κάθε ενέργεια έχει κάποια επίδραση στο περιβάλλον, και το περιβάλλον παρέχει ανατροφοδότηση που καθοδηγεί τον αλγόριθμο μάθησης.

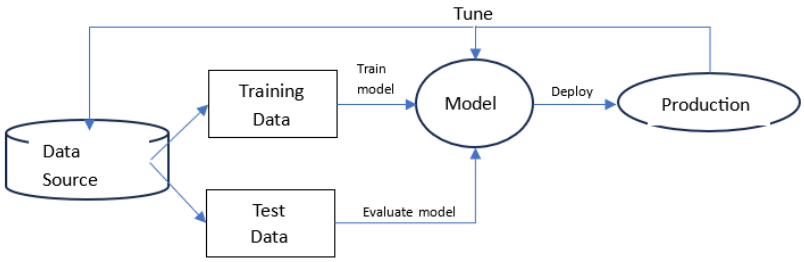


Σχήμα 3.3

3.2.2 Επιβλεπόμενη Μάθηση

Supervised Learning

Η Επιβλεπόμενη Μάθηση είναι το έργο μάθησης μιας συνάρτησης που αντιστοιχίζει μια είσοδο σε μια έξοδο έχοντας ως βάση παραδείγματα ζευγαριών εισόδου-εξόδου. Οι επιβλεπόμενοι αλγόριθμοι μηχανικής μάθησης χρειάζονται βοήθεια από εξωτερικά δεδομένα. Το σύνολο δεδομένων εισόδου διακρίνεται σε σύνολο δεδομένων εκπαίδευσης (training) και δοκιμής (test). Όλοι οι αλγόριθμοι μαθαίνουν κάποιο είδος προτύπων από το σύνολο δεδομένων εκπαίδευσης και τα εφαρμόζουν στο σύνολο δεδομένων δοκιμής για πρόβλεψη ή ταξινόμηση με στόχο την πρόβλεψη της σωστής ετικέτας για νέες εισόδους. Η ροή εργασίας των αλγορίθμων μηχανικής μάθησης με επίβλεψη δίνεται στο Σχήμα 3.4 (Batta Mahesh, 2018).



Σχήμα 3.4

Σύμφωνα με την Γεωργούλη Α. (2015) για τη συνάρτηση πρόγνωσης ισχύουν τα ακόλουθα:

- Η συνάρτηση δέχεται δεδομένα εισόδου, τα οποία χαρακτηρίζονται ως στιγμιότυπα (instance), δημιουργώντας έτσι ένα σύνολο στιγμιότυπων.
- Οι είσοδοι διαθέτουν γνωρίσματα (attributes) τα οποία χαρακτηρίζονται ως σημαντικά από την αρχή της μελέτης του προβλήματος.
- Ορισμένες είσοδοι συγκεντρώνονται από παρατηρήσεις και σχηματίζουν το σύνολο εκπαίδευσης (training set) το οποίο είναι υποσύνολο του συνόλου στιγμιότυπων.
- Το υπόλοιπο μέρος του συνόλου, αποτελεί το σύνολο δοκιμής (test set) το οποίο θα χρησιμοποιηθεί στο στάδιο της πιστοποίησης.
- Συνάρτηση στόχου (goal function), καλείται η συνάρτηση που απεικονίζει μια είσοδο από το σύνολο εκπαίδευσης στη γνωστή της έξοδο.
- Μεταβλητή στόχου (goal variable), δίνεται σε μια μεταβλητή και είναι η τιμή που επιστρέφει η συνάρτηση στόχου για ένα στιγμιότυπο από το σύνολο στιγμιότυπων.

- Η συμπεριφορά της συνάρτησης στόχου βελτιστοποιείται μέσω των διαδικασιών εκπαίδευσης με χρήση της συνάρτησης λάθους (error function) η οποία εντοπίζει τη διαφορά της μεταβλητής στόχου από την επιθυμητή έξοδο.

Οι τεχνικές με επίβλεψη μπορούν ταξινομούνται περαιτέρω σε δύο κύριες κατηγορίες: Ταξινόμηση και Παλινδρόμηση. Όταν η μεταβλητή-στόχος y που προσπαθούμε να προβλέψουμε είναι συνεχής, ονομάζουμε το πρόβλημα μάθησης, πρόβλημα παλινδρόμησης. Όταν το y μπορεί να πάρει μόνο ένα μικρό αριθμό διακριτών τιμών, το ονομάζουμε πρόβλημα ταξινόμησης (Andrew Ng & Tengyu Ma, 2007).

3.2.3 Παλινδρόμηση

Regression

Η τεχνική της παλινδρόμησης καθιερώθηκε από τις αρχές του 20ου, όταν ο Galton, σε αντίθεση με το νόμο της κανονικής κατανομής χρησιμοποιεί τον «νόμο των παρεκκλίσεων από τη μέση τιμή». Είναι αυτός που μετακίνησε την προσοχή από τη μέση τιμή και οι εργασίες του οδηγούν σε μια διαπίστωση, η οποία καταρρίπτει το σχήμα του Quetelet, υποστηρίζοντας ότι: μια κανονική κατανομή αντί να είναι ένδειξη ομοιογένειας, μπορεί να προκύπτει ως συνιστώσα πολλών διαφορετικών κατανομών που αντιστοιχούν σε επιμέρους πληθυσμούς με μέσες τιμές πολύ διαφορετικές μεταξύ τους. Έτσι, η μέση τιμή μιας κανονικής κατανομής γίνεται ένα μετακινούμενο κέντρο βάρους που συνδέεται με τη σύνθεση των εσωτερικών δυνάμεων των κοινωνιών. Η διαπίστωση αυτή οδήγησε στην ανακάλυψη της παλινδρόμησης (Αθανασιάδης Ηλίας, 2005).

Η Παλινδρόμηση (Regression) είναι μία ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίησης, η οποία ερευνά την συσχέτιση μεταξύ μίας εξαρτώμενης μεταβλητής y και μιας ή περισσότερων ανεξάρτητων μεταβλητών x (Κόκκινος Γιάννης, 2011). Συνεπώς, η παλινδρόμηση περιλαμβάνει την εκμάθηση μιας συνάρτησης $y = f(x)$ μέσω της οποίας ένα στοιχειώδες δεδομένο x απεικονίζεται σε μία πραγματική μεταβλητή πρόβλεψης y. Πριν την εφαρμογή της παλινδρόμησης, πρέπει ότι τα σχετικά δεδομένα να ταιριάζουν με μερικά γνωστά είδη συναρτήσεων (γραμμική, μη-γραμμική, πολυωνιμική κλπ.) και έπειτα μέσω αυτής προσδιορίζεται η καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα. Η συνάρτηση παλινδρόμησης προβλέπει την συνάρτηση συμμμετοχής του διανύσματος x στην κλάση με τιμή y. Ο πιο απλός αλλά πολύ διαδεδομένος τρόπος παλινδρόμησης είναι η γραμμική παλινδρόμηση:

$$y = c_0 + c_1 x_1 + \cdots + c_n x_n$$

η οποία υποθέτει γραμμικές συσχετίσεις και μπορεί να παράγει μία διαχωριστική συνάρτηση που διαχωρίζει έναν υποχώρο σε δύο περιοχές κλάσεων.

Υπάρχουν διάφοροι τύποι παλινδρόμησης ανάλογα με το είδος της συνάρτησης που ταιριάζουν καλύτερα τα σχετικά δεδομένα. Ορισμένοι από αυτούς είναι: Γραμμική παλινδρόμηση (Linear Regression), Λογιστική παλινδρόμηση (Logistic Regression), Πολυωνιμική παλινδρόμηση (Polynomial Regression), Παλινδρόμηση διανυσμάτων υποστήριξης (Support Vector Regression), Παλινδρόμηση δέντρων απόφασης (Decision Tree Regression), Παλινδρόμηση τυχαίων δασών (Random Forest Regression).

3.2.4 Λογιστική Παλινδρόμηση

Logistic Regression

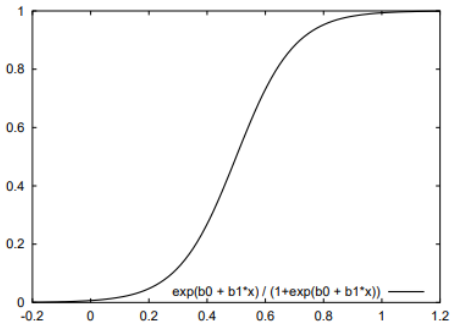
Υπάρχουν πολλές εφαρμογές για τις οποίες η διαδεδομένη γραμμική παλινδρόμηση δεν είναι κατάλληλη ή βέλτιστη (Paul Komarek, 2004). Επειδή το εύρος του γραμμικού μοντέλου στην εξίσωση είναι όλο το $\mathbb{R}$, η χρήση γραμμικής παλινδρόμησης για δεδομένα με συνεχή αποτελέσματα στο (0,1) ή δυαδικά αποτελέσματα στο {0,1} μπορεί να μην είναι κατάλληλη. Η λογιστική παλινδρόμηση (ΛΠ) είναι μια εναλλακτική τεχνική παλινδρόμησης που είναι κατάλληλη για τέτοια δεδομένα.

Η λογιστική παλινδρόμηση πήρε το όνομά της από τη λογιστική καμπύλη που χρησιμοποιεί για να μοντελοποιήσει την αναμενόμενη τιμή. Η προσαρμογή των χαρακτηριστικών στη λογιστική καμπύλης εξαρτάται από τη σιγμοειδή συνάρτηση και είναι επίσης ευαίσθητη στις συσχετίσεις με τα ανεξάρτητες μεταβλητές η μεταβλητή-στόχος της είναι μη γραμμική (M. Elnaggar, 2020). Επιπλέον, διαθέτει μια στατιστική βάση η οποία, στις κατάλληλες συνθήκες, θα μπορούσε να χρησιμοποιηθεί για την επέκταση των αποτελεσμάτων ταξινόμησης σε μια βαθύτερη ανάλυση (Paul Komarek, 2004).

Η λογιστική παλινδρόμηση είναι μία από τις σημαντικότερες στατιστικές τεχνικές που χρησιμοποιούνται από στατιστικολόγους και ερευνητές για την ανάλυση και ταξινόμηση δυαδικών και αναλογικών συνόλων δεδομένων. Ορισμένα από τα κύρια πλεονεκτήματα της ΛΠ, σύμφωνα με τον Maher Maalouf (2011), είναι ότι μπορεί να παρέχει πιθανότητες και να επεκταθεί σε προβλήματα ταξινόμησης πολλαπλών κατηγοριών. Ένα άλλο πλεονέκτημα είναι ότι οι μέθοδοι που χρησιμοποιούνται στην ανάλυση μοντέλων ΛΠ ακολουθούν τις ίδιες αρχές που χρησιμοποιούνται στη γραμμική παλινδρόμηση.

Το Μοντέλο της Λογιστικής Παλινδρόμησης

Κατά την λογιστική παλινδρόμηση, τα δεδομένα της μελέτης προσαρμόζονται στην εξίσωση της λογιστικής καμπύλης (Σχήμα 3.5). Όπως παρατηρείται, η καμπύλη είναι σιγμοειδής μορφής και χαρακτηρίζεται από ένα στάδιο εκθετικής ανάπτυξης.



Σχήμα 3.5 (Paul Komarek, Andrew W. Moore ,2003)

Η λογιστική καμπύλη μπορεί να οριστεί ως εξής

$$f(z_i) = \frac{e^{z_i}}{1 + e^{z_i}}$$

Όπου $z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_M x_{iM}$ η μεταβλητή εισόδου που εκφράζει το μέτρο της ολικής συνεισφοράς όλων των συμμετεχουσών ανεξάρτητων μεταβλητών στο μοντέλο και το $f(z_i)$ είναι το αποτέλεσμα της. Η μεταβλητή εισόδου z_i λαμβάνει πραγματικές τιμές ενώ το αποτέλεσμα αυτής $f(z_i)$ περιορίζεται σε εύρος τιμών μεταξύ 0 και 1.

Ο λόγος των συμπληρωματικών πιθανοτήτων (odds ratio) εκφράζει την πιθανότητα εμφάνισης ενός αποτελέσματος και είναι ο λόγος της πιθανότητας να συμβεί το αποτέλεσμα προς την πιθανότητα να μην συμβεί το αποτέλεσμα. Έτσι, αν p είναι η πιθανότητα εμφάνισης ενός γεγονότος και $1 - p$ είναι η πιθανότητα να μην συμβεί τότε

$$odds = \frac{p}{1 - p}$$

Εάν η παραπάνω σχέση ενσωματωθεί στο μοντέλο της παλινδρόμησης σε λογαριθμική μορφή τότε καταλήγουμε στο μοντέλο

$$\ln\left(\frac{f(z_i)}{1 - f(z_i)}\right) = \ln(e^{z_i}) = z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_M x_{iM}$$

Για την εκτίμηση των παραμέτρων του μοντέλου, στην περίπτωση της λογιστικής παλινδρόμησης χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας (maximum likelihood method). Είναι σημαντικό να σημειωθεί ότι, κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο του λόγου πιθανοφάνειας, σε αντίθεση με τη γραμμική παλινδρόμηση στην οποία η εκτίμηση των παραμέτρων γίνεται με μέθοδο ελάχιστων τετραγώνων. Ως συνέπεια, στη πρώτη, λόγω του μεταβαλλόμενου ποσοστού διακύμανσης αναπτύσσεται ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή, ενώ η δεύτερη δέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων.

Στην περίπτωση της λογιστικής παλινδρόμησης, η πιθανοφάνεια είναι:

$$L = \prod_{i=1}^{n} f(z_i)^{Y_i} \, (1 - f(z_i)^{Y_i})$$

$$\acute{\eta}$$

$$\ln(L) = \sum_{i=1}^{n} Y_i \ln\big(f(z_i)\big) + \left(n - \sum_{i=1}^{n} Y_i\right)\big(1 - f(z_i)\big)$$

Είναι σημαντικό να σημειωθεί η διαφορά μεταξύ της λογιστικής παλινδρόμησης και των μεθόδων ελαχίστων τετραγώνων με σιγμοειδές ή λογιστικό μοντέλο. Η λογιστική παλινδρόμηση είναι ένας τύπος γενικευμένου γραμμικού μοντέλου με μια συνάρτηση σφάλματος που επιλέγεται για να αντικατοπτρίζει τη μεταβαλλόμενη διακύμανση των διωνυμικά κατανεμημένων υπολειμματικών σφαλμάτων. Οι εφαρμογές που χρησιμοποιούν ένα κριτήριο ελαχίστων τετραγώνων ή παρόμοιο κριτήριο για τέτοια μοντέλα αγνοούν τη μη ομοιόμορφη διακύμανση των σφαλμάτων στο πεδίο. Αυτό μπορεί να έχει ή να μην έχει σημασία για την απόδοση ενός ad-hoc εκτιμητή συναρτήσεων όπως τα νευρωνικά δίκτυα. ωστόσο, ένα τέτοιο προσέγγιση αποφεύγει τα στατιστικά θεμέλια που δίνουν αξιοπιστία στη λογιστική παλινδρόμηση.

3.2.5 Μέτρα απόδοσης

Στη μηχανική μάθηση, για την μέτρηση της απόδοσης ενός μοντέλου ταξινόμησης συχνά χρησιμοποιούνται τα μέτρα απόδοσης: ακρίβεια, ανάκληση, f1-score, accuracy και AUC. Είναι αξιοσημείωτο ότι πολλοί επιστήμονες δεδομένων εξέφρασαν διαφωνίες σχετικά με τη σημασία κάθε μετρικής. Ένα μέρος από αυτούς υποστηρίζει ότι η AUC είναι η καλύτερη μετρική για να υποδείξει το βέλτιστο μοντέλο, ενώ άλλοι επιστήμονες πιστεύουν ότι άλλες μετρικές είναι καλύτερες.

Η θεμελιώδης έννοια των παραπάνω μέτρων απόδοσης βασίζεται στον συγκεντρωτικό πίνακα συχνοτήτων των προβλέψεων και των πραγματικών τιμών (Confusion matrix – Σχήμα 3.6).

		Predicted	
		No	**Yes**
Actually	**No**	*TN*	*FP*
	Yes	*FN*	*TP*

Σχήμα 3.6

Πιο ειδικά συμβολίζουμε ως:

PN (Predicted No): Πρόβλεψη μη ύπαρξης του χαρακτηριστικού που μελετάται

PY (Predicted Yes): Πρόβλεψη ύπαρξης του χαρακτηριστικού

AN (Actually No): Πραγματικά μη ύπαρξης του χαρακτηριστικού

AY (Actually Yes): Πραγματικά ύπαρξης του χαρακτηριστικού

Επιπλέον,

TN (True Negative): το χαρακτηριστικό ήταν αρνητικό και προβλεπόταν αρνητικό

FN (False Negative): το χαρακτηριστικό ήταν θετικό αλλά προβλεπόταν αρνητικό

FP (False Positive): το χαρακτηριστικό ήταν αρνητικό αλλά προβλεπόταν θετικό

TP (True Positive): το χαρακτηριστικό ήταν θετικό και προβλεπόταν θετικό.

Μέσω του παραπάνω πίνακα παίρνουμε τις εξής πληροφορίες:

Ακρίβεια (precision):

$Ακρίβεια\ θετικών\ = TP/(TP + FP)$, δηλαδή το ποσοστό των θετικών προβλέψεων που είναι πραγματικά θετικές,

$Ακρίβεια\ αρνητικών = TN/(TN + FN)$, δηλαδή το ποσοστό των αρνητικών προβλέψεων που είναι πραγματικά αρνητικές.

Ανάκληση (recall):

$Ανάκληση\ θετικών = TP/(TP + FN)$, είναι το ποσοστό των πραγματικά θετικών που αναγνωρίστηκαν σωστά,

$Ανάκληση\ αρνητικών = TN/(TN + FP)$, είναι το ποσοστό των πραγματικά αρνητικών που αναγνωρίστηκαν σωστά.

F1-Score: εκφράζει το ποσοστό των προβλέψεων που ήταν σωστές. Είναι ένας σταθμισμένος αρμονικός μέσος ακρίβειας και ανάκλησης έτσι ώστε η καλύτερη βαθμολογία είναι 1 και η χειρότερη είναι 0.

$$Βαθμολογία\ F1\ =\ 2 * (Ανάκληση\ *\ Ακρίβεια)/(Ανάκληση + Ακρίβεια)$$

Accuracy: αντιπροσωπεύει το ποσοστό των σωστών προβλέψεων.

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

Τέλος, η **Area Under Curve (AUC)**, χρησιμοποιείται ως σύνοψη της καμπύλης ROC (Receiver Operating Characteristic) και είναι το μέτρο της ικανότητας ενός δυαδικού ταξινομητή να διακρίνει μεταξύ κλάσεων. Όσο υψηλότερη τιμή έχει η AUC, τόσο καλύτερη είναι η απόδοση του μοντέλου για την αναγνώριση μεταξύ των θετικών και των αρνητικών κλάσεων.

3.3 Ασαφής Λογική

Fuzzy Logic

Τα μοντέλα πιθανοτήτων είναι διαδεδομένα στην ποσοτικοποίηση και την αξιολόγηση κινδύνων. Έχουν γίνει η θεμελιώδης βάση για τη λήψη τεκμηριωμένων αποφάσεων σχετικά με τον κίνδυνο σε πολλούς τομείς. Ωστόσο, ένα μοντέλο πιθανοτήτων που βασίζεται στην κλασική θεωρία συνόλων μπορεί να μην είναι σε θέση να περιγράψει ορισμένους κινδύνους με ουσιαστικό και πρακτικό τρόπο (Kailan Shang et al., 2013). Γεννήθηκε λοιπόν η ανάγκη για κατασκευή και εφαρμογή καταλληλότερων μοντέλων λειτουργικού κινδύνου, ένα εκ αυτών είναι η Ασαφής Λογική (Fuzzy Logic).

3.3.1 Ιστορική Ανασκόπηση

Από ιστορική άποψη, το ζήτημα της αβεβαιότητας δεν ήταν πάντα αποδεκτό από την επιστημονική κοινότητα (Timothy J. Ross, 2010). Από τα τέλη του 19ου αιώνα μέχρι τα τέλη του 20ου αιώνα, η θεωρία των πιθανοτήτων αποτελούσε την κορυφαία θεωρία στην ποσοτικοποίηση της αβεβαιότητας στα επιστημονικά μοντέλα. Ωστόσο, η έκφραση της αβεβαιότητας με τη χρήση της θεωρίας πιθανοτήτων αμφισβητήθηκε, πρωτίστως το

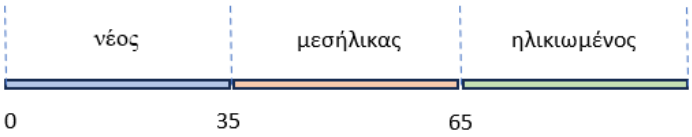
1937 από τον Max Black, με τις μελέτες του για την ασάφεια, και στη συνέχεια με την εισαγωγή των ασαφών συνόλων από τον Zadeh (1965). Η εργασία του Zadeh είχε βαθιά επίδραση στην σκέψη σχετικά με την αβεβαιότητα, διότι αμφισβήτησε τη θεωρία πιθανοτήτων που έως τότε αποτελούσε την μοναδική αναπαράσταση της αβεβαιότητας, αλλά και τα ίδια τα θεμέλια πάνω στα οποία βασίστηκε αυτή (Timothy J. Ross, 2010).

Οι πρώτες αντιδράσεις της επιστημονικής κοινότητας απέναντι στην ασαφή θεωρία ήταν αρνητικές, καθώς υπήρχε η άποψη ότι η κλασσική θεωρία πιθανοτήτων ήταν σε θέση να επιλύσει οποιοδήποτε πρόβλημα. Ωστόσο, η επιβίωση της ασαφούς θεωρίας οφείλεται κατά κύριο λόγο στην αφοσίωση του Zadeh, κατά τη διάρκεια της αμφισβήτησής της. Το 1968, πρότεινε την έννοια του ασαφούς αλγορίθμου (fuzzy algorithm), και δύο χρόνια αργότερα (1970), σε συνεργασία με τον Bellman, την έννοια των ασαφούς λήψης αποφάσεων (fuzzy decision making). Αργότερα προτάθηκαν έννοιες όπως την ασαφή διάταξη (fuzzy ordering), τη λεκτική μεταβλητή (linguistic variable) και τους ασαφείς κανόνες (fuzzy if–then rules) (Mastorokostas P., 2015). Η βαθμιαία αποδοχή της από την επιστημονική κοινότητα άρχισε μετά από με την εμφάνιση των πρώτων εφαρμογών της. Πιο συγκεκριμένα, το 1975, οι Mamdani και Assilian παρουσίασαν έναν ασαφή ελεγκτή για έλεγχο ατμομηχανής, το 1976, ο Tong πρότεινε έναν ασαφή ελεγκτή στη διαδικασία παραγωγής χάλυβα και το 1978 οι Holmblad και Østergaard δημιούργησαν έναν ασαφή ελεγκτή τσιμέντου. Τα επόμενα έτη, οι εφαρμογές της ξεπέρασαν τα όρια της Ευρώπης και προς την Άπω Ανατολή, με αποτέλεσμα το πλήθος και η σημαντικότητα των εφαρμογών με χρήση της ασαφούς λογικής να εξελιχθούν εκθετικά.

3.3.2 Ασαφής Λογική και Θεωρία Ασαφών Συνόλων

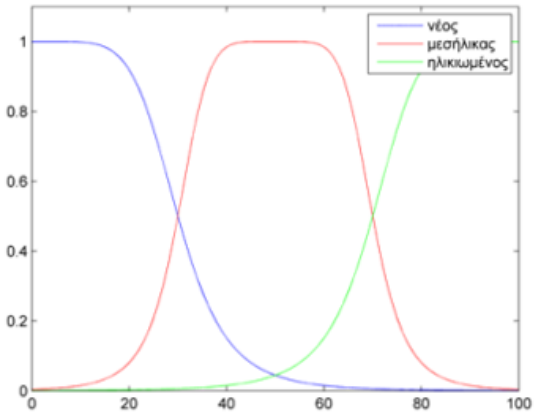
Fuzzy Logic and Fuzzy Set Theory

Στην κλασική θεωρία συνόλων, ένα μεμονωμένο αντικείμενο είτε αποτελεί μέλος, είτε δεν αποτελεί μέλος ενός συνόλου. Ωστόσο, στην πραγματικότητα, λόγω ανεπαρκούς γνώσης ή ασαφών δεδομένων, δεν είναι πάντα σίγουρο αν ένα αντικείμενο ανήκει σε ένα σύνολο ή όχι. Από την άλλη, τα Ασαφή Σύνολα (Fuzzy sets) ερμηνεύουν την αβεβαιότητα με έναν προσεγγιστικό τρόπο. Εννοιολογικά, η θεωρία ασαφών συνόλων επιτρέπει σε ένα αντικείμενο να ανήκει σε πολλαπλά σύνολα στο πλαίσιο συλλογισμού. Για κάθε ασαφές σύνολο, υπάρχει ένας βαθμός αλήθειας που ένα αντικείμενο ανήκει σε αυτό. Για παράδειγμα, ας υποθέσουμε ότι στις ηλικίες υπάρχουν τρία επίπεδα βαθμολογίας: νέος, μεσήλικας και ηλικιωμένος, τα οποία μπορούν να θεωρηθούν ως τρία σύνολα. Με βάση την κλασική θεωρία συνόλων, το πλήρες σύνολο αποτελείται από αυτά τα τρία αποκλειστικά σύνολα. Μόλις γίνει γνωστή η ηλικία ενός ατόμου, προσδιορίζεται το επίπεδο της ηλικίας του. Το Σχήμα 3.7 δείχνει ένα παράδειγμα κλασικών συνόλων. Αν ένας άνθρωπος είναι 36 ετών τότε είναι 100 τοις εκατό αληθές ότι χαρακτηρίζεται ως μεσήλικας.



Σχήμα 3.7

Το Σχήμα 3.8 δείχνει ένα παράδειγμα ασαφών συνόλων για τις ηλικίες. Κάθε σύνολο έχει την δική του συνάρτηση συμμετοχής (membership function), η οποία καθορίζει το βαθμό αλήθειας που ένα στοιχείο ανήκει στο σύνολο. Για παράδειγμα, για το ασαφές σύνολο «ηλικιωμένος» οι τιμές από τα 80 έτη και πάνω, δίνουν βαθμό συμμετοχής πάνω από το 0.95, γεγονός που δηλώνει ότι πως από αυτήν την ηλικία και πέρα, περιγράφεται ολοφάνερα το νόημα του συνόλου. Όσο πλησιάζουμε προς τις ηλικίες των 70 ετών, ο βαθμός συμμετοχής του συνόλου «ηλικιωμένος» μειώνονται σημαντικά, ενώ οι βαθμοί συμμετοχής του συνόλου «μεσήλικας» αυξάνουν σταδιακά. Συνεπώς, μπορεί αυτή η ηλικία να καλύπτεται από τα δύο σύνολα, ωστόσο άτομα με τέτοιες ηλικίες δεν μπορεί να καθορισθούν απόλυτα ότι είναι μεσήλικες ή ηλικιωμένοι καθώς διαθέτουν και τις δύο ιδιότητες κατά ένα ποσοστό. Πιο ειδικά, ένα άτομο 70 ετών συμπεριλαμβάνεται εξίσου στις κατηγορίες μεσήλικας και ηλικιωμένος, αφού όπως παρατηρείται, οι αντίστοιχες συναρτήσεις συμμετοχής έχουν τον ίδιο βαθμό 0.5. Η κατάσταση αυτή χαρακτηρίζεται ως κατάσταση μέγιστης ασάφειας, καθώς το άτομο δεν μπορεί να καταταχθεί με σαφήνεια ούτε στην μία αλλά ούτε στην άλλη κατηγορία.



$$A = \text{«νέος»} = \{(x,\mu_A(x)) | x \in U\},\ \mu_A(x) = \frac{1}{1+\left(\frac{x}{30}\right)^6}$$

$$B = \text{«μεσήλικας»} = \{(x,\mu_B(x)) | x \in U\},\ \mu_B(x) = \frac{1}{1+\left(\frac{x-50}{20}\right)^6}$$

$$C = \text{«ηλικιωμένος»} = \{(x,\mu_C(x)) | x \in U\},\ \mu_C(x) = \frac{1}{1+\left(\frac{x-100}{30}\right)^6}$$

Σχήμα 3.8

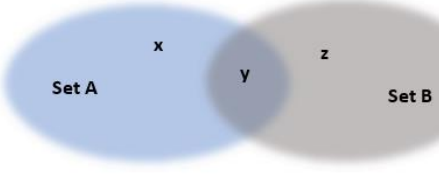
Ένα βασικό χαρακτηριστικό των ασαφών συνόλων είναι ότι δεν υπάρχουν αυστηροί κανόνες σχετικά με τον τρόπο ορισμού των συναρτήσεων συμμετοχής τους. Τόσο η μαθηματική μορφή της συνάρτησης όσο και οι παράμετροι εξαρτώνται από την είσοδο των εμπειρογνωμόνων. Εφόσον οι συναρτήσεις συμμετοχής είναι συνεπείς, σε

συγκριτική βάση, το συμπέρασμα που βασίζεται στα ασαφή σύνολα εξακολουθεί να έχει νόημα.

Οι συναρτήσεις συμμετοχής είναι συνήθως απλές για τα ασαφή σύνολα. Συχνά είναι γραμμικές και έχουν τη μορφή τριγώνου, τραπεζοειδούς, L ή r. Μπορεί επίσης να είναι γκαουσιανές ή γάμμα. Διαφορετικοί άνθρωποι μπορεί να έχουν τις δικές τους συναρτήσεις συμμετοχής για ένα ασαφές σύνολο λόγω διαφορετικών επιπέδων γνώσης και εμπειρίας. Ωστόσο, σε γενικές γραμμές, μπορεί να εννοούν παρόμοια πράγματα όταν αναφέρονται σε ένα ασαφές σύνολο. Για παράδειγμα, οι άνθρωποι μπορεί να έχουν την ίδια άποψη ότι ένας ο οποίος αιτείται δάνειο με υψηλό πιστωτικό βαθμό είναι πιθανό να εγκριθεί η αίτηση με σχετικά χαμηλό επιτόκιο ενυπόθηκου δανείου. Εδώ, το "υψηλό πιστωτικό σκορ" ταιριάζει φυσικά στην περιγραφή ενός ασαφούς συνόλου. Όμως, διαφορετικοί αξιολογητές κινδύνου μπορεί να έχουν διαφορετικές συναρτήσεις συμμετοχής για το ασαφές σύνολο "υψηλό πιστωτικό σκορ". Τα ασαφή σύνολα μας επιτρέπουν να δημιουργήσουμε ένα σύστημα χρησιμοποιώντας την καθημερινή μας γλώσσα και τις μεθόδους συλλογισμού.

3.3.3 Πράξεις Ασαφών Συνόλων

Όπως και στην κλασική θεωρία συνόλων, τα ασαφή σύνολα έχουν τις δικές τους πράξεις, όπως η ένωση, η τομή και συμπλήρωμα. Διαφορετικά από τις πράξεις στα κλασικά σύνολα, οι πράξεις στα ασαφή σύνολα βασίζονται στη συνάρτηση συμμετοχής. Η διαφορά τους σε σχέση με την κλασσική θεωρία συνόλων είναι ότι στην κλασσική θεωρία ο βαθμός συμμέτοχης περιορίζεται στο {0,1} ενώ στην ασαφή θεωρία μπορεί να παίρνει τιμές στο [0,1]. Επίσης, ενώ στην κλασσική θεωρία οι βασικές πράξεις είναι μοναδικές, στην ασαφή θεωρία για κάθε τελεστή υπάρχει μία κλάση συναρτήσεων. Ένας τύπος ασαφούς κανόνα εξαγωγής συμπερασμάτων ονομάζεται κανόνας max -min, που φαίνεται στο Σχήμα 3.9 και εφαρμόζεται σε εξαγωγή συμπερασμάτων.



$\mu_A(x) = 0.5 \quad \mu_B(x) = 0.1$
$\mu_A(y) = 0.6 \quad \mu_B(y) = 0.4$
$\mu_A(z) = 0.1 \quad \mu_B(z) = 0.7$

	$A \cup B = \max(\mu_A, \mu_B)$	$A \cap B = \min(\mu_A, \mu_B)$	$\overline{A} = 1 - \mu_A$
x	0.5	0.1	0.5
y	0.6	0.4	0.4
z	0.7	0.1	0.9

Σχήμα3.9

Θεωρούμε δύο ασαφή σύνολα Α και Β ενός συνόλου αναφοράς Χ. Έχουμε ότι:

$$A = \{(x, \mu_A(x))\}, \mu_A(x) \epsilon [0,1]$$
$$B = \{(x, \mu_B(x))\}, \mu_B(x) \epsilon [0,1]$$

Οι πράξεις των δύο παραπάνω ασαφών συνόλων περιγράφονται επί των συναρτήσεων συμμετοχής του $\mu_A(x)$ και $\mu_B(x)$.

Ισότητα

Δύο ασαφή σύνολα είναι ίσα αν και μόνο εάν για κάθε $x \epsilon X$ ισχύει ότι

$$\mu_A(x) = \mu_B(x)$$

και συμβολίζονται $A = B$.

Συμπλήρωμα

Τα ασαφή σύνολα A και $\bar{A}$ καλούνται συμπληρωματικά εάν

$$\mu_A(x) + \mu_{\bar{A}}(x) = 1$$

Ένωση

Η ένωση των ασαφών συνόλων Α και Β συμβολίζεται ως $A \cup B$ ορίζεται ως

$$\mu_{AUB}(x) = \max(\mu_A(x), \mu_B(x)) \ , x \epsilon X$$

Τομή

Η τομή των ασαφών συνόλων Α και Β συμβολίζεται ως $A \cap B$ και ορίζεται ως

$$\mu_{A \cap B}(x) = min(\mu_A(x), \mu_B(x)) \ , x \epsilon X$$

3.3.4 Ασαφής Συλλογιστική

Η ασαφής συλλογιστική είναι η διαδικασία με την οποία εξάγονται συμπεράσματα, δηλαδή είναι ένα υπολογιστικό σύστημα το οποίο αξιολογεί ασαφείς κανόνες της μορφής «$Αν\ το\ x\ είναι\ A\ τότε\ το\ y\ είναι\ B$».

Το σύστημα Ασαφούς Συμπερασματολογίας (Fuzzy Inference System – FIS) αποτελεί την πιο ευρέως διαδεδομένη εφαρμογή της ασαφούς λογικής και διακρίνεται σε

δύο στάδια, στην βάση γνώσης (knowledge base) και στο σύστημα επεξεργασίας. Κατά το πρώτο στάδιο καθορίζονται οι κανόνες συμμετοχής και οι ασαφείς κανόνες. Κατά το δεύτερο στάδιο, εισάγονται οι μεταβλητές οι οποίες είναι αριθμητικές, και μέσω την διαδικασίας Ασαφοποίησης (Fuzzification) μεταφράζονται σε ασαφής ποσότητες σε μορφή γλωσσικών μεταβλητών. Στην συνέχεια, παράγονται τα ασαφή συμπεράσματα μέσω των κανόνων που έχουν τεθεί στο πρώτο στάδιο. Τα συμπεράσματα αυτά, έχουν μορφή κλασσικών μεταβλητών και μέσω της Αποασαφοποίησης μεταφράζονται σε κλασσικούς αριθμούς.

3.3.5 Αποασαφοποίηση

Defuzzification

Η Αποασαφοποίηση (Defuzzification) είναι η διαδικασία εκτίμησης της τιμής της εξαρτημένης μεταβλητής με βάση το ασαφές σύνολο που προκύπτει μετά την εφαρμογή του ασαφούς κανόνα εξαγωγής συμπερασμάτων. Διαφορετικές μέθοδοι είναι κατάλληλες σε διαφορετικές καταστάσεις. Τρεις βασικές μορφές περιγράφονται κατωτέρω:

1**. Μέθοδος του μέσου όρου**: Η μέση αριθμητική τιμή της εξαρτημένης μεταβλητής στην έξοδο ασαφούς συνόλου.

2. **Μέθοδος του μέσου όρου του μέγιστου**: Η μέση αριθμητική τιμή της εξαρτημένης μεταβλητής με τον μέγιστο βαθμό αλήθειας στο ασαφές σύνολο εξόδου.

3. **Μέθοδος κεντροειδούς:** Η σταθμισμένη μέση αριθμητική τιμή της εξαρτημένης μεταβλητής στο το ασαφές σύνολο εξόδου. Το βάρος είναι ο βαθμός αλήθειας.

3.3.6 Σύστημα ασαφούς λογικής

Fuzzy Logic System

Με όλα τα στοιχεία, ένα σύστημα ασαφούς λογικής μπορεί να κατασκευαστεί με τα ακόλουθα βήματα:

Βήμα 1: Οι ανεξάρτητες μεταβλητές επιλέγονται ως οι βασικοί προσδιοριστικοί παράγοντες ή δείκτες της εξαρτημένης μεταβλητής.

Βήμα 2: Δημιουργούνται ασαφή σύνολα τόσο για τις ανεξάρτητες όσο και για τις εξαρτημένες μεταβλητές. Αντί της χρήσης αριθμητικής τιμής, χρησιμοποιούνται ασαφή σύνολα από την άποψη της ανθρώπινης γλώσσας για να για την περιγραφή μιας μεταβλητής. Ο βαθμός αλήθειας ότι κάθε μεταβλητή ανήκει σε ένα συγκεκριμένο ασαφές σύνολο καθορίζεται από τη συνάρτηση συμμετοχής.

Βήμα 3: Οι κανόνες εξαγωγής συμπερασμάτων κατασκευάζονται στο σύστημα. Μια ασαφής αντιστάθμιση μπορεί να χρησιμοποιηθεί για τη διόρθωση των συνάρτηση συμμετοχής σύμφωνα με την περιγραφή των κανόνων εξαγωγής συμπερασμάτων.

Βήμα 4: Το ασαφές σύνολο εξόδου της εξαρτημένης μεταβλητής δημιουργείται με βάση τους ανεξάρτητες μεταβλητές και τους κανόνες συμπερασμού. Μετά την Αποασαφοποίηση, ένα αριθμητικό τιμή μπορεί να χρησιμοποιηθεί για την αναπαράσταση του ασαφούς συνόλου εξόδου.

Βήμα5: Το αποτέλεσμα χρησιμοποιείται στη συνέχεια για τη λήψη τεκμηριωμένων αποφάσεων.

3.4 Διαχείριση Δεδομένων

Ένα βασικό κομμάτι της ανάλυσης δεδομένων και πριν την εφαρμογή όλων των παραπάνω τεχνικών που παρουσιάστηκαν, είναι η διαχείριση των δεδομένων, για να μην προκύψουν προβλήματα κατά την υπόλοιπη διαδικασία. Παρακάτω, παρουσιάζονται κάποιες βασικές τεχνικές που χρησιμοποιήθηκαν στην παρούσα εργασία.

3.4.1 Καθαρισμός Δεδομένων

Data cleaning

Ο καθαρισμός δεδομένων ορίζεται ως η διαδικασία διόρθωσης ή αφαίρεσης εσφαλμένων, κατεστραμμένων, εσφαλμένα μορφοποιημένων, διπλότυπων ή ελλιπών δεδομένων σε ένα σύνολο δεδομένων. Όταν συνδυάζονται πολλαπλές πηγές δεδομένων, υπάρχει πιθανότητα αντιγραφής ή εσφαλμένης επισήμανσης δεδομένων. Σε περίπτωση που τα δεδομένα είναι λανθασμένα, τότε τα αποτελέσματα και οι αλγόριθμοι δεν είναι αναξιόπιστοι, παρόλο που μπορεί να φαίνονται σωστά.

Αρχικά, πρέπει τα δεδομένα να ελεγχθούν ως προς το αν υπάρχουν διπλές γραμμές ή στήλες, για να διασφαλιστεί η ακεραιότητα των δεδομένων και για την αντιμετώπιση αυτού πρέπει να εξαλειφθούν από τα σύνολα δεδομένων. Αυτό μπορεί να συμβαίνει λόγω του ότι πολλές φορές όταν συνδυάζονται σύνολα δεδομένων από πολλά μέρη, υπάρχει δυνατότητα για τη δημιουργία διπλότυπων δεδομένων. Σε περίπτωση ύπαρξης διπλότυπων καταργούνται οι ανεπιθύμητες παρατηρήσεις από το σύνολο δεδομένων, συμπεριλαμβανομένων διπλών ή άσχετων παρατηρήσεων.

Επιπλέον, είναι σημαντικό να γίνει έλεγχος ελλειπουσών τιμών (missing values), διότι αν υπάρχουν χαμένα δεδομένα μπορεί να προκύψει μεροληψία στις εκτιμήσεις που προέρχονται από ένα στατιστικό μοντέλο και αναδεικνύουν ακατάλληλες κοινές στατιστικές μεθόδους ή δύσκολα εφαρμόσιμές. Υπάρχουν μερικοί τρόποι αντιμετώπισης δεδομένων που λείπουν. Κανένα από αυτά δεν είναι βέλτιστο, αλλά όλα μπορούν να ληφθούν υπόψη: (1) μπορούν να απορριφθούν οι παρατηρήσεις που έχουν ελλείπουσες τιμές, αλλά υπάρχει κίνδυνο να χαθούν πληροφορίες, επομένως χρειάζεται προσοχή, (2) μπορούν να εισαχθούν τιμές που λείπουν με βάση άλλες παρατηρήσεις αλλά και πάλι,

υπάρχει μια πιθανότητα να χαθεί η ακεραιότητα των δεδομένων επειδή μπορεί να λειτουργεί από υποθέσεις και όχι από πραγματικές παρατηρήσεις, (3) μπορεί να αλλάξει ο τρόπος που χρησιμοποιούνται τα δεδομένα για την αποτελεσματική πλοήγηση σε μηδενικές τιμές. Σε περίπτωση εύρεσης ελλειπουσών τιμών υπάρχουν πέντε ειδικές μέθοδοι για τα χαμένα δεδομένα, οι οποίες είναι η μέθοδος Listwise, Pairwise deletio, η Αντικατάσταση από τον Μέσο (Mean Substitution), η Hot-Deck μέθοδο, και η Παλινδρόμηση (Καλπινέλλη Ε., 2004).

Επιπρόσθετα, τα περισσότερα δεδομένα στην πραγματική ζωή έρχονται με κατηγορικές τιμές συμβολοσειρών ενώ τα περισσότερα μοντέλα μηχανικής μάθησης λειτουργούν μόνο με ακέραιες τιμές ή με άλλες διαφορετικές τιμές που μπορούν να είναι κατανοητές για το μοντέλο. Συνεπώς, στον τομέα της επιστήμης δεδομένων, κατά την προετοιμασία των δεδομένων πριν προχωρήσουμε στη μοντελοποίηση μία ακόμη εργασία που εκτελείται είναι η κωδικοποίηση κατηγορικών δεδομένων, η οποία είναι μια διαδικασία μετατροπής κατηγορικών δεδομένων σε ακέραια μορφή, έτσι ώστε τα δεδομένα με μετατρεπόμενες κατηγορικές τιμές να μπορούν να παρέχονται στα μοντέλα για να δώσουν και να βελτιώσουν τις προβλέψεις. Μερικές τεχνικές που μπορούν να χρησιμοποιηθούν για τον λόγο αυτό είναι One-Hot Encoding, Dummy Encoding, Label Encoding, Binary Encoding, Count Encoding, Target Encoding. Στην παρούσα διπλωματική προτείνεται η τεχνική Label Encoding η οποία αναθέτει αύξουσες αριθμητικές τιμές σε κάθε κατηγορία για τον χειρισμό κατηγορικών μεταβλητών. Αυτό το σύστημα κωδικοποίησης διευκόλυνε την ηλεκτρονική παρουσίαση των κατηγορικών μεταβλητών σε αριθμητική μορφή εντός του συνόλου δεδομένων.

3.4.2 Υπερδειγματοληψία - Υποδειγματοληψία

Επίσης, σε προβλήματα ταξινόμησης, είναι σημαντικό να διασφαλιστεί ότι οι κατηγορίες της μεταβλητής-στόχου είναι ισορροπημένες. Αυτή η ισορροπία είναι ζωτικής σημασίας για τη δημιουργία ισχυρών εκτιμήσεων και αξιόπιστων αξιολογήσεων της απόδοσης του μοντέλου. Στην περίπτωση που δεν υπάρχει ισορροπία, προτείνεται είτε η υπερδειγματιληψία (oversampling), είτε η υποδειγματοληψία (undersampling), οι οποίες είναι τεχνικές που χρησιμοποιούνται για την προσαρμογή της κατανομής κλάσεων ενός συνόλου δεδομένων. Στην παρούσα διπλωματική χρησιμοποιήθηκε υπερδειγματοληψία η οποία περιλαμβάνει τη συμπλήρωση των δεδομένων εκπαίδευσης με πολλαπλά αντίγραφα ορισμένων από τις τάξεις μειοψηφίας. Υπάρχει ένας αριθμός διαθέσιμων μεθόδων για την υπερδειγματοληψία ενός συνόλου δεδομένων που χρησιμοποιείται σε ένα τυπικό πρόβλημα ταξινόμησης όπως οι Augmentation, ADASYN, Random oversampling. Όμως, η πιο κοινή τεχνική είναι γνωστή ως SMOTE (Synthetic Minority Over-sampling Technique), όπου τα συνθετικά δείγματα παράγονται για την κατηγορία μειοψηφίας. Αυτός ο αλγόριθμος βοηθά να ξεπεραστεί το πρόβλημα υπερπροσαρμογής που δημιουργείται από την τυχαία υπερδειγματοληψία. Εστιάζει στον χώρο χαρακτηριστικών για τη δημιουργία νέων περιπτώσεων με τη βοήθεια παρεμβολής

μεταξύ των θετικών περιπτώσεων που βρίσκονται μαζί. Πιο ειδικά, ο αλγόριθμος της SMOTE ακολουθεί τα παρακάτω βήματα:

Βήμα 1: Ρύθμιση του συνόλου κλάσης μειοψηφίας A , για κάθε ένα $x \epsilon A$, οι k-πλησιέστεροι γείτονες του x λαμβάνονται υπολογίζοντας την Ευκλείδεια απόσταση μεταξύ του x και κάθε άλλου δείγματος στο σύνολο A .

Βήμα 2: Ο ρυθμός δειγματοληψίας N ρυθμίζεται σύμφωνα με την μη ισορροπημένη αναλογία. Για κάθε $x \epsilon A$, N παραδείγματα (δηλαδή$x_1, x_2, \dots, x_N$) επιλέγονται τυχαία από τους k-πλησιέστερους γείτονές του και κατασκευάζουν το σύνολο A_1.

Βήμα 3: Για κάθε παράδειγμα $x_k \in A_1$ $(k = 1, 2, 3 \dots N)$, ο ακόλουθος τύπος χρησιμοποιείται για τη δημιουργία ενός νέου παραδείγματος:

$$x' = x + rand(0,1) * |x - x_k|$$

στον οποίο το $rand(0, 1)$ αντιπροσωπεύει τον τυχαίο αριθμό μεταξύ 0 και 1.

3.4.3 Pearson Correlation

Στη στατιστική, ο συντελεστής συσχέτισης Pearson είναι ένας συντελεστής συσχέτισης που μετρά τη γραμμική συσχέτιση μεταξύ δύο συνόλων δεδομένων. Είναι ο λόγος μεταξύ της συνδιακύμανσης δύο μεταβλητών και του γινόμενου των τυπικών αποκλίσεων τους. Είναι ουσιαστικά, μια κανονικοποιημένη μέτρηση της συνδιακύμανσης, καταλήγοντας το αποτέλεσμα της πάντα σε μια τιμή μεταξύ του -1 και του 1. Ο συντελεστής συσχέτισης Pearson, όταν εφαρμόζεται σε έναν πληθυσμό, συμβολίζεται συνήθως με το ελληνικό γράμμα ρ και δίνεται από τον τύπο

$$\rho = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

όπου (X, Y) ζεύγος τυχαίων μεταβλητών, $cov(X, Y)$ είναι η συνδιακύμανση, σ_X η τυπική απόκλιση του X και σ_Y η τυπική απόκλιση του Y.

3.4.4 Μέθοδοι Επιλογής Μεταβλητών

Feature Selection

Η επιλογή μεταβλητών είναι η διαδικασία μείωσης του αριθμού των μεταβλητών εισόδου κατά την ανάπτυξη ενός μοντέλου πρόβλεψης. Η μείωση του αριθμού των μεταβλητών εισόδου, συμβάλλει στην μείωση του υπολογιστικού κόστους της μοντελοποίησης, και σε ορισμένες περιπτώσεις στη βελτίωση της απόδοσης του μοντέλου. Οι πιο γνωστοί μέθοδοι είναι οι Βηματικές μέθοδοι (Forward Feature Selection και ο Backward Feature Elimination) και Μέθοδοι ποινικοποιημένης παλινδρόμησης (ο

Ridge Regression και ο Lasso Regression). Στην παρούσα εργασία χρησιμοποιήθηκε η τεχνική Backward elimination η οποία υλοποιήται ως εξής:

- Αρχικά, συμπεριλαμβάνονται στο μοντέλο όλες οι επεξηγηματικές μεταβλητές p, που είναι διαθέσιμες.
- Στην συνέχεια, αφαιρούνται μία-μία οι επεξηγηματικές μεταβλητές, αρχίζοντας από αυτή που δίνει τη μικρότερη τιμή κριτηρίου σύγκρισης του μοντέλου (σύμφωνα με τα κριτήρια πληροφορίας AIC ή BIC ή της στατιστικής συνάρτησης t).
- Όλη η παραπάνω διαδικασία επαναλαμβάνεται, μέχρις ότου το κριτήριο σύγκρισης του μοντέλου για την απομάκρυνση οποιασδήποτε άλλης επεξηγηματικής μεταβλητής να μην βελτιώνεται περισσότερο, συνεπώς και σταματάει.

Είναι φανερό ότι, όταν μια επεξηγηματική μεταβλητή έχει αφαιρεθεί τότε δεν μπορεί να την συμπεριληφθεί πάλι στο μοντέλο, ακόμα και αν αυτή εμφανίζεται σε κάποιο βήμα ως στατιστικά σημαντική. Επιπλέον, έχοντας ως δεδομένο ότι ξεκινάει από το μοντέλο που περιέχει όλες τις ανεξάρτητες μεταβλητές, η διαδικασία απαλοιφής θα είναι αργή όταν υπάρχει πολύ μεγάλο πλήθος μεταβλητών.

Έχοντας ως κριτήριο απόρριψης τη στατιστική συνάρτηση t, για να εξαχθεί μία μεταβλητή από το μοντέλο η μέθοδος Backward elimination ξεκινάει με την προσαρμογή του πλήρους μοντέλου:

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i \, , i = 1, \dots , n$$

Έπειτα, εντοπίζεται η μεταβλητή που θα εξαχθεί από το μοντέλο, $Xj1$, η οποία δίνει τη μικρότερη στατιστικά σημαντική τιμή της t, αφού πρώτα υπολογιστούν οι λόγοι, δηλαδή ισχύει:

$$\left| \frac{\hat{b}_{j_1}}{s(\hat{b}_{j_1})} \right| = min \left\{ \left| \frac{\hat{b}_j}{s(\hat{b}_j)} \right| , j = 1, \dots , p \right\}$$

και

$$\left| \frac{\hat{b}_{j_1}}{s(\hat{b}_{j_1})} \right| \leq \left(\frac{\alpha}{2} \right) \times t_{n-p+1}$$

Σε περίπτωση όμως που δεν πληρείται το κριτήριο πιο πάνω, δηλαδή έχουμε

$$\left|\frac{\hat{b}_{j_1}}{s(\hat{b}_{j_1})}\right| > \left(\frac{\alpha}{2}\right) \times t_{n-p+1}$$

σημαίνει ότι όλες οι επεξηγηματικές μεταβλητές θεωρούνται σημαντικές οπότε η αναζήτηση τερματίζει, με βέλτιστο να είναι το πλήρες μοντέλο.

Στην περίπτωση που ικανοποιείται όμως η παραπάνω σχέση και η μεταβλητή $Xj1$ έχει αφαιρεθεί, η μέθοδος συνεχίζει στην εξαγωγή της δεύτερης μεταβλητής, προσαρμόζοντας τα γραμμικά μοντέλα της μορφής:

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i \, , i = 1, \dots \, , n \, , j \neq j_1$$

Έπειτα, εντοπίζεται η δεύτερη μεταβλητή που θα εξαχθεί από το μοντέλο, $Xj2$, αφού ξανά υπολογιστούν οι λόγους, για την οποία ισχύει:

$$\left|\frac{\hat{b}_{j_2}}{s(\hat{b}_{j_2})}\right| = min\left\{\left|\frac{\hat{b}_j}{s(\hat{b}_j)}\right| \, , j = 1, \dots, p, j \neq j_1\right\}$$

και

$$\left|\frac{\hat{b}_{j_2}}{s(\hat{b}_{j_2})}\right| \leq \left(\frac{\alpha}{2}\right) \times t_{n-p+1}$$

Η παραπάνω διαδικασία επαναλαμβάνεται, μέχρι να μην ικανοποιηθεί μία ανισότητα της παραπάνω μορφής.

Κεφάλαιο 4ο:
Πολυκριτήρια Ανάλυση Αποφάσεων

4.1 Εισαγωγή

Η Πολυκριτήρια Ανάλυση Αποφάσεων (Multicriteria Decision Making - MCDM) είναι ένας σημαντικός τομέας της επιχειρησιακής έρευνας, και τα τελευταία χρόνια έχει ευδοκιμήσει αρκετά. Σημαντικό παράγοντα της ανάπτυξης της αποτέλεσε η διαπίστωση του ότι, η επίλυση πολύπλοκων προβλημάτων λήψης αποφάσεων δεν μπορεί να επιτευχθεί μέσω μίας μονοδιάστατης ανάλυσης. Τις περισσότερες φορές πρέπει να εξεταστούν πολλοί παράμετροι και παράγοντες για τη λήψη της καταλληλότερης απόφασης με αποτέλεσμα την αποθάρρυνση των αναλυτών για την λήψη της πιο ρεαλιστικής προσέγγισης. Η επίλυση του προβλήματος αυτού, συνιστά το κύριο αντικείμενο της πολυκριτήριας ανάλυσης αποφάσεων.

Το βασικό χαρακτηριστικό της πολυκριτήριας ανάλυσης, δεν είναι απλά η σύνθεση των παραμέτρων σε ένα πρόβλημα, αλλά η σύνθεση τους υπό το πρίσμα των προτιμήσεων κα των αποφάσεων του αποφασίζοντα. Δηλαδή, το αποτέλεσμα της ανάλυσης για την επίλυση ενός προβλήματος, εν τέλη εξαρτάται από τον ίδιο τον αποφασίζοντα. Υπό τις συνθήκες αυτές, η πολυκριτήρια ανάλυση εμβαθύνει στην έρευνα μεθοδολογιών που έχουν σχέση με την ανάλυση, την μαθηματική μοντελοποίηση και την αναπαράσταση των προτιμήσεων του εκάστοτε αποφασίζοντα. Απώτερος στόχος είναι ο εντοπισμός των κύριων χαρακτηριστικών γνωρισμάτων του υπάρχοντος προβλήματος καθώς και των ιδιομορφιών των διαθέσιμων εναλλακτικών λύσεων.

4.2 Ιστορική Αναδρομή

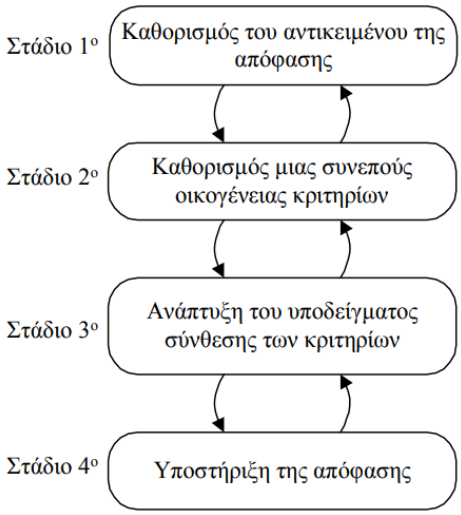
Η λήψη αποφάσεων αποτελεί μία σημαντική διαδικασία από τα πρώτα χρόνια της ανθρωπότητας καθώς, αν και δεν είχε μαθηματική μορφή, σχετίζεται με την ανάλυση διάφορων παραγόντων για την λήψη μίας απόφασης, η οποία θα εξαρτάται από την εμπειρία του κάθε ατόμου. Η πρώτη επιστημονική αντιμετώπιση του προβλήματος, θεωρείται η εργασία του Pareto (1896), μέσω της οποίας τέθηκαν οι βάσεις και εισήχθη μία από τις βασικότερες έννοιες της σύγχρονης πολυκριτήριας ανάλυσης, η έννοια της

αποτελεσματικότητας. Αργότερα, με την εργασία του Koopmans (1951), εισήχθη η έννοια του αποτελεσματικού συνόλου, δηλαδή του συνόλου των εναλλακτικών δραστηριοτήτων οι οποίες δεν ελέγχονται από κάποια άλλη εναλλακτική, ενώ παράλληλα με την συμβολή Von Neumann και Morgenstern (1944), αναπτυχθηκε η θεωρία χρησιμότητας, η οποία αποτελεί τη βάση της πολυκριτήριας ανάλυσης αποφάσεων. Οι παραπάνω έρευνες αποτέλεσαν την βάση για τα επόμενα χρόνια και έναυσμα για περισσότερη έρευνα πάνω στον τομέα. Το 1961, οι Charnes και Cooper έκαναν έρευνα για τη σύνδεση της θεωρίας του γραμμικού προγραμματισμού με την πολυκριτήρια ανάλυση, καθώς επίσης το 1965, ο Fishburn ασχολήθηκε με την διεύρυνση του τομέα της θεωρίας χρησιμότητας σε προβλήματα λήψης αποφάσεων υπό το πρίσμα πολλαπλών κριτηρίων. Κατά τα τέλη της δεκαετία του 1960, ο Roy (1968), ανέπτυξε τη θεωρία των σχέσεων υπεροχής. Τις επόμενες δύο δεκαετίες (1970-1990), τα πολύπλοκα προβλήματα λήψης αποφάσεων γνώρισαν τεράστια άνθηση, και σε θεωρητικό αλλά και σε πρακτικό επίπεδο. Σημαντική υπήρξε η ταχύτατη τεχνολογική πρόοδος και η συμβολή της πληροφορικής, ώστε να υλοποιηθούν όλες οι μεθοδολογικές εξελίξεις της πολυκριτήριας ανάλυσης σε ολοκληρωμένα πληροφορικά συστήματα, τα οποία ταυτόχρονα βοήθησαν και στην ανάπτυξη των πρακτικών εφαρμογών της πολυκριτήριας ανάλυσης.

4.3 Βασικές έννοιες και μεθοδολογίες

Βασικό αντικείμενο της πολυκριτήριας ανάλυσης αποφάσεων αποτελεί η αποσαφήνιση και η αξιοποίηση υποδειγμάτων, αποτελούμενα από όλες τις παραμέτρους ενός προβλήματος, με σκοπό να υποστηριχθεί ο αποφασίζοντας στη λήψη ορθολογικών αποφάσεων με βάση τις αξίες και τις προτιμήσεις που τον διέπουν. Ο παραπάνω στόχος είναι μια περίπλοκη διαδικασία η οποία δεν έχει τις βέλτιστες λύσεις, αλλά τις πιο ικανοποιητικές, οι οποίες ανταποκρίνονται όσο γίνεται στις προτιμήσεις του ατόμου που καλείται να αποφασίσει. Ο Roy (1985) ανέπτυξε ένα γενικό πλαίσιο με στόχο την αντιμετώπιση πολυδιάστατων προβλημάτων λήψης αποφάσεων, το οποίο περιλαμβάνει τέσσερα στάδια (Σχήμα 4.1). Μέσω αυτού, αποτυπώνεται η φιλοσοφία των μεθοδολογιών του χώρου και συνθέτη το θεμέλιο κάθε πολυκριτήριας προσέγγισης.

Κατά το πρώτο στάδιο, καθορίζεται το πρόβλημα και το σύνολο όλων των δυνατών λύσεων Α και αποφάσεων. Το σύνολο Α μπορεί να είναι συνεχές ή διακριτό. Στην πρώτη περίπτωση, τα όρια του θέτονται είτε από τον ίδιο τον ειδικό λήψης αποφάσεων είτε από το σύνολο δυνατών λύσεων. Στην άλλη περίπτωση, όπου το σύνολο Α είναι διακριτό, θεωρείται ότι υπάρχει ένα συγκεκριμένο σύνολο εναλλακτικών, οι οποίες λαμβάνονται υπόψιν και αναλύονται με στόχο την λήψη της καλύτερης δυνατής απόφαση. Συνεπώς, έπειτα από προσδιορισμό του συνόλου Α, καθορίζεται και ο τρόπος με τον οποίο εξετάζονται οι εναλλακτικές αποφάσεις, ώστε το υπάρχον πρόβλημα να επιλύεται έπειτα από το αποτέλεσμα της ανάλυσης.



Σχήμα 4.1 (Roy B, 1985)

Η εξέταση των εναλλακτικών στρατηγικών επίλυσης ενός προβλήματος μπορεί να πραγματοποιηθεί με μια από τις παρακάτω τέσσερις προβληματικές:

- Επιλογής (choice), όπου αναφέρεται στην επιλογή μίας ή περισσότερων εναλλακτικών λύσεων από τις πλέον κατάλληλες**.**
- Ταξινόμησης (classification/sorting), όπου αναφέρεται στην ταξινόμηση των αποφάσεων σε προκαθορισμένες ομοιογενείς κλάσεις.
- Κατάταξης (ranking), όπου αναφέρεται στην κατάταξη των δραστηριοτήτων σε φθίνουσα σειρά**.**
- Περιγραφή (description), όπου αναφέρεται στην περιγραφή των εναλλακτικών δραστηριοτήτων όπως αυτά προκύπτουν από τις επιδόσεις τους στα επιμέρους κριτήρια αξιολόγησης.

Ανάλογα το πρόβλημα που εξετάζεται, επιλέγεται και η κατάλληλη προβληματική. Υπάρχουν αρκετές περιπτώσεις όπου για την αντιμετώπιση ενός προβλήματος δε χρησιμοποιείται αποκλειστικά μία προβληματική αλλά πιθανόν να απαιτηθεί ο συνδυασμός δύο προβληματικών για την αποτελεσματικότερη αντιμετώπιση του.

Κατά το δεύτερο στάδιο της διαδικασίας, προσδιορίζονται τα κριτήρια που καθορίζουν το αποτέλεσμα της ανάλυσης των εναλλακτικών του συνόλου Α. Ως κριτήριο, ορίζεται κάθε πραγματική συνάρτηση g μέσω της οποίας η συμπεριφορά των εναλλακτικών αποφάσεων αναπαρίσταται σε έναν πραγματικό αριθμό, έτσι ώστε για δύο οποιεσδήποτε εναλλακτικές x και y να ισχύουν:

$$g(x) > g(y) \Leftrightarrow x \succ y \text{ (Η x είναι προτιμότερη της y)}$$

$$g(x) = g(y) \Leftrightarrow x \sim y \text{ (Η x είναι αδιάφορη της y)}$$

Παρόλο που η αριθμητική περιγραφή της εναλλακτικής x σε ένα χαρακτηριστικό είναι μεγαλύτερη σε σχέση με την αντίστοιχη αριθμητική περιγραφή μιας άλλης εναλλακτικής y, αυτό δεν σημαίνει πως η x υπερέχει έναντι της y. Το χαρακτηριστικό δεν ορίζει καμία προτιμησιακή συμπεριφορά. Σε αυτό το στάδιο της διαδικασίας ανάλυσης το σύνολο των κριτηρίων $G = \{g_1, g_2, \dots, g_n\}$ που εντοπίζονται πρέπει ικανοποιεί τις παρακάτω βασικές ιδιότητες ώστε να αποτελεί μια συνεπή οικογένεια κριτηρίων:

- Μονοτονία (monotonicity), όπου ένα σύνολο κριτηρίων χαρακτηρίζεται με την ιδιότητα της μονοτονίας, αν και μόνο αν για οποιεσδήποτε δυο εναλλακτικές x' και x'' τέτοιες ώστε $x'_i > x''_i$ για κάποιο κριτήριο x_i και $x'_j > x''_j$ για όλα τα υπόλοιπα κριτήρια, συμβολίζεται ως $x'_i \succ x''_i$.
- Επάρκεια (exhaustivity), όπου ένα σύνολο κριτηρίων χαρακτηρίζεται με την ιδιότητα της επάρκειας, αν και μόνο αν για κάθε ζεύγος εναλλακτικών x' και x'' τέτοιες ώστε $x'_i = x''_i$ για όλα τα κριτήρια x_i, συμβολίζεται ως $x'_i \sim x''_i$.
- Μη πλεονασμός (non-redundancy) όπου ένα σύνολο κριτηρίων χαρακτηρίζεται με την ιδιότητα του μη πλεονασμού, αν και μόνο αν η παράλειψη ενός οποιουδήποτε κριτηρίου καταλήγει σε παραβίαση των ιδιοτήτων της μονοτονίας ή της επάρκειας.

Αφού ολοκληρωθεί η διαδικασία καθορισμού όλων των κριτηρίων μεταβαίνουμε στο τρίτο στάδιο της διαδικασίας. Στο σημείο αυτό προσδιορίζεται το υπόδειγμα σύνθεσης όλων των κριτηρίων στο οποίο θα αντιμετωπιστεί το αντικείμενο του προβλήματος όπως αυτό έχει χαρακτηριστεί στο αρχικό στάδιο της διαδικασίας (επιλογή, κατάταξη, ταξινόμηση, περιγραφή). Στο τέταρτο και τελευταίο στάδιο, συμπεριλαμβάνονται όλες εκείνες οι στρατηγικές επίλυσης οι οποίες θα συνδράμουν στην κατανόηση των αποτελεσμάτων του υποδείγματος σύνθεσης των κριτηρίων, όπως αυτά καθορίστηκαν στο προηγούμενο στάδιο καθώς και τη διαδικασία με την οποία έγινε η εξαγωγή των αποτελεσμάτων αυτών. Επομένως, αυτός που καλείται να λάβει την απόφαση θα έχει την δυνατότητα να υλοποιήσει με επιτυχία τα αποτελέσματα της ανάλυσης και να επιχειρηματολογήσει υπέρ των αποφάσεων του, εάν και όταν αυτό κριθεί αναγκαίο.

4.4 Κύρια θεωρητικά ρεύματα

Για την επίλυση προβλημάτων λήψης αποφάσεων, παρατηρούνται αρκετές διαφοροποιήσεις στο χώρο της πολυκριτήριας ανάλυσης ως προς τις μεθοδολογικές προσεγγίσεις. Οι κύριες διαφοροποιήσεις μεταξύ των προσεγγίσεων αυτών εμφανίζονται στη μορφή των υποδειγμάτων τα οποία αναπτύσσονται καθώς και στη διαδικασία που χρησιμοποιείται για την ανάπτυξη αυτών. Αξιοποιώντας το στοιχείο αυτό, αρκετοί ερευνητές έχουν προτείνει διάφορες ομαδοποιήσεις. Αρχικά, ο Roy (1985) έχοντας ως

γνώμονα τη μορφή των υποδειγμάτων που αναπτύσσονται πρότεινε μια ομαδοποίηση με τις ακόλουθες τρεις βασικές κατηγορίες:

- Προσεγγίσεις μοναδικής σύνθεσης των κριτηρίων, αγνοώντας κάθε συγκριτικότητα μεταξύ των εναλλακτικών απόφασης,
- Προσεγγίσεις που βασίζονται στις σχέσεις υπεροχής, λαμβάνοντας υπόψιν την πιθανή συγκριτικότητα μεταξύ των εναλλακτικών απόφασης,
- Αλληλεπιδραστικές προσεγγίσεις.

Αργότερα, οι Pardalos et al. (1995) πρότειναν μια διαφορετική ομαδοποίηση των πολυκριτήριων προσεγγίσεων, η οποία συμπεριλαμβάνει τη μορφή και τον τρόπο των υποδειγμάτων που αναπτύσσονται. Η ομαδοποίηση αυτή αποτελείτε από τέσσερις κατηγορίες:

- Πολυκριτήριος μαθηματικός προγραμματισμός (multiobjective mathematical programming),
- Πολυκριτήρια θεωρία χρησιμότητας (multiattribute utility theory),
- Θεωρία των σχέσεων υπεροχής (outranking relations),
- Αναλυτική-συνθετική προσέγγιση (preference disaggregation approach).

Μεταξύ αυτών, ο πολυκριτήριος μαθηματικός προγραμματισμός χρησιμοποιείται κυρίως σε συνεχή προβλήματα και αποτελεί ένα υπερσύνολο του μαθηματικού προγραμματισμού σε περιπτώσεις όπου χρειάζεται να αναπτυχθούν πολλές αντικειμενικές συναρτήσεις, παρ' όλα αυτά μπορεί να συμβάλλει στην αντιμετώπιση διακριτών προβλημάτων λήψης αποφάσεων. Από την άλλη, οι τρεις τελευταίοι μέθοδοι χρησιμοποιούνται κυρίως σε περιπτώσεις διακριτών προβλημάτων λήψης αποφάσεων. Στόχος τους, για την αξιολόγηση ενός πεπερασμένου συνόλου εναλλακτικών, αποτελεί ο σχηματισμός όλων των κριτηρίων σύμφωνα με τις προβληματικές. Επιπλέον, αξίζει να σημειωθεί ότι, χρησιμοποιούνται και ως εργαλεία για την αντιμετώπιση συνεχών προβλημάτων.

4.5 Πολυκριτήρια θεωρία χρησιμότητας

Από τα αρχικά στάδια εξέλιξης της πολυκριτήριας ανάλυσης, η πολυκριτήρια θεωρία χρησιμότητας αποτέλεσε και συνεχίζει να αποτελεί έναν από τους κύριους θεμελιωτές της. Σκοπό της είναι, με χρήση μίας συνάρτησης αξιών/χρησιμότητας U, να μοντελοποιήσει και αναπαραστήσει το σύστημα αποφάσεων που ακολουθεί ο αποφασίζοντας (συνειδητά ή ασυνείδητα). Αυτή η συνάρτηση εξαρτάται από το σύνολο των κριτηρίων αξιολόγησης τα οποία καθορίζουν το αποτέλεσμα της αξιολόγησης: $U(G) = U(g_1, g_2, \ldots, g_n)$. Οι συναρτήσεις χρησιμότητας, είναι μη γραμμικές,

αύξουσες, ορίζονται στο πεδίο τιμών των κριτηρίων αξιολόγησης και ακολουθούν τις δύο παρακάτω βασικές ιδιότητες:

$$U(G_x) > U(G_y) \Leftrightarrow x \succ y \text{ (Η x είναι προτιμότερη της y)}$$

$$U(G_x) = U(G_y) \Leftrightarrow x \sim y \text{ (Η x είναι αδιάφορη της y)}$$

Η συνηθισμένη μορφή που χρησιμοποιείτε ευρέως, τόσο σε πρακτικό όσο και σε ερευνητικό επίπεδο είναι η ακόλουθη:

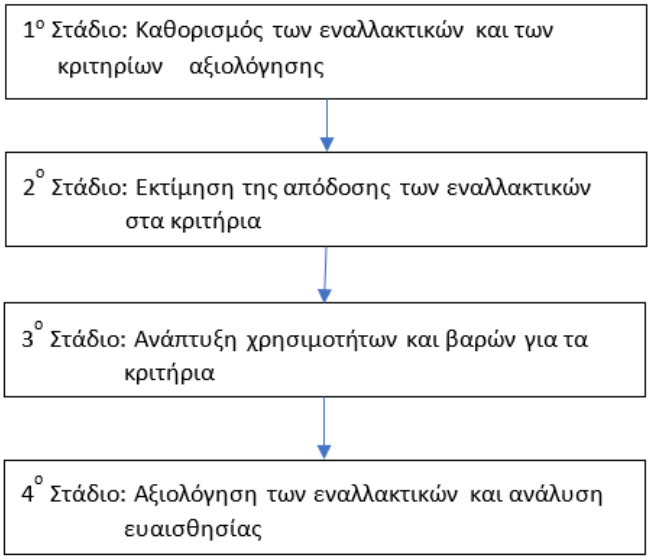
$$U(G) = \sum_{i=1}^{n} w_i u_i(g_i)$$

όπου $u_1, u_2, \dots, u_n$ είναι οι συναρτήσεις μερικών χρησιμοτήτων των κριτηρίων αξιολόγησης και $w_1, w_2, \dots, w_n \geq 0$ είναι συντελεστές στάθμισης (βάρη) των κριτηρίων. Κάθε συνάρτηση μερικής χρησιμότητας $u_i(g_i)$ καθορίζει τη χρησιμότητα των εναλλακτικών δραστηριοτήτων βάσει των επιδόσεών τους στο κριτήριο g_i, ενώ κάθε συντελεστής στάθμισης w_i υποδεικνύει την παραχώρηση (tradeoff) που είναι διατεθειμένος να κάνει ο αποφασίζοντας σε ένα κριτήριο αναφορά ώστε να επιτευχθεί αύξηση μιας μονάδας στο κριτήριο g_i. Η πλέον πιο διαδεδομένη μορφή συνολικής συνάρτησης χρησιμότητας είναι η προσθετική, οποία εκφράζεται από την σχέση

$$U(g) = w_1 u_1(g_1) + w_2 u_2(g_2) + \cdots + w_n u_n(g_n)$$

Η χρησιμοποίηση της παραπάνω συνάρτησης χρησιμότητας αναφέρεται στην προτιμησιακή ανεξαρτησία των κριτηρίων αξιολόγησης. Προτιμησιακά ανεξάρτητο (preferential independent) των υπολοίπων κριτηρίων θεωρείται ένα υποσύνολο του συνόλου των κριτηρίων αξιολόγησης $(G' \subset G)$ εάν και μόνο εάν οι προτιμήσεις σχετικά με τις εξεταζόμενες εναλλακτικές δραστηριότητες αυτού που λαμβάνει την απόφαση δεν επηρεάζονται από τα υπόλοιπα κριτήρια. Την υπόθεση της αμοιβαίας προτιμησιακής ανεξαρτησίας πληρεί το σύνολο G των κριτηρίων αξιολόγησης αν και μόνο αν κάθε υποσύνολο $(G' \subset G)$ είναι προτιμησιακά ανεξάρτητο των υπολοίπων κριτηρίων (Fishburn, 1970, Keeney and Raiffa, 1993). Στη συνεργασία μεταξύ του ειδικού αναλυτή και αυτού που λαμβάνει τις αποφάσεις στηρίζεται η διαδικασία ανάπτυξης μίας συνάρτησης ώστε να καθοριστεί με σαφήνεια το επίπεδο σημαντικότητας των κριτηρίων αξιολόγησης καθώς επίσης, ο τύπος των συναρτήσεων. Οι συντελεστές βαρύτητας χρησιμοποιούνται ώστε αυτός που αποφασίζει να κάνει παραχωρήσεις σε ένα κριτήριο αξιολόγησης με σκοπό την βελτίωση κάποιου άλλου. Για τον καθορισμό των συναρτήσεων μερικών χρησιμοτήτων οι τεχνικές που χρησιμοποιούνται, έχοντας ως βάση απλά ερωτήματα, προσπαθούν να λάβουν από τον αποφασίζοντα τις αναγκαίες πληροφορίες με σκοπό να οριστεί ο τρόπος που αξιολογεί τις εναλλακτικές στρατηγικές

σε κάθε ένα από τα κριτήρια. Η πιο διαδεδομένη τεχνική είναι αυτή της μέσης αξίας (Keeney & Raiffa, 1993) ενώ έχουν δημιουργηθεί και ποικίλα συστήματα υποστήριξης αποφάσεων που επιτρέπουν την αλληλεπιδραστική ανάπτυξη και χρησιμοποίηση συναρτήσεων χρησιμότητας. Ο αποφασίζοντας συγκρίνοντας την συνολική χρησιμότητα των εναλλακτικών αποφάεων, όπως αυτή υπολογίζεται μέσω της συνάρτησης χρησιμότητας, έχει τη δυνατότητα ταξινόμησης των εναλλακτικών δραστηριοτήτων από τις καλύτερες (δηλαδή αυτές με την υψηλότερη ολική χρησιμότητα) προς τις χειρότερες (δηλαδή αυτές με τη χαμηλότερη ολική χρησιμότητα), να τις ταξινομλησει σε κατηγορίες ή να επιλέξει κάποια ή κάποιες από αυτές.
Παρακάτω δίνεται μία σχεδιαγραμματική αναπαράσταση των 4 σταδίων την πολυκριτήριας Θεωρίας Χρησιμότητας (Σχήμα 4.2).



Σχήμα 4.2

4.6 Αναλυτική Ιεραρχική Διαδικασία

Analytic Hierarchy Process (AHP)

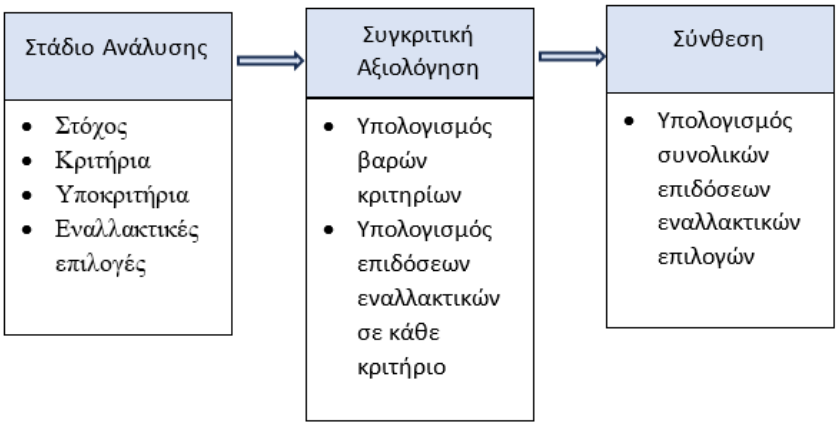
Η Αναλυτική Ιεραρχική Διαδικασία (Analytic Hierarchy Process – AHP) είναι μία από τις πιο ευρέως χρησιμοποιούμενες μεθόδους πολυκριτήριας ανάλυσης αποφάσεων και πιο ειδικά ανάγεται στις μεθόδους πολυκριτήριας θεωρίας χρησιμότητας. Η μέθοδος AHP χρησιμοποιεί σαφή ή αντικειμενικά μαθηματικά για να παρουσιάσει τις υποκειμενικές προτιμήσεις ενός ατόμου ή μιας ομάδας έχοντας ως στόχο την λήψη ομαδικών αποφάσεων με βάση τους στόχους και την κατανόηση τους για το πρόβλημα.

Η AHP μπορεί να συνοψιστεί σε τρία βασικά στάδια (Σχήμα 4.3):

- **Στάδιο Ανάλυσης:** Κατά το πρώτο στάδιο προσδιορίζεται το πρόβλημα και ο στόχος του, οι διαθέσιμες εναλλακτικές επιλογές του, δηλαδή οι επιλογές ανάμεσα στις οποίες καλούνται να επιλέξουν οι ειδικοί για την αντιμετώπιση του προβλήματος, καθώς επίσης και τα κριτήρια, δηλαδή τα χαρακτηριστικά των εναλλακτικών, βάσει των οποίων οι αποφασίζοντες θα κληθούν να επιλέξουν τη βέλτιστη λύση.
- **Συγκριτική Αξιολόγηση:** Κατά το δεύτερο στάδιο, περιλαμβάνεται ο καθορισμός των βαρών των κριτηρίων και ο υπολογισμός των επιδόσεων των εναλλακτικών σε κάθε κριτήριο. Έπειτα, την αξιολόγηση των εναλλακτικών για κάθε κριτήριο, προσδιορίζεται η σημαντικότητα κάθε κριτηρίου μέσω της απόδοσης βαρών σε αυτά. Μετά την απόδοση, υπολογίζονται τα κανονικοποιημένα βάρη:

$$w_i = \frac{w_i'}{\sum_{i=1}^{n} w_i'}$$

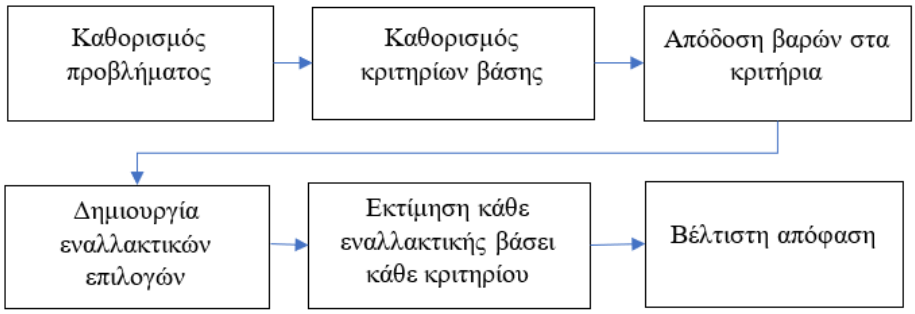
- **Σύνθεση:** Κατά το τελευταίο στάδιο, υπολογίζονται οι συνολικές επιδόσεις των εναλλακτικών επιλογών, με στόχο την βέλτιστη απόφαση.



Σχήμα 4.3

Πιο αναλυτικά (Σχήμα 4.4), οι ειδικοί λήψης αποφάσεων αρχικά αποσυνθέτουν το πρόβλημα της απόφασής τους σε μια ιεραρχία κατανοητών και ανεξάρτητων υποπροβλημάτων, τα οποία μπορεί να σχετίζονται με οποιαδήποτε πτυχή του προβλήματος. Μόλις χτιστεί η ιεραρχία, οι υπεύθυνοι λήψης αποφάσεων αξιολογούν τα διάφορα στοιχεία συγκρίνοντάς τα ανά ζεύγη, με τη χρήση αριθμών εντός της κλίμακας 1-9. Έπειτα, ένα αριθμητικό βάρος προκύπτει για κάθε στοιχείο της ιεραρχίας, επιτρέποντας τη σύγκριση διαφορετικών και ασύγκριτων στοιχείων μεταξύ τους με λογικό και συνεπή τρόπο. Αυτή η ικανότητα διακρίνει την AHP από άλλες τεχνικές λήψης αποφάσεων. Στο τελικό βήμα της διαδικασίας, υπολογίζονται αριθμητικές προτεραιότητες για κάθε μία από τις εναλλακτικές αποφάσεις. Αυτοί οι αριθμοί

αντιπροσωπεύουν τη σχετική ικανότητα των εναλλακτικών λύσεων να επιτύχουν τον στόχο της απόφασης, επομένως επιτρέπουν μια απλή εξέταση των διαφόρων τρόπων δράσης.



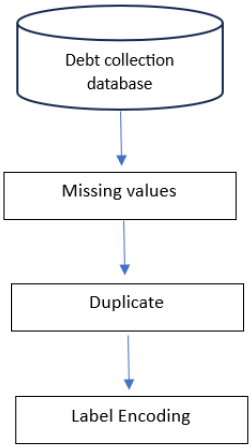
Σχήμα 4.4

Κεφάλαιο 5ο:
Προτεινόμενη Μεθοδολογία

Στο παρόν κεφάλαιο παρουσιάζεται η μεθοδολογία η οποία προτείνεται στην παρούσα διπλωματική και η οποία εφαρμόστηκε και αναλύθηκε σε δεδομένα του πραγματικού κόσμου, στο επόμενο κεφάλαιο. Η εφαρμογή της παρούσας μεθοδολογίας χωρίζεται σε 5 μέρη και εφαρμόζεται σε δεδομένα πελατών που οφείλουν ληξιπρόθεσμους λογαριασμούς σε εταιρείες. Μέσω αυτής και των δεδομένων του κάθε πελάτη, θα προκύπτει η πιθανότητα θετικής υπόσχεσης πληρωμής (P2P), η πιθανότητα αθέτησης μίας πληρωμής για πάνω από 30 ημέρες (DPD), το σκορ ενός πελάτη για την ταξινόμηση του σε μία από τις 3 κατηγορίες (bad – good – very good) και τέλος η πρόταση και βαθμολόγηση διάφορων στρατηγικών με σκοπό την επιλογή της καλύτερης στρατηγικές επικοινωνίας για την είσπραξη ληξιπρόθεσμων λογαριασμών ανάλογα με τον εκάστοτε πελάτη.

5.1 Προεργασία και Προετοιμασία Δεδομένων

Το 1ο μέρος της μεθοδολογίας αφορά ένα βασικό κομμάτι της ανάλυσης δεδομένων το οποίο είναι η προεργασία και η προετοιμασία των μεταβλητών των δεδομένων (Σχήμα 5.1). Κατά την εισαγωγή του Dataset είναι βασικό να ελεγχθούν κάποια βασικά χαρακτηριστικά των δεδομένων για να μην προκύψουν προβλήματα κατά την υπόλοιπη διαδικασία.



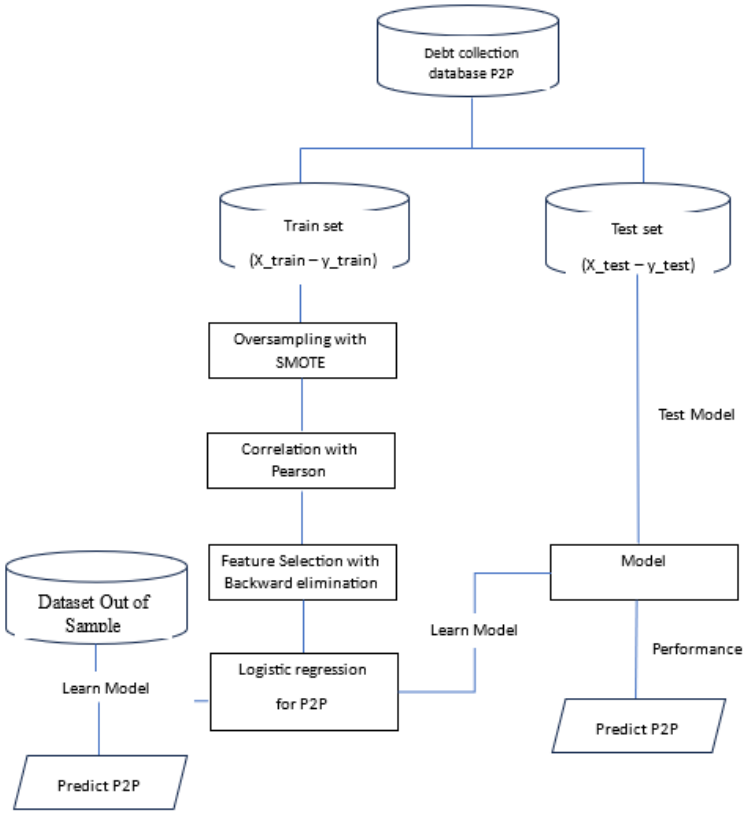
Σχήμα 5.1

Αρχικά, πρέπει τα δεδομένα να ελεγχθούν ως προς το αν υπάρχουν διπλές γραμμές ή στήλες, για να διασφαλιστεί η ακεραιότητα των δεδομένων και για την αντιμετώπιση αυτού πρέπει να εξαλειφθούν από τα σύνολα δεδομένων. Έπειτα, είναι σημαντικό να γίνει έλεγχος ελλειπουσών τιμών (missing values), διότι αν υπάρχουν χαμένα δεδομένα μπορεί να προκύψει μεροληψία στις εκτιμήσεις που προέρχονται από ένα στατιστικό μοντέλο και καθιστούν τις κοινές στατιστικές μεθόδους ακατάλληλες ή δύσκολα εφαρμόσιμές. Επιπλέον, σημαντικό είναι οι ποιοτικές μεταβλητές να μετατραπούν σε ποσοτικές, με σκοπό την ευκολότερη επεξεργασία τους. Στην παρούσα διπλωματική προτείνεται η τεχνική Label Encoding (παράγραφος 3.4.1) η οποία χρησιμοποιείται για τη μετατροπή κατηγορικών στηλών σε αριθμητικές, ώστε να μπορούν να τοποθετηθούν από μοντέλα μηχανικής εκμάθησης που λαμβάνουν μόνο αριθμητικά δεδομένα. Είναι ένα σημαντικό βήμα προεπεξεργασίας σε ένα έργο μηχανικής μάθησης. Πιο συγκεκριμένα, το Label Encoding αναθέτει αύξουσες αριθμητικές τιμές σε κάθε κατηγορία για τον χειρισμό κατηγορικών μεταβλητών. Αυτό το σύστημα κωδικοποίησης διευκόλυνε την ηλεκτρονική παρουσίαση των κατηγορικών μεταβλητών σε αριθμητική μορφή εντός του συνόλου δεδομένων.

5.2 Λογιστική Παλινδρόμηση για Πρόβλεψη Υπόσχεσης Πληρωμής (P2P)

Το 2ο μέρος της μεθοδολογίας περιλαμβάνει την προετοιμασία αλλά και την εφαρμογή την Λογιστικής Παλινδρόμησης (ΛΠ) στα δεδομένα με σκοπό την πρόβλεψη της θετικής υπόσχεσης πληρωμής ενός πελάτη (Promise to Pay – P2P) Ως θετική υπόσχεση πληρωμής ορίζεται η πιθανότητα ο πελάτης να υποσχεθεί ότι θα πληρώσει το ληξιπρόθεσμο ποσό οφειλής έπειτα από επικοινωνία του με τον εισπράκτορα οφειλών. Η διαγραμματική αναπαράσταση της μεθοδολογίας παρουσιάζεται στο Σχήμα 5.2.

Αρχικά χρησιμοποιούμε ένα σύνολο δεδομένων (Development Dataset) το οποίο θα περιέχει στοιχεία από τις πληρωμές των πελατών και την μεταβλητή-στόχο P2P. Διαχωρίζουμε το Dataset σε σύνολο εκπαίδευσης (train_set) και σύνολο δοκιμής (test_set) με αναλογία 80% και 20% αντίστοιχα. Σκοπός αυτού του διαχωρισμού είναι να εφαρμοστεί η λογιστική παλινδρόμηση για την πρόβλεψη της θετικής υπόσχεσης πληρωμής (P2P) σε νέα δεδομένα που θα προκύψουν στο μέλλον.



Σχήμα 5.2

Επίσης, σε προβλήματα ταξινόμησης, είναι σημαντικό να διασφαλιστεί ότι οι κατηγορίες της μεταβλητής-στόχου είναι ισορροπημένες. Αυτή η ισορροπία είναι ζωτικής σημασίας για τη δημιουργία ισχυρών εκτιμήσεων και αξιόπιστων αξιολογήσεων της απόδοσης του μοντέλου. Στην περίπτωση που δεν υπάρχει ισορροπία, προτείνεται η τεχνική Synthetic Minority Over-sampling Technique (SMOTE) για την εξισορρόπηση των δεδομένων εκπαίδευσης του μοντέλου (παράγραφος 3.4.2).

Επιπλέον, είναι πολύ σημαντικό το σύνολο εκπαίδευσης να μην περιέχει υψηλά συσχετιζόμενες μεταβλητές. Για τον λόγο αυτό, στην συνέχεια προτείνεται η μέθοδος Pearson Correlation και η αφαίρεση όλων των στηλών -πλην μίας- που παρουσιάζουν υψηλή συσχέτιση (παράγραφος 3.4.3). Στην παρούσα εργασία το όριο για τις υψηλά συσχετιζόμενες μεταβλητές ορίστηκε αν η απόλυτη τιμή του είναι πάνω από το 0,7 αλλά κάθε φορά μπορεί να καθορίζεται από τις προτιμήσεις του κάθε ειδικού. Το Pearson Correlation ένα μέτρο γραμμικής συσχέτισης μεταξύ δύο μεταβλητών. Όσο πιο κοντά βρίσκεται στο 0, τόσο πιο μικρή συσχέτιση υπάρχει μεταξύ των δύο μεταβλητών.

Για να δημιουργηθεί ένα μοντέλο Λογιστικής Παλινδρόμησης (ΛΠ), πρέπει πρώτα να επιλεχθούν τα χαρακτηριστικά που θα χρησιμοποιηθούν για την πρόβλεψη της μεταβλητής στόχου. Η επιλογή των χαρακτηριστικών είναι ένα σημαντικό βήμα στη διαδικασία μοντελοποίησης, καθώς μπορεί να επηρεάσει σημαντικά την απόδοση του

μοντέλου. Στην παρούσα εργασία χρησιμοποιήθηκε η τεχνική Backward elimination, η οποία συνέβαλλε στο να βρεθούν οι 10 σημαντικότερες μεταβλητές για την εφαρμογή της Λογιστική Παλινδρόμησης (παράγραφος 3.4.4). Η τεχνική αυτή είναι χρήσιμη γιατί μπορεί να βοηθήσει στη μείωση των πιθανοτήτων υπερπροσαρμογής των δεδομένων και να κάνει το μοντέλο γραμμικής παλινδρόμησης πιο ερμηνεύσιμο. Η τεχνική Backward elimination χρησιμοποιείται στη μηχανική εκμάθηση για την εύρεση του καλύτερου υποσυνόλου χαρακτηριστικών από ένα δεδομένο σύνολο χαρακτηριστικών. Ξεκινά με την προσαρμογή ενός μοντέλου παλινδρόμησης με όλες τις ανεξάρτητες μεταβλητές και αφαιρεί επαναληπτικά χαρακτηριστικά που δεν είναι προγνωστικά για τη μεταβλητή στόχο ή έχουν τη μικρότερη προγνωστική ισχύ.

Έπειτα από όλα τα παραπάνω, προτείνεται η εφαρμογή της Λογιστικής Παλινδρόμησης (ΛΠ), όπως περιεγράφηκε στην παράγραφο 3.2.4, με στόχο την πρόβλεψη του P2P (1 αν λάβουμε θετική απάντηση , 0 αν λάβουμε αρνητική). Το μοντέλο αυτό εφαρμόζεται έπειτα στο σύνολο δοκιμής ώστε να υπολογίζουμε την ακρίβεια του στο ποσοστό των προβλέψεων. Τέλος, το μοντέλο αυτό χρησιμοποιείται και εφαρμόζεται σε ένα μελλοντικό σύνολο δεδομένων (Dataset out of sample), το οποίο με βάση τα χαρακτηρίστηκα των πελατών υπολογίζει την πιθανότητα P2P, πριν ακόμη ο συλλέκτης πληρωμών επικοινωνήσει με τον πελάτη.

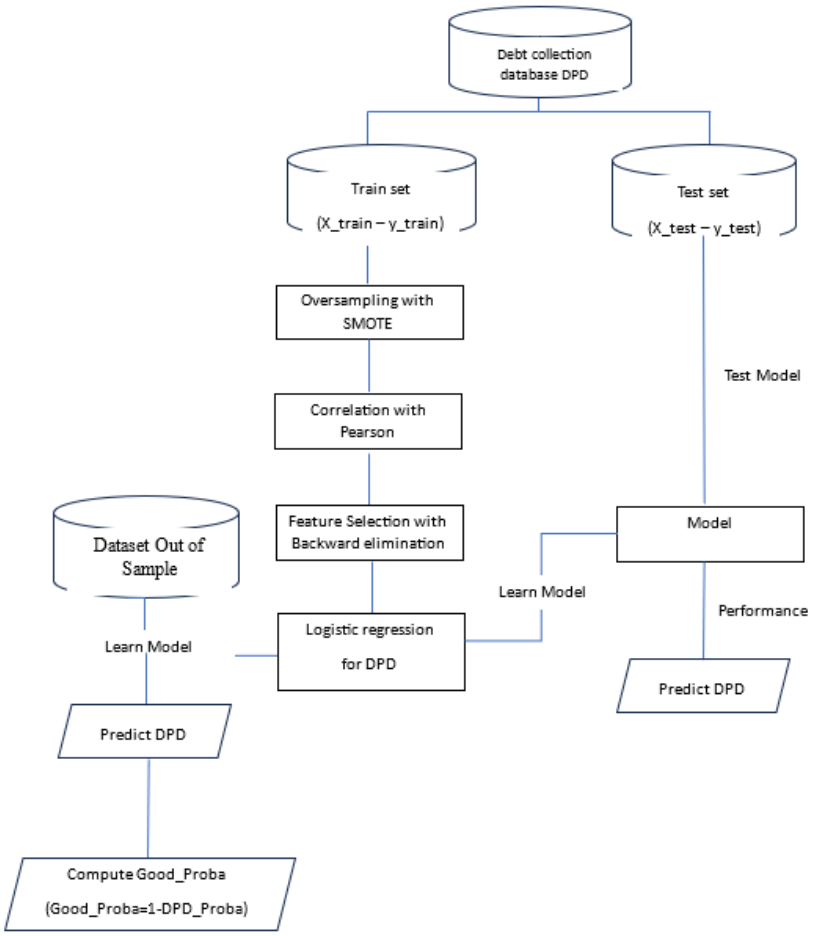
5.3 Λογιστική Παλινδρόμηση για Πρόβλεψη Αθέτησης Πληρωμής (DPD)

Το 3^{ο} μέρος της μεθοδολογίας, είναι παρόμοιο με το 2^{ο} μέρος, με την διαφορά ότι περιλαμβάνει την προετοιμασία αλλά και την εφαρμογή την Λογιστικής Παλινδρόμησης (ΛΠ) στα δεδομένα με σκοπό την πρόβλεψη της αθέτησης πληρωμής ενός πελάτη για διάστημα μεγαλύτερο των 30 ημερών (DPD). Η διαγραμματική αναπαράσταση της μεθοδολογίας παρουσιάζεται στο Σχήμα 5.3.

Στο προτεινόμενό μοντέλο εξίσου χρησιμοποιείται συνόλου δεδομένων (Development Dataset) το οποίο θα περιέχει στοιχεία από τις πληρωμές των πελατών και την μεταβλητή-στόχο DPD. Το σύνολο αυτό διαχωρίζεται όπως και το προηγούμενο σε σύνολο εκπαίδευσης (train_set) και σύνολο δοκιμής (test_set) με αναλογία 80% και 20% αντίστοιχα. Επίσης, στην περίπτωση που δεν υπάρχει ισορροπία, προτείνεται η τεχνική Synthetic Minority Over-sampling Technique (SMOTE) για την εξισορρόπηση των δεδομένων εκπαίδευσης του μοντέλου. Επιπλέον, στην συνέχεια προτείνεται η μέθοδος Pearson Correlation και η αφαίρεση όλων των στηλών -πλην μίας- που παρουσιάζουν υψηλή συσχέτιση πάλι αν η απόλυτη τιμή του ορίου είναι πάνω από το 0,7. Στην συνέχεια, για την επιλογή των χαρακτηριστικών για την πρόβλεψη της μεταβλητής στόχου εφαρμόζουμε με την τεχνική Backward elimination, Έπειτα από όλη την παραπάνω μεθοδολογία, προτείνεται η εφαρμογή της Λογιστικής Παλινδρόμησης (ΛΠ), με στόχο την πρόβλεψη του DPD (1 αν υπάρχει αθέτηση πληρωμής, 0 δεν υπάρχει

αθέτηση πληρωμής). Το μοντέλο αυτό εφαρμόζεται έπειτα στο σύνολο δοκιμής ώστε να υπολογίζουμε την ακρίβεια του στο ποσοστό των προβλέψεων. Τέλος, το μοντέλο αυτό χρησιμοποιείται και εφαρμόζεται σε ένα μελλοντικό σύνολο δεδομένων (Dataset out of sample), το οποίο με βάση τα χαρακτηρίστηκα των πελατών υπολογίζει την πιθανότητα αθέτησης πληρωμών. Σε συνέχεια του μέρους αυτού, μέσω των τιμών DPD που υπολογίστηκαν, δημιουργείται μία νέα μεταβλητή Good_Proba, ώστε οι μικρές τιμές να δηλώνουν την μεγάλη πιθανότητα αθέτησης πληρωμής ενώ οι μεγάλες τιμές να δηλώνουν μικρή πιθανότητα αθέτησης (δηλαδή όσο πιο κοντά στο 1 είναι η πιθανότητα τόσο πιο καλός πελάτης χαρακτηρίζεται). Συνεπώς,

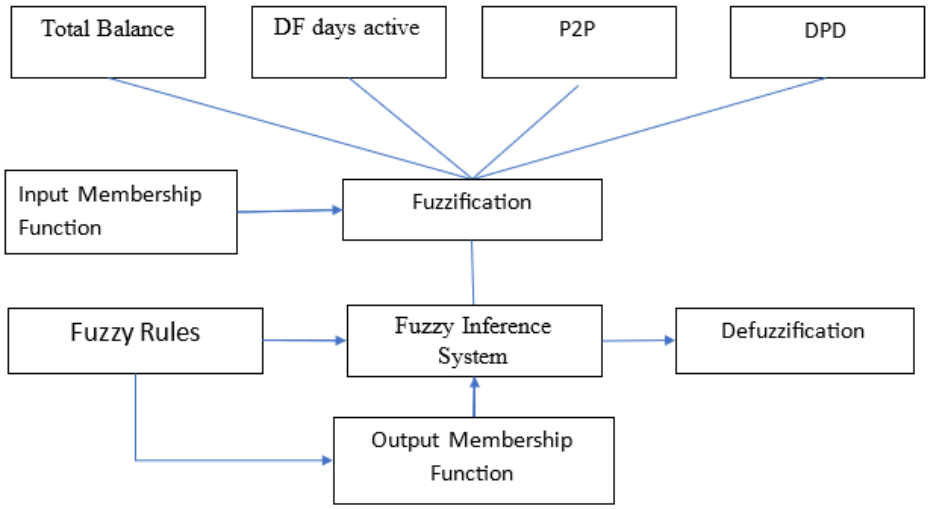
$$Good\_Proba = 1 - DPD\_Proba$$



Σχήμα 5.3

5.4 Ασαφής Λογική

Σε συνέχεια της μεθοδολογίας, θα εφαρμοστεί η μέθοδος Mamdani της Fuzzy Logic, όπως περιεγράφηκε στην παράγραφο 3.3 της παρούσας εργασίας, με σκοπό να προκύψει ένα σκορ για τον κάθε πελάτη, στην κλίμακα του 0-100. Η διαγραμματική αναπαράσταση της όλης διαδικασίας αναπαρίσταται στο Σχήμα 5.4.

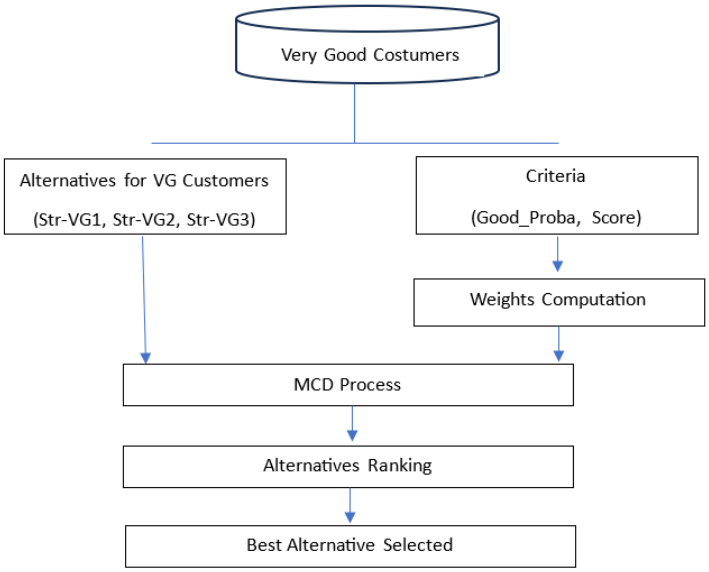


Σχήμα 5.4

Στο προαναφερθέν σύνολο δεδομένων Dataset Out of Sample για την ασαφοποίηση και για τους κανόνες ασαφούς συμπερασμού προτείνεται να χρησιμοποιηθούν 4 βασικές μεταβλητές εισόδου. Αυτές είναι: το συνολικό ανεξόφλητο ποσό (Total Balance – TB), το σύνολο των πληρωμών από την τελευταία πληρωμή (DF days active – DA) και οι δύο νέες μεταβλητές , P2P και Good_Proba, που προέκυψαν από την εφαρμογή των μοντέλων ΛΠ. Στα πλαίσια επέκτασης της παρούσας διπλωματικής, είναι στην ευχέρεια του κάθε ειδικού να χρησιμοποιήσει όποιες μεταβλητές θεωρεί πιο σημαντικές και να ορίσει τους ανάλογου κανόνες ασάφειας. Αρχικά τα δεδομένα τροποποιούνται σε λεκτικές μεταβλητές (ασαφοποίηση – Fuzzification), έτσι τα εισαγόμενα στοιχεία μετασχηματίζονται σε ένα βαθμό συμμετοχής στο 0-1. Το δεύτερο βήμα είναι η ασαφής συμπερασματολογία (Fuzzy Inference System). Τα εξαγόμενα στοιχεία από την διαδικασία αποτελούν ένα ασαφές σύνολο. Το τρίτο βήμα είναι η αποκωδικοποίηση από έναν crisp αριθμό ,μέσω της Αποασαφοποίησης (Defuzzification). Συνεπώς προκύπτει ένα Score για τον κάθε πελάτη ανάλογα με τις τιμές εισόδου στην κλίμακα 0-100 και μέσω αυτού άλλη μία στήλη με τους γλωσσικούς όρους που χαρακτήριζαν τον κάθε πελάτη (Customer) ως Bad (B), Good (G) ή Very Good(VG).

5.5 Πολυκριτήρια Ανάλυση Αποφάσεων

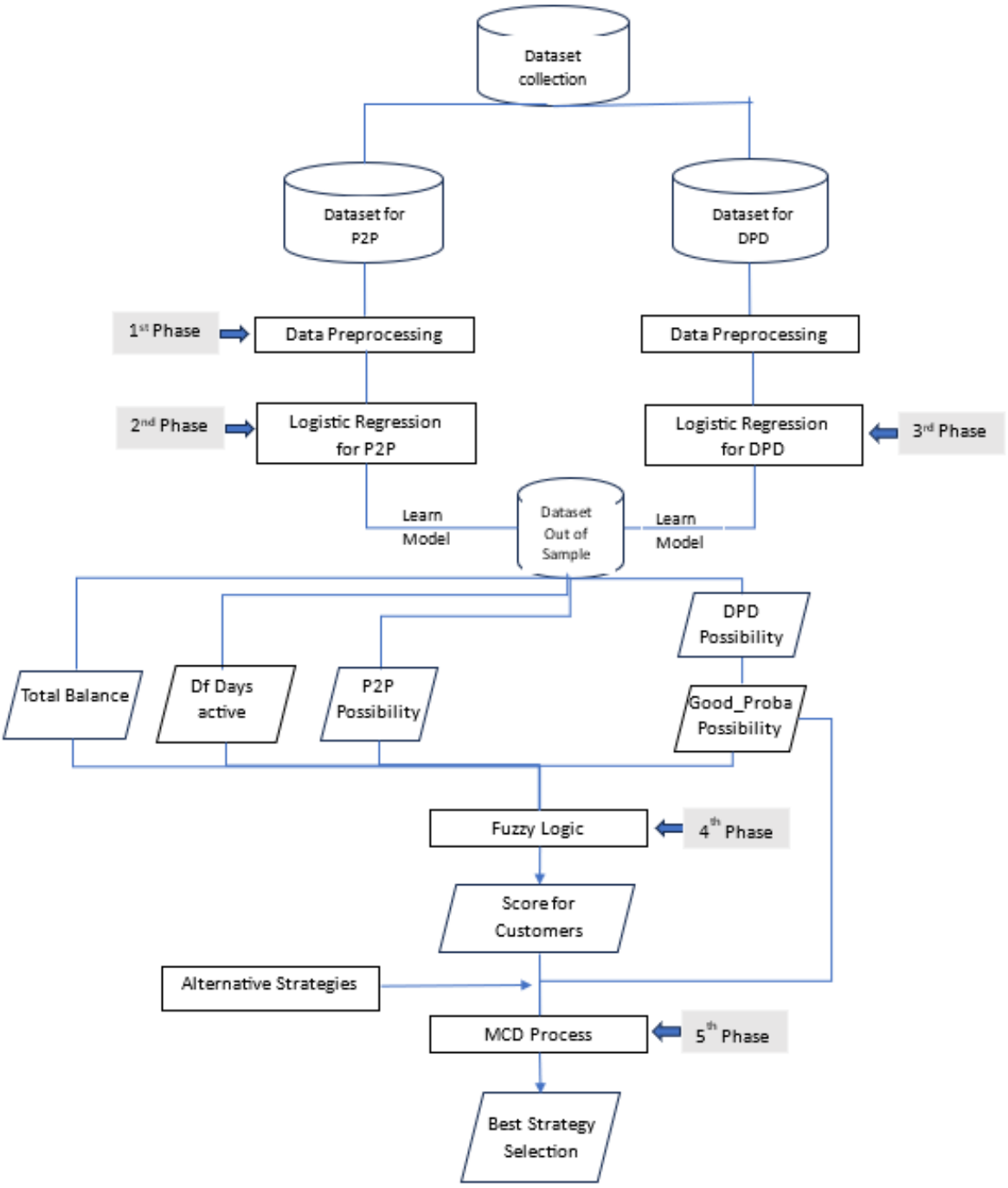
Κατά το τελικό στάδιο της εργασίας, σκοπός είναι να βρεθεί η πιο κατάλληλη στρατηγική προσέγγισης του κάθε πελάτη για την είσπραξη ληξιπρόθεσμων λογαριασμών με την μέθοδο AHP. Κατά το στάδιο αυτό, προτείνεται να οριστούν από 3 στρατηγικές (Str-1 - Str-2 - Str-3) για κάθε κατηγορία πελάτη (B – G – VG). Ως στρατηγική εννοείται μία σειρά καθορισμένων ενεργειών επικοινωνίας με τον πελάτη μέσω email, sms ή τηλεφωνικής κλήσης, με στόχο της είσπραξη οφειλών. . Κάθε μία από τις στρατηγικές βαθμολογείτε ως προς δύο κάποια κριτήρια το Good_Proba που προέκυψε παραπάνω μέσω της ΛΠ και το Score που προέκυψε παραπάνω από την μέθοδο Fuzzy. Οι λήπτες αποφάσεων εισάγουν τις προτιμήσεις τους όσον αφορά την αξιολόγηση των στρατηγικών σε έναν τετραγωνικό πίνακα με κλίμακα από το 1 έως το 9. Οι στρατηγικές και οι ανάλογες βαθμολογήσεις κάθε φορά καθορίζονται από τον εκάστοτε ειδικό ο οποίος τις καθορίζει ανάλογα την εμπειρία και τις γνώσεις του. Τα βάρη που προκύπτουν χρησιμοποιούνται για την εύρεση καλύτερης στρατηγικής για κάθε πελάτη ανάλογα με τις τιμές κριτηρίων. Στο Σχήμα 5.5 δίνεται η διαγραμματική αναπαράσταση που ακολουθείτε στην περίπτωση των πολύ καλών πελατών (VG). Ανάλογα ισχύει και για τις άλλες δύο κατηγορίες πελατών (Bad - B & Good - G).



Σχήμα 5.5

Συμπερασματικά, η προτεινόμενη μεθοδολογία αποτελείται από 5 στάδια αλληλένδετα μεταξύ τους. Η διαγραμματική της αναπαράσταση αποτυπώνεται στο Σχήμα 5.6. Κατά το 1ο στάδιο περιλαμβάνεται h προετοιμασία των δεδομένων, στο 2ο στάδιο μέσω της ΛΠ η οποία εφαρμόζεται σε στοιχεία πελατών προκύπτει η πιθανότητα να λάβουμε θετική απάντηση αποπληρωμής ληξιπρόθεσμου λογαριασμού, στο 3ο στάδιο επίσης εφαρμόζεται η ΛΠ μέσω της οποίας προκύπτει η πιθανότητα αθέτησης μία

πληρωμής. Στο 4ο στάδιο, χρησιμοποιούμε τις μεταβλητές P2P και Good_Proba των προηγούμενων σταδίων με στόχο το σκοράρισμα των πελατών και της ταξινόμησης τους με χρήση ασαφούς λογικής. Τέλος, κατά το 5ο στάδιο ορίζονται η κατάλληλες στρατηγικές και τα κριτήρια τα οποία είναι δύο από τις μεταβλητές των προηγούμενων σταδίων με στόχο την εφαρμογή πολυκριτήριας ανάλυσης αποφάσεων για την εύρεση της καλύτερης στρατηγικής επικοινωνίας με τον πελάτη με βάση τα διαφορετικά αποτελέσματα του καθενός. Συνεπώς, στην παρούσα διπλωματική προτείνεται μία σειρά μεθοδολογιών οι οποίες είναι ευρέως γνωστές στην επιστημονική κοινότητα όπου τα αποτελέσματα της κάθε μίας χρησιμοποιούνται για την εφαρμογή της επόμενης, καθιστώντας τες αλληλένδετες μεταξύ τους.


Dataset collection
Dataset for P2P
Dataset for DPD
1st Phase
Data Preprocessing
Data Preprocessing
2nd Phase
Logistic Regression for P2P
Logistic Regression for DPD
3rd Phase
Learn Model
Dataset Out of Sample
Learn Model
DPD Possibility
Total Balance
Df Days active
P2P Possibility
Good_Proba Possibility
Fuzzy Logic
4th Phase
Score for Customers
Alternative Strategies
MCD Process
5th Phase
Best Strategy Selection

Σχήμα 5.6

Κεφάλαιο 6ο:
Μελέτη Περίπτωσης και Υπολογιστικά Αποτελέσματα

6.1 Περιβάλλον Ανάπτυξης και Σετ Δεδομένων

Στο παρόν κεφάλαιο γίνεται η εφαρμογή της προαναφερθείσας μεθοδολογίας σε δύο διαφορετικά πλαίσια δεδομένων, το 1ο αφορά εταιρείες και το 2ο φυσικά πρόσωπα, δεδομένα του πραγματικού κόσμου, με χρήση, κυρίως της γλώσσας προγραμματισμού Python. Πιο συγκεκριμένα, προτείνεται μία σειρά ενεργειών πάνω σε δεδομένα με σκοπό την εύρεση της καλύτερης στρατηγικής επικοινωνίας με πελάτες που έχουν ληξιπρόθεσμους λογαριασμούς.

Τα δεδομένα τις παρούσας εργασίας αποτελούν πραγματικά δεδομένα μίας εταιρείας η οποία είναι πάροχος ηλεκτρικής ενέργειας στην Ελλάδα και περιλαμβάνουν στοιχεία πελατών, οι οποίοι οφείλουν ληξιπρόθεσμου λογαριασμούς σε αυτήν. Αναλύθηκαν και μελετήθηκαν 4 διαφορετικά σύνολα δεδομένων (Datasets), εκ των οποίων τα πρώτα δύο αφορούν επιχειρήσεις (Business Dataset 1 – BD1 & Business Dataset 2 – BD2) και τα άλλα δύο αφορούν φυσικά πρόσωπα (Retail Dataset 1 – RD1 & Retail Dataset 2 – RD2). Πιο ειδικά, στα BD1 και RD1, η τελευταία στήλη (LABEL) αναφέρεται στη θετική υπόσχεση πληρωμής (Promise to Pay - P2P), δηλαδή στο αν τελικά μετά από μία επιτυχημένη επικοινωνία, οι πελάτες έδωσαν θετική υπόσχεση πληρωμής (1) ή όχι (0). Ενώ, στα BD2 και RD2, η τελευταία στήλη αναφέρεται στην αθέτηση της πληρωμής (1) ή όχι (0). Επιπλέον, οι βάσεις δεδομένων αρχικά είχαν 188 ακατέργαστα χαρακτηρίστηκα εκ των οποίων μόνο 6 από αυτά χρησιμοποιούνται πραγματικά. Τα 176 τα χαρακτηριστικά έχουν αριθμητική αξία, δηλαδή είναι ποσοτικά, ενώ τα υπόλοιπα 12 είναι κατηγορικά. Επιπλέον, αξίζει να σημειωθεί ότι τα παραπάνω δεδομένα για λόγους της διπλωματικής, θα διαχωριστούν σε Dataset_Out_of_Sample, αποτελούμενο περίπου από το 25% των δεδομένων και σε Development_Dataset με το 75% των δεδομένων. Για λόγους ερευνητικούς, θεωρούμε ότι τα δεδομένα από το Dataset_Out_of_Sample δεν έχουν τιμές ως προς τις μεταβλητές στόχου των ΛΠ και θα κρατήσουμε τις τιμές που θα προκύψουν μέσω των μοντέλων ΛΠ. Αυτό γίνεται, διότι θέλουμε να δούμε την εφαρμογή του μοντέλου σε μελλοντικώς ερχόμενα δεδομένα πελατών ώστε μέσω της μεθοδολογίας να γνωρίζουμε εξ' αρχής τον σωστό τρόπο προσέγγισης τους. Τέλος, αξίζει να επισημανθεί ότι με βάση το μεθοδολογικό πλαίσιο που περιεγράφηκε στον 5ο κεφάλαιο, κατά την εφαρμογή της μεθόδου Fuzzy και κατά

την Πολυκριτήρια ανάλυση αποφάσεων, κάθε φορά απαιτείται από τον εκάστοτε ειδικό να αποφασίζει και καθορίζει τις τιμές ανάλογα την εμπειρία, τις γνώσεις και τις προτιμήσεις του. Στα πλαίσια της παρούσας διπλωματικής οι προτιμήσεις διατίθενται ως ενδεικτικό παράδειγμα και κάθε φορά μπορούν να προσαρμοστούν κατάλληλα.

Να σημειωθεί ότι η προτεινόμενη μεθοδολογία αναπτύχθηκε στην γλώσσα προγραμματισμού Python με χρήση προσωπικού φορητού υπολογιστή HP με επεξεργαστή Intel Core i3 με μνήμη RAM 8 GB και με χρήση των Windows 10.

6.2 Εφαρμογή της Μεθόδου σε Επιχειρήσεις

6.2.1 Μοντέλο Λογιστικής Παλινδρόμησης για P2P

Business Dataset 1 (BD1)

Παρακάτω παρουσιάζεται η βασική μεθοδολογία και τα αποτελέσματα της που εφαρμόστηκε στο πρώτο dataset (BD1) που αφορά τις επιχειρήσεις.

Α΄ Μέρος: Παρουσίαση και Προεργασία των Δεδομένων

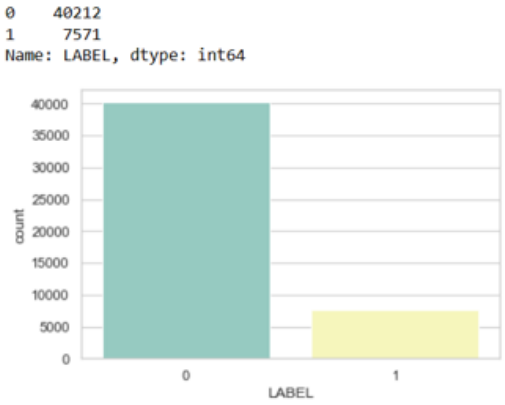
Το αρχικό σύνολο δεδομένων του Business Dataset 1 (BD1) περιέχει στοιχεία 79.729 επιχειρήσεων με 188 μεταβλητές για την κάθε μία από αυτές τα οποία δίνουν πληροφορίες για το χρωστούμενο ποσό, τον χρονικό διάστημα της οφειλής, το χρονικό διάστημα από την τελευταία επικοινωνία με τον πελάτη κ.ο.κ. Η τελευταία στήλη (LABEL), όπως αναφέρθηκε πιο πάνω, αφορά την θετική υπόσχεση πληρωμής (P2P).

Αρχικά, ελέγχθηκαν τα δεδομένα για τυχόν ελλείπουσες τιμές, και διαπιστώθηκε ότι όλες οι στήλες και οι γραμμές είναι πλήρως συμπληρωμένες. Έπειτα, ελέγχθηκαν για τυχόν διπλές γραμμές ή στήλες, ώστε να αφαιρεθούν τα ίδια δεδομένα, αν τυχόν υπάρχουν, αλλά επίσης διαπιστώθηκε ότι και οι 79.729 γραμμές αφορούν διαφορετικές επιχειρήσεις. Στην συνέχεια, βρέθηκαν οι ποιοτικές μεταβλητές (12 στο σύνολο) και μέσω του Label Encoding μετατράπηκαν σε ποσοτικές, με σκοπό την ευκολότερη επεξεργασία του ώστε όλες οι μεταβλητές να είναι αριθμητικές.

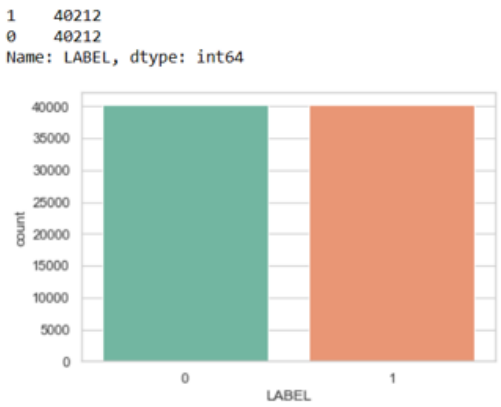
Έπειτα από την βασικό έλεγχο και προεπεξεργασία των δεδομένων, το BD1 χωρίστηκε σε Dataset_Out_of_Sample, αποτελούμενο από 20.000 στοιχεία (περίπου το 25% των δεδομένων) και σε Development_Dataset (75% των δεδομένων). Ο διαχωρισμός αυτός έγινε με σκοπό την επεξεργασία του Development_Dataset και την χρήση του κατά την λογιστική παλινδρόμηση και έπειτα, το μοντέλο αυτό να εφαρμοστει στο Dataset_Out_of_Sample με σκοπό να ελεγχθεί η λειτουργεία του σε ένα σύνολο εκτός του δείγματος δεδομένων. Έπειτα, το Development_Dataset διαχωρίστηκε σε

σύνολο εκπαίδευσης (train_set) και σύνολο δοκιμής (test_set) με αναλογία 80% και 20% αντίστοιχα. Σκοπός αυτού του διαχωρισμού είναι να εφαρμοστεί η λογιστική παλινδρόμηση για την πρόβλεψη της θετικής υπόσχεσης πληρωμής (P2P) σε νέα δεδομένα που θα προκύψουν στο μέλλον.

Σε προβλήματα ταξινόμησης, είναι σημαντικό να διασφαλιστεί ότι οι κατηγορίες της μεταβλητής-στόχου (LABEL) είναι ισορροπημένες. Παρατηρήθηκε λοιπόν ότι η μεταβλητής-στόχος έχει διαφορετικό αριθμό περιπτώσεων (Εικόνα 6.1), για τον λόγο αυτό εφαρμόζουμε την τεχνική Synthetic Minority Over-sampling Technique (SMOTE) για την εξισορρόπηση των δεδομένων (Εικόνα 6.2).



Εικόνα 6.1



Εικόνα 6.2

Ακόμη, είναι πολύ σημαντικό το σύνολο δεδομένων να μην περιέχει υψηλά συσχετιζόμενες μεταβλητές. Για τον λόγο αυτό, την συνέχεια εφαρμόστηκε Pearson Correlation, και αφαιρέθηκαν από τα δεδομένα όλες οι στήλες -πλην μίας- που είχαν συσχέτιση από 0,7 και πάνω ή από -0,7 και κάτω

Β΄ Μέρος: Λογιστική Παλινδρόμηση

Πριν την δημιουργία του μοντέλου Λογιστικής Παλινδρόμησης επιλέχθηκαν τα 10 σημαντικότερα χαρακτηριστικά που θα χρησιμοποιηθούν για την πρόβλεψη της μεταβλητής στόχου, με χρήση της τεχνικής Backward elimination (Πίνακας 6.1 - Παράρτημα). Στην συνέχεια, εφαρμόστηκε η Λογιστικής Παλινδρόμησης (ΛΠ), όπως περιεγράφηκε στην παράγραφο 3.2.4, με στόχο την πρόβλεψη του P2P (1 αν λάβουμε θετική απάντηση, 0 αν λάβουμε αρνητική). Τα αποτελέσματα της φαίνονται στους Πίνακες 6.2, 6.3 και 6.4.

Πιο συγκεκριμένα, ο Πίνακας 6.2 είναι ο συγκεντρωτικός πίνακας των προβλέψεων και των πραγματικών τιμών. Πιο ειδικά παρατηρούμε:

9.471 TN (True Negative) δηλαδή η P2P ήταν αρνητική και προβλεπόταν αρνητική

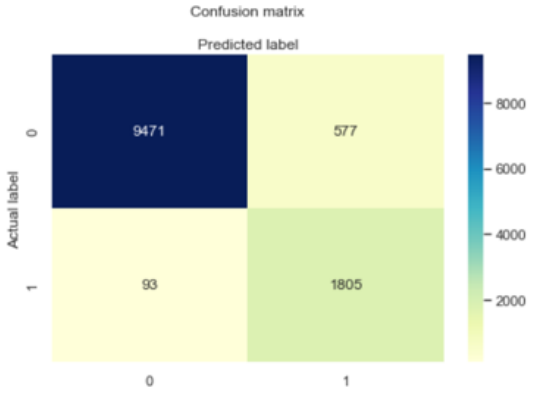
93 FN (False Negative) δηλαδή η P2P ήταν θετική αλλά προβλεπόταν αρνητική

577 FP (False Positive) δηλαδή η P2P ήταν αρνητική αλλά προβλεπόταν θετική και

1.805 TP (True Positive) δηλαδή η P2P ήταν θετική και προβλεπόταν θετική.

	PN	PY
AN	9.471	577
AY	93	1.805

Πίνακας 6.2



Εικόνα 6.3

Οι Πίνακα 6.3 και 6.4 είναι συγκεντρωτικοί πίνακες για την αξιολόγηση της ποιότητας των προβλέψεων της ΛΠ. Ο πίνακας αυτός περιλαμβάνει τις εξής πληροφορίες Ακρίβεια (precision), Ανάκληση (recall), F1-Score, Accuracy και AUC.

Όπως παρατηρείται στους Πίνακες 6.3 και 6.4, για τις αρνητικές υποσχέσεις πληρωμής (without P2P) η ακρίβεια είναι 0,99, η ανάκληση 0,94 και το f1-score 0,97. Ανάλογα για τις θετικές υποσχέσεις (with P2P), οι αντίστοιχες τιμές είναι 0,76, 0,95 και 0,84. Ακόμη, να σημειωθεί ότι η ακρίβεια του μοντέλου βγήκε 94,4%, το AUC 98,1% και το f1-Score 94%, αρκετά μεγάλα ποσοστά για την εύρεση σωστών μελλοντικών προβλέψεων.

	Precision	Recall	F1-Score
Without P2P	0.99	0.94	0.97
With P2P	0.76	0.95	0.84

Πίνακας 6.3

	Logistic Regression Model for P2P
Accuracy	0.944
AUC	0.981
F1-Score	0.94

Πίνακας 6.4

Τέλος, το παραπάνω μοντέλο εφαρμόστηκε στα δεδομένα του Dataset_Out_of_Sample, δημιουργώντας στα δεδομένα μία επιπλέον στήλη P2P_Proba με τις ανάλογες τιμές που προέκυψαν από την εφαρμογή του καθώς και άλλη μία P2P με τις ανάλογες τιμές 1 ή 0 για την πρόβλεψη της θετικής ή όχι υπόσχεσης πληρωμής.

6.2.2 Μοντέλο Λογιστικής Παλινδρόμησης για DPD

Business Dataset 2 (BD2)

Παρακάτω παρουσιάζεται η βασική μεθοδολογία και τα αποτελέσματα της που εφαρμόστηκε στο δεύτερο dataset (BD2) που αφορά τις επιχειρήσεις.

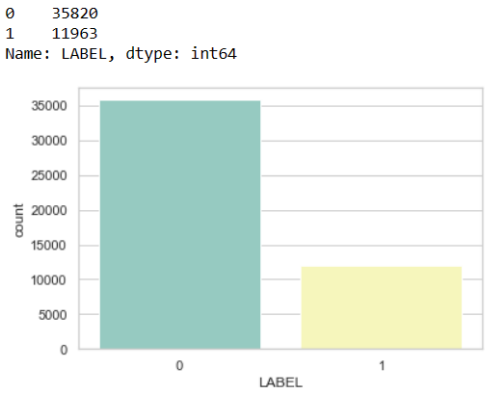
Α΄ Μέρος: Παρουσίαση και Προεργασία των Δεδομένων

Το δεύτερο σύνολο δεδομένων Business Dataset 2 (BD2) περιέχει στοιχεία για τις ίδιες 79.729 επιχειρήσεις με το BD1. Η διαφορά στην συγκεκριμένη περίπτωση είναι ότι η τελευταία στήλη (LABEL) εκφράζει την πιθανότητα να αθετήσει ο πελάτης την υπόσχεση πληρωμής (DPD).

Ομοίως με το BD1, ελέγχθηκαν τα δεδομένα για τυχόν ελλείπουσες τιμές, και διαπιστώθηκε ότι όλες οι στήλες και οι γραμμές είναι πλήρως συμπληρωμένες. Έπειτα, ελέγχθηκαν για τυχόν διπλές γραμμές ή στήλες, ώστε να αφαιρεθούν τα ίδια δεδομένα, αν τυχόν υπάρχουν, αλλά επίσης διαπιστώθηκε ότι όλες γραμμές αφορούν διαφορετικές επιχειρήσεις. Στην συνέχεια, βρέθηκαν οι ποιοτικές μεταβλητές και μέσω του Label Encoding μετατράπηκαν σε ποσοτικές.

Έπειτα, το BD2 χωρίστηκε σε Dataset_Out_of_Sample, αποτελούμενο από 20.000 στοιχεία (περίπου το 25% των δεδομένων) και σε Development_Dataset (75% των δεδομένων). Έπειτα, το Development_Dataset διαχωρίστηκε σε σύνολο εκπαίδευσης (train_set) και σύνολο δοκιμής (test_set) με αναλογία 80% και 20% αντίστοιχα. Σκοπός αυτού του διαχωρισμού είναι να εφαρμοστεί η λογιστική παλινδρόμηση για την πρόβλεψη της αθέτησης ή όχι πληρωμής (DPD) σε νέα δεδομένα που θα προκύψουν στο μέλλον.

Στην συνέχεια, παρατηρήθηκε ότι δεν υπάρχει ισορροπία ανάμεσα στα δεδομένα για την μεταβλητή-στόχο (Εικόνα 6.4), επομένως όπως και πριν εφαρμόστηκε η τεχνική SMOTE για την εξισορρόπηση τους. Τέλος, εφαρμόστηκε η Pearson Correlation, και αφαιρέθηκαν από τα δεδομένα όλες οι στήλες -πλην μίας- που είχαν συσχέτιση από 0,7 και πάνω ή από -0,7 και κάτω.



Εικόνα 6.4

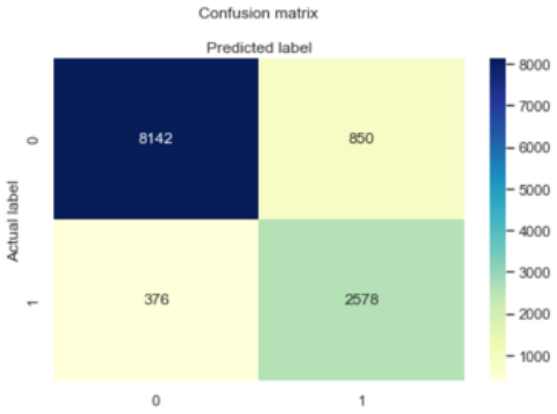
Β΄ Μέρος: Λογιστική Παλινδρόμηση

Η μεθοδολογία που ακολουθήθηκε και εδώ είναι ίδια με το BD1. Αρχικά εφαρμόστηκε η τεχνική Backward elimination, η οποία συνέβαλλε στο να βρεθούν οι 10 σημαντικότερες μεταβλητές για την εφαρμογή της Λογιστική Παλινδρόμησης (Πίνακας 6.5 - Παράρτημα). Στην συνέχεια λοιπόν, εφαρμόστηκε η Λογιστικής Παλινδρόμησης (ΛΠ), όπως περιεγράφηκε στην παράγραφο 3.2.4, με στόχο την πρόβλεψη του DPD (1 αν ο πελάτης αθετήσει την πληρωμή , 0 αν τελικά δεν την αθετήσει και πληρώσει). Τα αποτελέσματα της φαίνονται στους Πίνακες 6.6 και 6.7.

Πιο συγκεκριμένα, ο Πίνακας 6.6 είναι ο συγκεντρωτικός πίνακας των προβλέψεων και των πραγματικών τιμών.

	PN	PY
AN	8.142	850
AY	376	2.578

Πίνακας 6.6



Εικόνα 6.5

Σύμφωνα με τον παραπάνω πίνακα προκύπτουν οι εξής τιμές:

8.142 TN (True Negative) δηλαδή DPD ήταν αρνητική και προβλεπόταν αρνητική

376 FN (False Negative) δηλαδή η DPD ήταν θετική αλλά προβλεπόταν αρνητική

850 FP (False Positive) δηλαδή η DPD ήταν αρνητική αλλά προβλεπόταν θετική και

2.578 TP (True Positive) δηλαδή η DPD ήταν θετική και προβλεπόταν θετική.

Όπως παρατηρείται στους Πίνακες 6.7 και 6.8, παρουσιάζεται ο συγκεντρωτικός πίνακας για την αξιολόγηση της ποιότητας των προβλέψεων της ΛΠ. Όπως παρατηρείται, στην περίπτωση μη αθέτησης της πληρωμής (without DPD) η ακρίβεια είναι 0,96, η ανάκληση 0,91 και το f1-score 0,93. Ανάλογα για την αθέτηση πληρωμής (with DPD), οι αντίστοιχες τιμές είναι 0,75, 0,87 και 0,81. Τέλος, να σημειωθεί ότι η ακρίβεια του μοντέλου βγήκε 89,7%, το AUC 96,7% και το f1-Score 90%, αρκετά μεγάλα ποσοστά για την εύρεση σωστών μελλοντικών προβλέψεων.

	Precision	Recall	F1-Score
Without DPD	0.96	0.91	0.93
With DPD	0.75	0.87	0.81

Πίνακας 6.7

	Logistic Regression Model for DPD
Accuracy	0.897
AUC	0.967
F1-Score	0.90

Πίνακας 6.8

Τέλος, το παραπάνω μοντέλο εφαρμόστηκε στα δεδομένα του Dataset_Out_of_Sample, δημιουργώντας στα δεδομένα μία επιπλέον στήλη DPD_Proba με τις ανάλογες τιμές που προέκυψαν από την εφαρμογή του καθώς και άλλη μία DPD με τις ανάλογες τιμές 1 ή 0 για την πρόβλεψη της αθέτησης ή όχι πληρωμής.

6.2.3 Ασαφές Σύστημα Τύπου Mamdani

Σε συνέχεια της μεθοδολογίας θα εφαρμοστεί η μέθοδος Mandani της Fuzzy Logic, όπως περιεγράφηκε στην παράγραφο 3.3 της παρούσας εργασίας, με σκοπό να προκύψει ένα σκορ για τον κάθε πελάτη, στην κλίμακα του 0-100. Πιο συγκεκριμένα, χρησιμοποιήθηκε το Dataset_Out_of_Sample με τις 20.000 επιχειρήσεις, αφού πρώτα προστέθηκαν σε αυτό οι δύο νέες στήλες P2P_Proba και DPD_Proba, οι οποίες προέκυψαν από την ΛΠ, που περιεγράφηκε παραπάνω, των BD1 και BD2.

Αρχικά, η μεταβλητή DPD_Proba μετατράπηκε σε Good_Proba, ώστε οι μικρές τιμές να δηλώνουν την μεγάλη πιθανότητα αθέτησης πληρωμής ενώ οι μεγάλες τιμές να δηλώνουν μικρή πιθανότητα αθέτησης (δηλαδή όσο πιο κοντά στο 1 είναι η πιθανότητα τόσο πιο καλός πελάτης χαρακτηρίζεται). Συνεπώς,

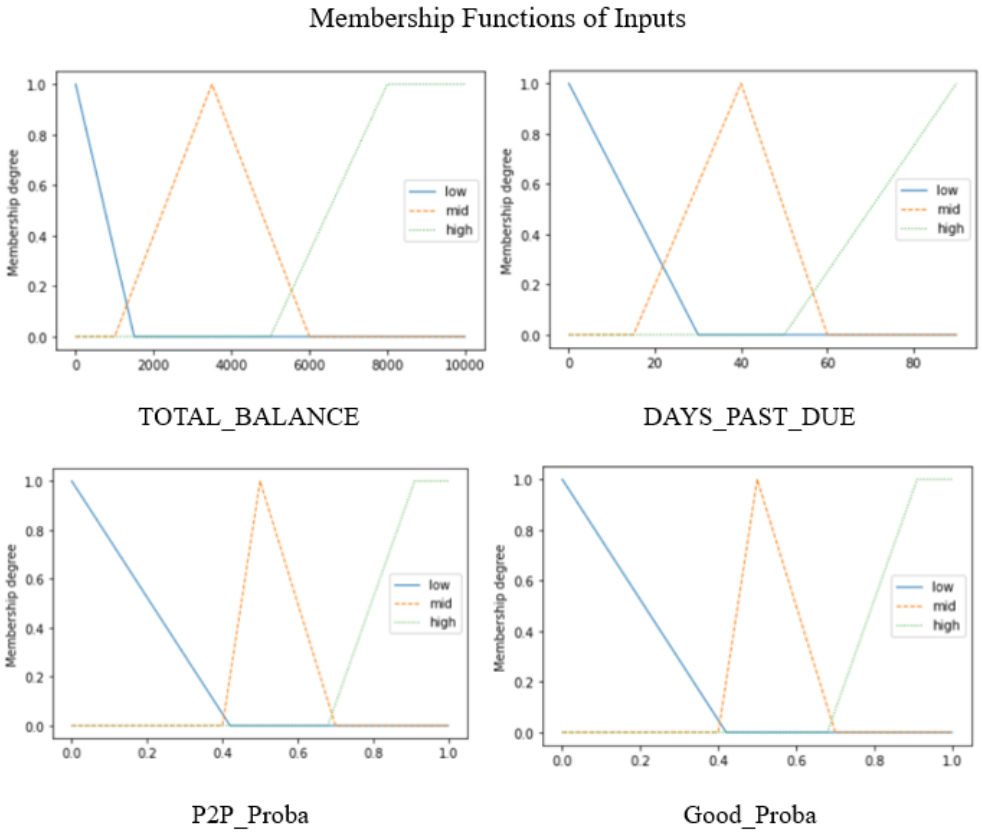
$$Good\_Proba = 1 - DPD\_\text{Proba}$$

Επιπλέον, από τις 190 στήλες του Dataset, κρατήθηκαν σε αυτό μόνο οι πιο σημαντικές (Πίνακας 6.9) οι οποίες θα χρησιμοποιηθούν ως είσοδοι των ασαφών συστημάτων συμπερασμού για το σκοράρισμα του κάθε πελάτη.

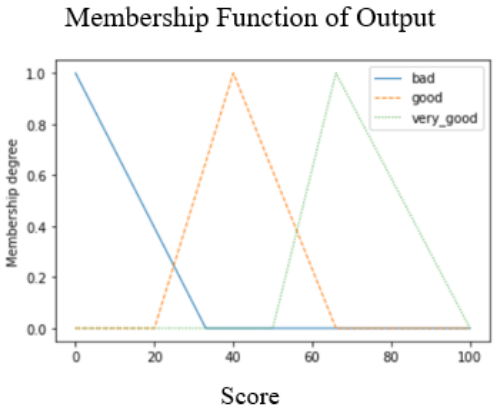
ΜΕΤΑΒΛΗΤΗ	ΠΕΡΙΓΡΑΦΗ
CASE_ID	ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ
TOTAL_BALANCE (TB)	ΣΥΝΟΛΙΚΟ ΑΝΕΞΟΦΛΗΤΟ ΠΟΣΟ
DF_days_active (DA)	ΣΥΝΟΛΟ ΗΜΕΡΩΝ ΑΠΟ ΤΗΝ ΤΕΛΕΥΤΑΙΑ ΠΛΗΡΩΜΗ
P2P_Proba (P)	ΠΙΘΑΝΟΤΗΤΑ ΥΠΟΣΧΕΣΗΣ ΠΛΗΡΩΜΗΣ
Good_Proba (G)	ΠΙΘΑΝΟΤΗΤΑ ΕΓΚΑΙΡΗΣ ΠΛΗΡΩΜΗΣ

Πίνακας 6.9

Η έξοδο (Score) αλλά και κάθε είσοδο έχει μία τριγωνική συνάρτηση συμμετοχής (triangular membership function) που αποτελείται από γλωσσικές μεταβλητές (Εικόνα 6.7 - Παράρτημα). Οι συναρτήσεις συμμετοχής έχουν οριστεί χρησιμοποιώντας προσεγγίσεις που βασίζονται σε δεδομένα και γνώμες εμπειρογνωμόνων. Τα δεδομένα του παρελθόντος έχουν αναλυθεί για κάθε είσοδο και οι κλίμακες και οι συναρτήσεις συμμετοχής στις Εικόνες 6.8.α και 6.8.β έχουν κατασκευαστεί από κοινού με τους εμπειρογνώμονες.



Εικόνα 6.8.α



Εικόνα 6.8.β

Η κάθε συνάρτηση συμμετοχής για κάθε μία από τις εισόδους είναι τριγωνική και ορίζεται από τρείς γλωσσικούς όρους (low – mid – high). Για παράδειγμα, η συνάρτηση συμμετοχής για τον γλωσσικό όρο high της μεταβλητής TOTAL BALANCE ορίζεται από την παρακάτω εξίσωση. Οι υπόλοιπες εξισώσεις των μεταβλητών βρίσκονται στους Πίνακας 6.10.α και 6.10.β (Παράρτημα).

$$\mu_H(x) = \begin{cases} 0 & ,if\ x \leq 1000 \\ \dfrac{x - 5000}{3000} & ,if\ 5000 \leq x \leq 8000 \\ 1 & ,if\ x \geq 8000 \end{cases}$$

Στην συνέχεια δημιουργήθηκαν 49 κανόνες fuzzy (fuzzy rules) (Πίνακας 6.11 – Παράρτημα), οι οποίοι οδηγούν ο κάθε ένας και σε ένα γλωσσικό όρο του Score ανά πελάτη (bad – good – very good). Με βάση αυτά, δημιουργήθηκε ένα Score για την κάθε εταιρεία ανάλογα με τις τιμές εισόδου της στην κλίμα 0-100 και μέσω αυτού άλλη μία στήλη με τους γλωσσικούς όρους που χαρακτήριζαν τον κάθε πελάτη (Customer) ως bad, good ή very good.

6.2.4 Πολυκριτήρια Ανάλυση Αποφάσεων

Κατά το τελικό στάδιο της εργασίας, σκοπός είναι να βρεθεί η πιο κατάλληλη στρατηγική προσέγγισης του κάθε πελάτη για την είσπραξη ληξιπρόθεσμων λογαριασμών με την μέθοδο AHP. Έπειτα από έρευνα παλαιότερων εργασιών (Sezi Cevik Onar et al., 2015) διαμορφώθηκαν 9 διαφορετικές στρατηγικές- 3 ανά κάθε κατηγορία πελάτη (Πίνακας 6.12- Παράρτημα) και κάθε μία από τις τρεις στρατηγικές βαθμολογήθηκε ως προς δύο κριτήρια, το Good_Proba που προέκυψε μέσω της ΛΠ και

το Score που προέκυψε από την μέθοδο Fuzzy. Οι λήπτες αποφάσεων εισάγουν τις προτιμήσεις τους όσον αφορά την αξιολόγηση των στρατηγικών σε έναν τετραγωνικό πίνακα με κλίμακα από το 1 έως το 9. Οι στρατηγικές και οι ανάλογες βαθμολογήσεις κάθε φορά καθορίζονται από τον εκάστοτε ειδικό ο οποίος τις καθορίζει ανάλογα την εμπειρία και τις γνώσεις του. Έπειτα, μέσω του λογισμικού PriEsT (Priority Estimation Tool) βρήκαν τα βάρη για κάθε μία στρατηγική (Πίνακας 6.13).

	STRATEGY 1	STATEGY 2	STRATEGY 3
Good_Proba	0.674	0.226	0.101
Score	0.238	0.625	0.136

Πίνακας 6.13

Τα βάρη αυτά χρησιμοποιήθηκαν με σκοπό την εύρεση καλύτερης στρατηγικής για κάθε πελάτη ανάλογα με τα τις τιμές κριτηρίων που προέκυψαν από τις παραπάνω μεθοδολογίες.

	STRATEGIES								
	Str-VG1	Str-VG2	Str-VG3	Str-G1	Str-G2'	Str-G3'	Str-B1	Str-B2	Str-B3
COUNT	19.719	0	0	0	281	0	0	0	0

Πίνακας 6.14

Συμπεραίνουμε λοιπόν από τον Πίνακα 6.14 ότι στις 19.719 εταιρείες συστήνεται να προτείνουμε την $1^{η}$ στρατηγική των πολύ καλών πελατών και στις υπόλοιπες 281 συστήνεται η 2 στρατηγική των καλών πελατών.

6.3 Εφαρμογή της Μεθόδου σε Φυσικά Πρόσωπα

Όμοια επεξεργασία και μεθοδολογία, με την παραπάνω, εφαρμόστηκε και σε πελάτες της ίδια εταιρεία που αφορούν λογαριασμούς φυσικών προσώπων. Παρακάτω επισημαίνονται τα βασικά στοιχεία με τις ανάλογες διαφοροποιήσεις που αφορούν την εφαρμογή αυτή.

6.3.1 Μοντέλο Λογιστικής Παλινδρόμησης για P2P

Retail Dataset 1 (RD1)

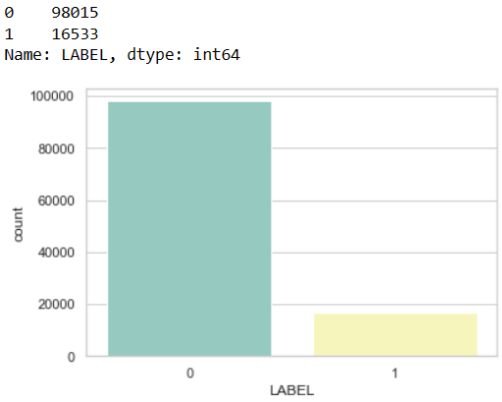
Παρακάτω παρουσιάζεται η βασική μεθοδολογία και τα αποτελέσματα της που εφαρμόστηκε στο πρώτο dataset (RD1) που αφορά τα σπίτια.

Α΄ Μέρος: Παρουσίαση και Προεργασία των Δεδομένων

Το αρχικό σύνολο δεδομένων του Rome Dataset 1 (RD1) περιέχει στοιχεία 191.186 φυσικών προσώπων με 188 δεδομένα για το κάθε ένα από αυτά τα οποία δίνουν πληροφορίες για το χρωστούμενο ποσό, τον χρονικό διάστημα της οφειλής, το χρονικό διάστημα από την τελευταία επικοινωνία με τον πελάτη κ.ο.κ. Η τελευταία στήλη (LABEL), όπως αναφέρθηκε πιο πάνω, αφορά την θετική υπόσχεση πληρωμής (P2P).

Ομοίως με την μέθοδο που εφαρμόστηκε στις επιχειρήσεις, ελέγχθηκαν ελλείπουσες τιμές, διπλές γραμμές και στήλες και εφαρμόστηκε Label Encoding στις 12 ποιοτικές μεταβλητές. Έπειτα, το σύνολο δεδομένων χωρίστηκε σε Dataset_Out_of_Sample, αποτελούμενο από 48.000 στοιχεία (περίπου το 25% των δεδομένων) και σε Development_Dataset (75% των δεδομένων). Έπειτα, το Development_Dataset διαχωρίστηκε σε σύνολο εκπαίδευσης και σύνολο δοκιμής με αναλογία 80% και 20% αντίστοιχα ώστε να εφαρμοστεί η λογιστική παλινδρόμηση για την πρόβλεψη της θετικής υπόσχεσης πληρωμής (P2P).

Πριν την μέθοδο της λογιστικής παλινδρόμησης, εξασφαλίστηκε η ισορροπία της μεταβλητής στόχου (LABEL) με την μέθοδο SMOTE, καθώς παρατηρήθηκε ανομοιομορφία (Εικόνα 6.9). Επιπλέον, αφαιρέθηκαν οι υψηλά συσχετιζόμενες μεταβλητές με απόλυτη συσχέτιση πάνω από 0,7, οι οποίες βρέθηκαν με Pearson Correlation.



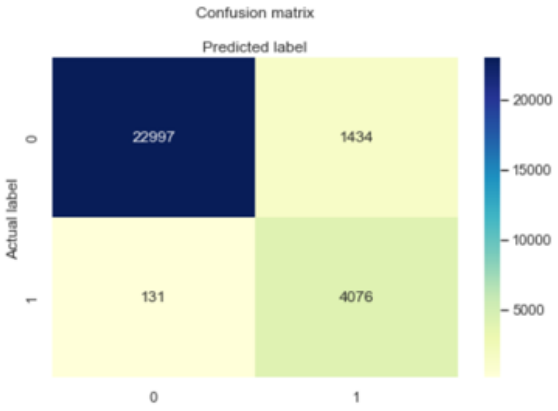
Εικόνα 5.9

Β΄ Μέρος: Λογιστική Παλινδρόμηση

Ομοίως με την παραπάνω εφαρμογή, για την επιλογή των σημαντικότερων χαρακτηριστικών επιλέχθηκε η τεχνική Backward elimination (Πίνακας 6.15 – Παράρτημα). Στην συνέχεια, εφαρμόστηκε η Λογιστικής Παλινδρόμησης (ΛΠ), όπως περιεγράφηκε στην παράγραφο 3.2.4, με στόχο την πρόβλεψη του P2P (1 αν λάβουμε θετική απάντηση πληρωμής, 0 αν λάβουμε αρνητική). Τα αποτελέσματα της φαίνονται στους Πίνακες 6.16 και 6.17.

	PN	PY
AN	22.997	1.434
AY	131	4.076

Πίνακας 6.16



Εικόνα 6.10

Πιο συγκεκριμένα, από τον συγκεντρωτικό πίνακας 5.14 προκύπτει ότι:

22.997 TN (True Negative) δηλαδή η P2P ήταν αρνητική και προβλεπόταν αρνητική

131 FN (False Negative) δηλαδή η P2P ήταν θετική αλλά προβλεπόταν αρνητική

1.434 FP (False Positive) δηλαδή η P2P ήταν αρνητική αλλά προβλεπόταν θετική και

4.076 TP (True Positive) δηλαδή η P2P ήταν θετική και προβλεπόταν θετική.

Όπως παρατηρείται από τον συγκεντρωτικό πίνακας για την αξιολόγηση της ποιότητας των προβλέψεων (Πίνακας 6.17), για τις αρνητικές υποσχέσεις πληρωμής (without P2P) η ακρίβεια είναι 0,99, η ανάκληση 0,94 και το f1-score 0,97. Ανάλογα για τις θετικές υποσχέσεις (with P2P), οι αντίστοιχες τιμές είναι 0,74, 0,97 και 0,84. Τέλος, να σημειωθεί ότι η ακρίβεια του μοντέλου βγήκε 94,5%, το AUC 96,8% και το f1-Score 95%, αρκετά μεγάλα ποσοστά για την εύρεση σωστών μελλοντικών προβλέψεων.

	Precision	Recall	F1-Score
Without P2P	0.99	0.94	0.97
With P2P	0.74	0.97	0.84

Πίνακας 6.17

	Logistic Regression Model for P2P
Accuracy	0.945
AUC	0.968
F1-Score	0.95

Πίνακας 6.18

Τέλος, το παραπάνω μοντέλο εφαρμόστηκε στα δεδομένα του Dataset_Out_of_Sample, δημιουργώντας στα δεδομένα μία επιπλέον στήλη P2P_Proba με τις ανάλογες τιμές που προέκυψαν από την εφαρμογή του καθώς και άλλη μία P2P με τις ανάλογες τιμές 1 ή 0 για την πρόβλεψη της θετικής ή όχι υπόσχεσης πληρωμής.

6.3.2 Μοντέλο Λογιστικής Παλινδρόμησης για DPD

Retail Dataset 2 (RD2)

Παρακάτω παρουσιάζεται η βασική μεθοδολογία και τα αποτελέσματα της που εφαρμόστηκε στο δεύτερο dataset (HD2) που αφορά τα σπίτια.

Α΄ Μέρος: Παρουσίαση και Προεργασία των Δεδομένων

Όσον αφορά το δεύτερο σύνολο δεδομένων Retail Dataset 2 (RD2) περιέχει στοιχεία για τις ίδια 191.186 σπίτια με το RD1. Η διαφορά στην συγκεκριμένη περίπτωση είναι ότι η τελευταία στήλη (LABEL) εκφράζει την πιθανότητα να αθετήσει ο πελάτης την υπόσχεση πληρωμής (DPD). Η μεθοδολογία που εφαρμόστηκε είναι ακριβώς ίδια με τις προαναφερθείσες δηλαδή έλεγχος ελλειπουσών, έλεγχος διπλών γραμμών/ στηλών, Label encoding, τεχνική SMOTE για την μεταβλητή στόχο και Pearson Correlation με σκοπό της παράλειψη ισχυρά συσχετιζόμενων στηλών.

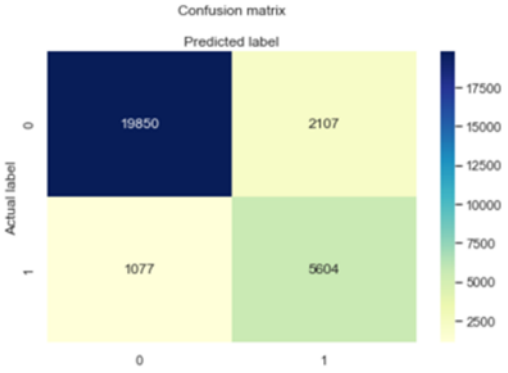
Β΄ Μέρος: Λογιστική Παλινδρόμηση

Ομοίως λοιπόν, εφαρμόστηκε Backward elimination για την εύρεση των 10 σημαντικότερων χαρακτηριστικών (Πίνακας 6.19 - Παράρτημα) και έπειτα εφαρμόστηκε ΛΠ.

Πιο συγκεκριμένα, ο Πίνακας 6.20 είναι ο συγκεντρωτικός πίνακας των προβλέψεων και των πραγματικών τιμών.

	PN	PY
AN	19.850	2.107
AY	1.077	5.604

Πίνακας 6.20



Εικόνα 6.11

Σύμφωνα με τον παραπάνω πίνακα προκύπτουν οι εξής τιμές:

19.850 TN (True Negative) δηλαδή DPD ήταν αρνητική και προβλεπόταν αρνητική

1.077 FN (False Negative) δηλαδή η DPD ήταν θετική αλλά προβλεπόταν αρνητική

2.107 FP (False Positive) δηλαδή η DPD ήταν αρνητική αλλά προβλεπόταν θετική και

5.604 TP (True Positive) δηλαδή η DPD ήταν θετική και προβλεπόταν θετική.

Στον Πίνακα 6.21 παρατηρείται ότι στην περίπτωση μη αθέτησης της πληρωμής (without DPD) η ακρίβεια είναι 0,95, η ανάκληση 0,90 και το f1-score 0,93. Ανάλογα για την αθέτηση πληρωμής (with DPD), οι αντίστοιχες τιμές είναι 0,73, 0,84 και 0,78. Τέλος, να σημειωθεί ότι η ακρίβεια του μοντέλου βγήκε 88,9%, το AUC 95,4% και το f1-Score 89%, αρκετά μεγάλα ποσοστά για την εύρεση σωστών μελλοντικών προβλέψεων.

	Precision	Recall	F1-Score
Without DPD	0.95	0.90	0.93
With DPD	0.73	0.84	0.78

Πίνακας 6.21

	Logistic Regression Model for DPD
Accuracy	0.889
AUC	0.954
F1-Score	0.89

Πίνακας 6.22

Τέλος, το παραπάνω μοντέλο εφαρμόστηκε στα δεδομένα του Dataset_Out_of_Sample, δημιουργώντας στα δεδομένα μία επιπλέον στήλη DPD_Proba

με τις ανάλογες τιμές που προέκυψαν από την εφαρμογή του καθώς και άλλη μία DPD με τις ανάλογες τιμές 1 ή 0 για την πρόβλεψη της αθέτησης ή όχι πληρωμής.

6.3.3 Ασαφές Σύστημα Τύπου Mamdani

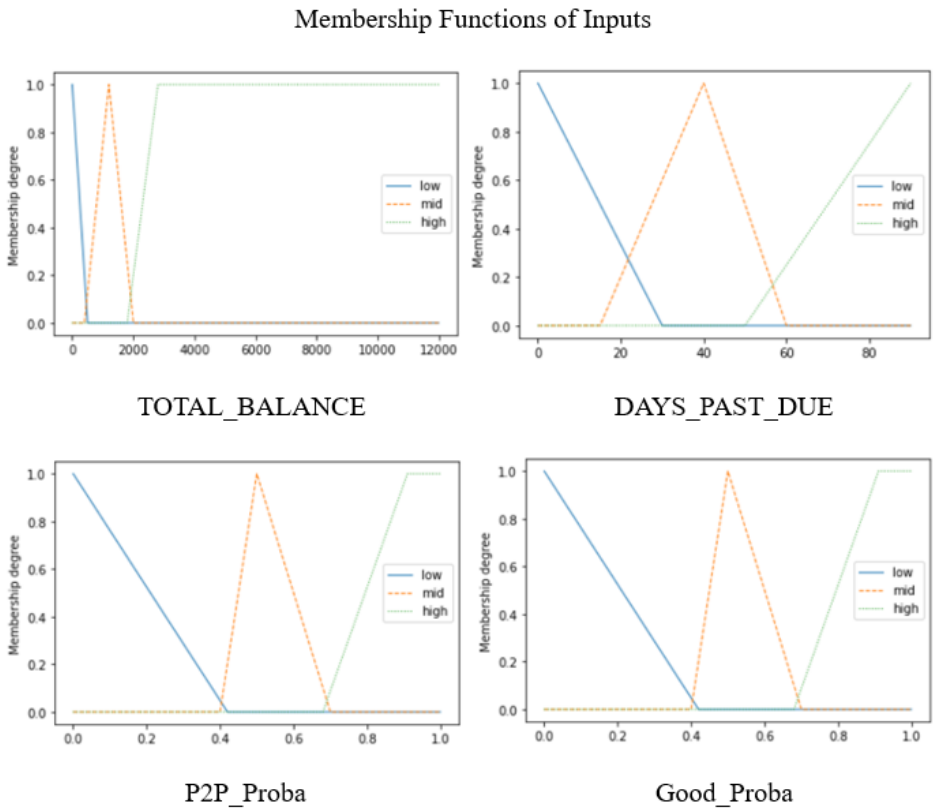
Στην συνέχεια, έχοντας όλα τα δεδομένα του Dataset_Out_of_Sample με τα 48.000 φυσικά πρόσωσπα και έχοντας προσθέσει τις στήλες P2P_Proba, DPD_Proba και Good_Proba (όπου Good_Proba = 1 - DPD_Proba) εφαρμόζεται η μέθοδος Mandani της Fuzzy Logic με σκοπό να προκύψει ένα σκορ για τον κάθε πελάτη, στην κλίμακα του 0-100. Επιπλέον, από τις 190 στήλες του Dataset, κρατήθηκαν μόνο οι πιο σημαντικές (Πίνακας 6.23) οι οποίες θα χρησιμοποιηθούν ως είσοδοι των ασαφών συστημάτων συμπερασμού για το σκοράρισμα του κάθε πελάτη.

ΜΕΤΑΒΛΗΤΗ	ΠΕΡΙΓΡΑΦΗ
CASE_ID	ΚΩΔΙΚΟΣ ΠΕΛΑΤΗ
TOTAL_BALANCE (TB)	ΣΥΝΟΛΙΚΟ ΑΝΕΞΟΦΛΗΤΟ ΠΟΣΟ
DF_days_active (DA)	ΣΥΝΟΛΟ ΗΜΕΡΩΝ ΑΠΟ ΤΗΝ ΤΕΛΕΥΤΑΙΑ ΠΛΗΡΩΜΗ
P2P_Proba (P)	ΠΙΘΑΝΟΤΗΤΑ ΥΠΟΣΧΕΣΗΣ ΠΛΗΡΩΜΗΣ
Good_Proba (G)	ΠΙΘΑΝΟΤΗΤΑ ΕΓΚΑΙΡΗΣ ΠΛΗΡΩΜΗΣ

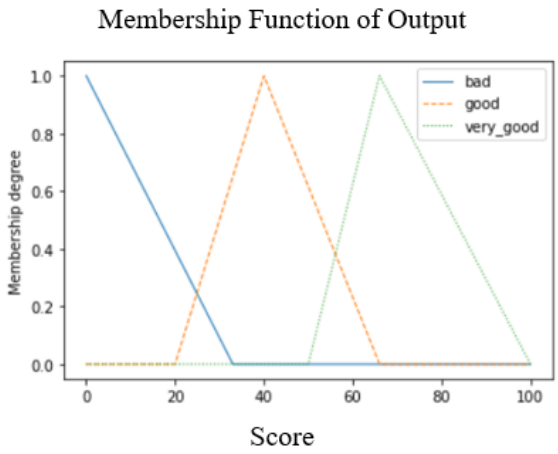
Πίνακας 6.23

Η έξοδο (Score) αλλά και κάθε είσοδο έχει μία τριγωνική συνάρτηση συμμετοχής που αποτελείται από γλωσσικές μεταβλητές (Εικόνα 6.7 - Παράρτημα). Οι συναρτήσεις συμμετοχής έχουν οριστεί χρησιμοποιώντας προσεγγίσεις που βασίζονται σε δεδομένα με γνώμονα παλιότερες έρευνες αλλά και μετά από εξέταση βασικών στατιστικών αναλύσεων των τιμών που έχει κάθε μία από τις μεταβλητές. Παρ' όλα αυτά ο κάθε χρήστης μπορεί να ορίσει τις δικιές του τιμές ανάλογα με την εμπειρία και τις γνώσεις του και να λειτουργεί εξίσου καλά σαν μέθοδος. Τα Δεδομένα έχουν αναλυθεί για κάθε είσοδο και οι κλίμακες και οι συναρτήσεις συμμετοχής στις Εικόνες 6.12.α και 6.12.β έχουν κατασκευαστεί. Ομοίως με το BD κάθε συνάρτηση συμμετοχής για κάθε μία από τις εισόδους είναι τριγωνική και ορίζεται από τρείς γλωσσικούς όρους (low – mid – high). Οι εξισώσεις των μεταβλητών βρίσκονται στους Πίνακας 6.24.α και 6.24.β (Παράρτημα).

Στην συνέχεια χρησιμοποιήθηκαν οι ίδιοι 49 κανόνες fuzzy (Πίνακας 6.11 – Παράρτημα), οι οποίοι οδηγούν ο κάθε ένας και σε ένα γλωσσικό όρο του Score ανά πελάτη (bad – good – very good). Με βάση αυτά, δημιουργήθηκε ένα Score για την κάθε εταιρεία ανάλογα με τις τιμές εισόδου της στην κλίμα 0-100 και μέσω αυτού άλλη μία στήλη με τους γλωσσικούς όρους που χαρακτήριζαν τον κάθε πελάτη (Customer) ως bad, good ή very good.



Εικόνα 6.12.α



Εικόνα 6.12.β

6.3.4 Πολυκριτήρια Ανάλυση Αποφάσεων

Κατά το τελικό στάδιο της εργασίας, σκοπός είναι να βρεθεί η πιο κατάλληλη στρατηγική προσέγγισης του κάθε πελάτη για την είσπραξη ληξιπρόθεσμων λογαριασμών με την μέθοδο AHP. Οι στρατηγικές και οι βαθμολογήσεις που χρησιμοποιήθηκαν στην παρούσα εργασία είναι ίδιες με την περίπτωση του BD, παρ' όλα

αυτά, κάθε φορά μπορούν να προσαρμόζονται από τον εκάστοτε ειδικό. Έπειτα, μέσω του λογισμικού PriEsT (Priority Estimation Tool) βρήκαν τα βάρη για κάθε μία στρατηγική (Πίνακας 6.25).

	STRATEGY 1	STATEGY 2	STRATEGY 3
Good_Proba	0.674	0.226	0.101
Score	0.238	0.625	0.136

Πίνακας 6.25

Τα βάρη αυτά χρησιμοποιήθηκαν με σκοπό την εύρεση καλύτερης στρατηγικής για κάθε πελάτη ανάλογα με τα τις τιμές κριτηρίων που προέκυψαν από τις παραπάνω μεθοδολογίες.

	STRATEGIES								
	Str-VG1	Str-VG2	Str-VG3	Str-G1	Str-G2'	Str-G3'	Str-B1	Str-B2	Str-B3
COUNT	46.301	0	0	48	1.651	0	0	0	0

Πίνακας 6.26

Συμπεραίνουμε λοιπόν από τον Πίνακα 6.26 ότι από τα 48.000 φυσικά πρόσωπα συστήνεται για τους πολύ καλούς πελάτες (very good) να προτείνουμε την $1^{η}$ στρατηγική και στις υπόλοιπες 1699 που έχουν χαρακτηριστεί ως καλοί πελάτες (good) συστήνεται η 1η στρατηγική στους 48 και η $2^{η}$ στους 1.651.

Κεφάλαιο 7ο:
Συμπεράσματα και Μελλοντικές Επεκτάσεις

Συνοψίζοντας, το πρόβλημα των ληξιπρόθεσμων οφειλών είναι ευρέως γνωστό στον χρηματοοικονομικό κλάδο (π.χ., τράπεζες, εταιρείες παροχής ενέργειας, κλπ.). Για παράδειγμα, μία τράπεζα μπορεί να έχει εγκρίνει μεγάλο αριθμό καταναλωτικών δανείων ή ένα νοικοκυριό μπορεί να έχει καταναλώσει υψηλό αριθμό kWh, οπότε η μη καταβολή οφειλών να προκαλέσει οικονομική ζημία για την τράπεζα ή την εταιρεία παροχής ηλεκτρικού ρεύματος, αντίστοιχα. Συνεπώς, είναι απαραίτητη η σωστή αξιολόγηση των πελατών λαμβάνοντας υπόψη πολλαπλούς παράγοντες (ποσοτικούς, ποιοτικούς). Η παρούσα μελέτη εξετάζει τα στοιχεία πελατών μίας εταιρείας και προσπαθεί να κατηγοριοποιήσει τους πελάτες με βάση τα χαρακτηριστικά τους και να βρει την κατάλληλη μέθοδο προσέγγισης με στόχο την είσπραξη ληξιπρόθεσμων οφειλών. Τα αποτελέσματα που προέκυψαν από την παρούσα μεταπτυχιακή διατριβή συμβάλουν στην προώθηση της επιστήμης στον τομέα της είσπραξης οφειλών, καθιερώνοντας ένα μεθοδολογικό πλαίσιο για ακριβέστερες και αποτελεσματικότερες στρατηγικές επικοινωνίας με τον πελάτη.

Στόχος της έρευνας αποτέλεσε η προσπάθεια βελτιστοποίησης του τομέα της είσπραξης οφειλών διότι παρά την πληθώρα προγνωστικών μοντέλων που έχουν σχεδιαστεί για τη διαχείριση κινδύνων η είσπραξη οφειλών έχει λάβει ελάχιστη προσοχή στην ακαδημαϊκή κοινότητα. Επιπλέον, κάθε πελάτης είναι ξεχωριστός και οι εισπράκτορες, έχοντας πρωτίστως διασφαλίσει ηθικές βάσεις, πρέπει να είναι πολύ προσεκτικοί όσο αναφορά τη συμπεριφορά τους προς τους πελάτες και λόγω των νομοθεσιών σχετικά με την είσπραξη οφειλών.

Για την επίτευξη του στόχου χρησιμοποιήθηκε μια καινοτόμα και πολύπλευρη μεθοδολογία. Εμβάθυνε στις θεωρητικές βάσεις της είσπραξης οφειλών και εξέτασε διάφορες τεχνικές κατηγοριοποίησης και αξιολόγησης πελατών. Η ανάλυση της μελέτης περίπτωσης πραγματοποιήθηκε με τη χρήση πραγματικών δεδομένων και ανέδειξε την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας στην ακριβή αξιολόγηση του πελάτη. Αξίζει να σημειωθεί ότι το προτεινόμενο μεθοδολογικό πλαίσιο, λαμβάνει υπόψιν την εμπειρία και γνώση ενός ειδικού στον χώρο των ληξιπρόθεσμων οφειλών μοντελοποιώντας κατάλληλα τις προτιμήσεις του και διασφαλίζοντας ταυτόχρονα μία ορθολογική διαδικασία λήψης αποφάσεων (rational decision making). Στο πλαίσιο της διπλωματικής εργασίας, οι προτιμήσεις του ειδικού δόθηκαν ως ενδεικτικό παράδειγμα

εφαρμογής διατηρώντας χαμηλό δείκτη ασυνέπειας. Ωστόσο, αξίζει να αναφερθεί πως το προτεινόμενο μεθοδολογικό πλαίσιο επιτρέπει στον οποιοδήποτε ειδικό του χώρου να εκφράσει τις προτιμήσεις του με έναν ελεγχόμενο τρόπο ούτως ώστε ο δείκτης ασυνέπειας να διατηρείται κάτω από το 10%. Ως εκ τούτου, δεν υπάρχει λάθος ή σωστή απόφαση, καθώς οποιοδήποτε αποτέλεσμα που προκύπτει μετά από σωστή μοντελοποίηση των προτιμήσεων ενός ειδικού και το οποίο ικανοποιεί τα κατάλληλα στατιστικά μέτρα απόδοσης (π.χ., δείκτης ασυνέπειας < 10%) είναι αποδεκτό.

Μερικοί από τους περιορισμούς της διατριβής σχετίζονται με την μη ύπαρξη πολλών πηγών (σχετικών με ληξιπρόθεσμες οφειλές) δεδομένων στην βιβλιογραφία. Στην παρούσα διατριβή, παρά το γεγονός ότι αξιοποιήθηκαν πραγματικά δεδομένα, αυτά αφορούσαν περιορισμένο αριθμό περιπτώσεων από το χαρτοφυλάκιο μίας εταιρείας παροχής ενέργειας και για συγκεκριμένη χρονική περίοδο. Ως εκ τούτου, περισσότερα σε όγκο, αλλά και σε βάθος χρόνου, δεδομένα θα ήταν χρήσιμο να αξιοποιηθούν κατά την μοντελοποίηση των προβλημάτων. Επιπρόσθετα, θα παρουσίαζε ιδιαίτερο ενδιαφέρον η διαθεσιμότητα ενός ή μίας ομάδας ειδικών στο χώρο προκειμένου να ορίσει με βέλτιστο τρόπο (βάσει εμπειρίας) στρατηγικές ληξιπρόθεσμων οφειλών και να δηλώσει τις σχετικές προτιμήσεις στο πλαίσιο της πολυκριτήριας ανάλυσης αποφάσεων.

Ως μελλοντικές επεκτάσεις της παρούσας μεταπτυχιακής διατριβής, θα μπορούσαν να αναφερθούν εναλλακτικά υπολογιστικά πειράματα χρησιμοποιώντας: (1) περισσότερα από ένα (πλέον της Λογιστικής Παλινδρόμησης) μοντέλα μηχανικής μάθησης (π.χ., Νευρωνικά Δίκτυα, Τυχαία Δάση, κλπ.), (2) επιπλέον τεχνικές επιλογής μεταβλητών (π.χ., Γενετικοί Αλγόριθμοι), (3) επιπλέον τεχνικές πολυκριτήριας ανάλυσης αποφάσεων (π.χ., TOPSIS, VIKOR, WSM, κλπ.), (4) μεγαλύτερο όγκο δεδομένων, καθώς επίσης και (5) πραγματικές προτιμήσεις ειδικών του χώρου.

Παράρτημα

6ο ΚΕΦΑΛΑΙΟ:

Πίνακας 6.1: Σύνολο επιλεγόμενων μεταβλητών για το P2P του BD1 έπειτα από Backward Selection

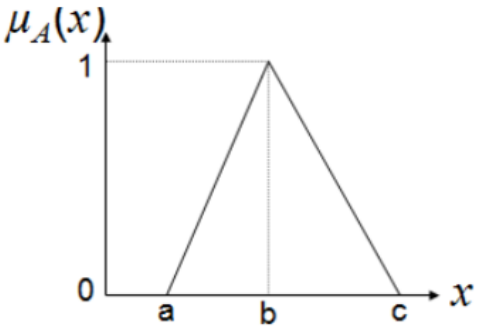
Σύνολο επιλεγόμενων μεταβλητών για το P2P του BD1
'DF_times_a_payment_has_been_related_with_contact'
'DF_is_payment_related_to_contact_now'
'DF_number_of_related_to_the_case_queues'
DAYS_TILL_SETTLE_FRACTION'
'DF_collector_debtor_actions'
'DF_case_is_related_to_contact_settlement_now'
'DF_number_of_PTPs_given'
'DF_last_percentage_of_keeping_promise'
'DF_max_perc_paid_to_promised_amount']
'DF_person_ex_lookup1'

Πίνακας 6.5: Σύνολο επιλεγόμενων μεταβλητών για το DPD του BD2 έπειτα από Backward Selection

Σύνολο επιλεγόμενων μεταβλητών για το DPD του BD2
'DF_current_phase_id'
'DF_min_phone_contact_usage'
'DF_product_description'
'DAYS_IN_HANDLING_QUEUE'
'TIMES_IN_COLLECTION'
'DAYS_PAST_DUE'
'CAPTION'
'DF_collector_debtor_actions'
'DF_last_contact_day_of_week'
'DF_last_status'

Εικόνα 6.7: Τριγωνική συνάρτηση συμμετοχής

$$\mu_A(x) = \begin{cases} 0 & if\ x \le a \\ \dfrac{x-a}{b-a} & if\ a \le x \le b \\ \dfrac{c-x}{c-b} & if\ b \le x \le c \\ 0 & if\ x \ge c \end{cases}$$



Πίνακας 6.10.α: Συναρτήσεις συμμετοχής μεταβλητών εισόδου

TOTAL BALANCE	DAYS_PAST_DUE
$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{1500 - x}{1500} & ,if\ 0 \le x \le 1500 \\ 0 & ,if\ x \ge 8000 \end{cases}$	$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{30 - x}{30} & ,if\ 0 \le x \le 30 \\ 0 & ,if\ x \ge 30 \end{cases}$
$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 1000 \\ \frac{x - 1000}{2500} & ,if\ 1000 \le x \le 3500 \\ \frac{6000 - x}{2500} & ,if\ 3500 \le x \le 6000 \\ 0 & ,if\ x \ge 6000 \end{cases}$	$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 15 \\ \frac{x - 15}{25} & ,if\ 15 \le x \le 40 \\ \frac{60 - x}{20} & ,if\ 40 \le x \le 60 \\ 0 & ,if\ x \ge 60 \end{cases}$
$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 1000 \\ \frac{x - 5000}{3000} & ,if\ 5000 \le x \le 8000 \\ 1 & ,if\ x \ge 8000 \end{cases}$	$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 50 \\ \frac{x - 50}{40} & ,if\ 50 \le x \le 90 \\ 1 & ,if\ x \ge 90 \end{cases}$

P2P_Proba	Good_Proba
$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{0.42 - x}{0.42} & ,if\ 0 \le x \le 0.42 \\ 0 & ,if\ x \ge 0.42 \end{cases}$	$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{0.42 - x}{0.42} & ,if\ 0 \le x \le 0.42 \\ 0 & ,if\ x \ge 0.42 \end{cases}$
$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 0.4 \\ \frac{x - 0.4}{0.1} & ,if\ 0.4 \le x \le 0.5 \\ \frac{0.7 - x}{0.2} & ,if\ 0.5 \le x \le 0.7 \\ 0 & ,if\ x \ge 0.7 \end{cases}$	$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 0.4 \\ \frac{x - 0.4}{0.1} & ,if\ 0.4 \le x \le 0.5 \\ \frac{0.7 - x}{0.2} & ,if\ 0.5 \le x \le 0.7 \\ 0 & ,if\ x \ge 0.7 \end{cases}$
$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 0.68 \\ \frac{x - 0.68}{0.23} & ,if\ 0.68 \le x \le 0.91 \\ 1 & ,if\ x \ge 0.91 \end{cases}$	$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 0.68 \\ \frac{x - 0.68}{0.23} & ,if\ 0.68 \le x \le 0.91 \\ 1 & ,if\ x \ge 0.91 \end{cases}$

Πίνακας 6.10.β: Συνάρτηση συμμετοχής μεταβλητής εξόδου

Score
$\mu_L(x) = \begin{cases} 1 & ,if\ x \leq 0 \\ \frac{33-x}{33} & ,if\ 0 \leq x \leq 33 \\ 0 & ,if\ x \geq 33 \end{cases}$
$\mu_M(x) = \begin{cases} 0 & ,if\ x \leq 20 \\ \frac{x-20}{20} & ,if\ 20 \leq x \leq 40 \\ \frac{66-x}{26} & ,if\ 40 \leq x \leq 66 \\ 0 & ,if\ x \geq 66 \end{cases}$
$\mu_H(x) = \begin{cases} 0 & ,if\ x \leq 50 \\ \frac{x-50}{16} & ,if\ 50 \leq x \leq 66 \\ \frac{100-x}{34} & ,if\ 66 \leq x \leq 100 \\ 0 & ,if\ x \geq 100 \end{cases}$

Πίνακας 6.11: Fuzzy Rules for BD

R1="IF (TOTAL_BALANCE IS low) AND (Good_Proba IS high) THEN (SCORE IS very_good)"

R2="IF (TOTAL_BALANCE IS low) AND (Good_Proba IS mid) THEN (SCORE IS very_good)"

R3="IF (TOTAL_BALANCE IS low) AND (Good_Proba IS low) THEN (SCORE IS good)"

R4="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS high) AND (Good_Proba IS high) THEN (SCORE IS good)"

R5="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS high) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R6="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS high) AND (Good_Proba IS low) THEN (SCORE IS good)"

R7="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS mid) AND (Good_Proba IS high) THEN (SCORE IS good)"

R8="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS mid) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R9="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS mid) AND (Good_Proba IS low) THEN (SCORE IS good)"

R10="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS low) AND (Good_Proba IS high) THEN (SCORE IS good)"

R11="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS low) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R12="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS low) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R13="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS high) AND (Good_Proba IS high) THEN (SCORE IS good)"

R14="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS high) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R15="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS high) AND (Good_Proba IS low) THEN (SCORE IS good)"

R16="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS mid) AND (Good_Proba IS high) THEN (SCORE IS good)"

R17="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS mid) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R18="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS mid) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R19="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS low) AND (Good_Proba IS high) THEN (SCORE IS good)"

R20="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS low) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R21="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS low) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R22="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS high) AND (Good_Proba IS high) THEN (SCORE IS good)"

R23="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS high) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R24="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS high) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R25="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS mid) AND (Good_Proba IS high) THEN (SCORE IS good)"

R26="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS mid) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R27="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS mid) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R28="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS low) AND (Good_Proba IS high) THEN (SCORE IS bad)"

R29="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS low) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R30="IF (TOTAL_BALANCE IS mid) AND (DAYS_PAST_DUE IS high) AND (P2P_Proba IS low) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R31="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS high) AND (Good_Proba IS high) THEN (SCORE IS good)"

R32="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS high) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R33="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS high) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R34="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS mid) AND (Good_Proba IS high) THEN (SCORE IS good)"

R35="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS mid) AND (Good_Proba IS mid) THEN (SCORE IS good)"

R36="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS mid) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R37="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS low) AND (Good_Proba IS high) THEN (SCORE IS good)"

R38="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS low) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R39="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS low) AND (P2P_Proba IS low) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R40="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS high) AND (Good_Proba IS high) THEN (SCORE IS good)"

R41="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS high) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R42="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS high) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R43="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS mid) AND (Good_Proba IS high) THEN (SCORE IS bad)"

R44="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS mid) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R45="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS mid) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R46="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS low) AND (Good_Proba IS high) THEN (SCORE IS good)"

R47="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS low) AND (Good_Proba IS mid) THEN (SCORE IS bad)"

R48="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS mid) AND (P2P_Proba IS low) AND (Good_Proba IS low) THEN (SCORE IS bad)"

R49="IF (TOTAL_BALANCE IS high) AND (DAYS_PAST_DUE IS high) THEN (SCORE IS bad)"

Πίνακας 6.12: Στρατηγικές ανά κατηγορία πελάτη

VERY GOOD SCORE STRATEGIES					
STRATEGY-VG1		STRATEGY-VG2		STRATEGY-VG3	
3rd day	email	3rd day	email	3rd day	email
10th day	email	10th day	email	5th day	email
25th day	email	20th day	email	10th day	email
40th day	sms	30th day	sms	20th day	sms
60th day	sms	40th day	sms	30th day	sms
80th day	call	60th day	sms	40th day	sms
		80th day	call	60th day	call
				80th day	call

GOOD SCORE STRATEGIES					
STRATEGY-G1		STRATEGY-G2		STRATEGY-G3	
5th day	email	5th day	email	3rd day	email
10th day	email	10th day	email	7th day	email
20th day	email	20th day	email	15th day	sms
40th day	sms	35th day	sms	35th day	sms
60th day	sms	55th day	sms	55th day	call
80th day	sms	70th day	sms	80th day	call
		80th day	call		

BAD SCORE STRATEGIES					
STRATEGY-B1		STRATEGY-B2		STRATEGY-B3	
3rd day	email	3rd day	email	7th day	email
10th day	email	5th day	email	15th day	sms
20th day	sms	15th day	email	30th day	call
40th day	sms	30th day	sms	60th day	call
60th day	sms	40th day	sms	80th day	call
80th day	call	50th day	call		
		60th day	call		
		80th day	call		

Πίνακας 6.15: Σύνολο επιλεγόμενων μεταβλητών για το P2P του RD1 έπειτα από Backward Selection

Σύνολο επιλεγόμενων μεταβλητών για το P2P του BD1
'DF_times_a_payment_has_been_related_with_contact'
'DF_number_of_related_accounts_with_available_mobile'
'DF_number_of_related_accounts_with_available_telephone'
'EX_LOOKUP10',
'DAYS_TILL_SETTLE_FRACTION',
'TIMES_IN_COLLECTION',
'NEXT_INSTALLMENT_AMOUNT',
'DAYS_BEFORE_NEXT_BILLING',
'DF_number_of_PTPs_given',
'DF_last_percentage_of_keeping_promise'

Πίνακας 6.19: Σύνολο επιλεγόμενων μεταβλητών για το DPD του RD2 έπειτα από Backward Selection

Σύνολο επιλεγόμενων μεταβλητών για το DPD του BD2
'DF_current_phase_id'
'DF_1_month_prelegal_count'
'DF_current_past_due_amount_bcy'
'DF_number_of_related_accounts_with_available_telephone'
'DAYS_IN_HANDLING_QUEUE'
'TIMES_IN_COLLECTION'
'NEXT_INSTALLMENT_AMOUNT'
'CASE_DAYS_ACTION_RANGE'
'DF_collector_debtor_actions',
'DF_last_contact_day_of_week'

Πίνακας 6.24.α: Συναρτήσεις συμμετοχής μεταβλητών εισόδου

TOTAL BALANCE	DAYS_PAST_DUE
$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{500 - x}{500} & ,if\ 0 \le x \le 500 \\ 0 & ,if\ x \ge 500 \end{cases}$	$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{30 - x}{30} & ,if\ 0 \le x \le 30 \\ 0 & ,if\ x \ge 30 \end{cases}$
$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 400 \\ \frac{x - 400}{800} & ,if\ 400 \le x \le 1200 \\ \frac{2000 - x}{800} & ,if\ 1200 \le x \le 2000 \\ 0 & ,if\ x \ge 2000 \end{cases}$	$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 15 \\ \frac{x - 15}{25} & ,if\ 15 \le x \le 40 \\ \frac{60 - x}{20} & ,if\ 40 \le x \le 60 \\ 0 & ,if\ x \ge 60 \end{cases}$
$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 1800 \\ \frac{x - 1800}{1000} & ,if\ 1800 \le x \le 2800 \\ 1 & ,if\ x \ge 2800 \end{cases}$	$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 50 \\ \frac{x - 50}{40} & ,if\ 50 \le x \le 90 \\ 1 & ,if\ x \ge 90 \end{cases}$

P2P_Proba	Good_Proba
$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{0.42 - x}{0.42} & ,if\ 0 \le x \le 0.42 \\ 0 & ,if\ x \ge 0.42 \end{cases}$	$\mu_L(x) = \begin{cases} 1 & ,if\ x \le 0 \\ \frac{0.42 - x}{0.42} & ,if\ 0 \le x \le 0.42 \\ 0 & ,if\ x \ge 0.42 \end{cases}$
$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 0.4 \\ \frac{x - 0.4}{0.1} & ,if\ 0.4 \le x \le 0.5 \\ \frac{0.7 - x}{0.2} & ,if\ 0.5 \le x \le 0.7 \\ 0 & ,if\ x \ge 0.7 \end{cases}$	$\mu_M(x) = \begin{cases} 0 & ,if\ x \le 0.4 \\ \frac{x - 0.4}{0.1} & ,if\ 0.4 \le x \le 0.5 \\ \frac{0.7 - x}{0.2} & ,if\ 0.5 \le x \le 0.7 \\ 0 & ,if\ x \ge 0.7 \end{cases}$
$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 0.68 \\ \frac{x - 0.68}{0.23} & ,if\ 0.68 \le x \le 0.91 \\ 1 & ,if\ x \ge 0.91 \end{cases}$	$\mu_H(x) = \begin{cases} 0 & ,if\ x \le 0.68 \\ \frac{x - 0.68}{0.23} & ,if\ 0.68 \le x \le 0.91 \\ 1 & ,if\ x \ge 0.91 \end{cases}$

Πίνακας 6.24.β: Συνάρτηση συμμετοχής μεταβλητής εξόδου

Score
$\mu_L(x) = \begin{cases} 1 & ,if\ x \leq 0 \\ \frac{33-x}{33} & ,if\ 0 \leq x \leq 33 \\ 0 & ,if\ x \geq 33 \end{cases}$
$\mu_M(x) = \begin{cases} 0 & ,if\ x \leq 20 \\ \frac{x-20}{20} & ,if\ 20 \leq x \leq 40 \\ \frac{66-x}{26} & ,if\ 40 \leq x \leq 66 \\ 0 & ,if\ x \geq 66 \end{cases}$
$\mu_H(x) = \begin{cases} 0 & ,if\ x \leq 50 \\ \frac{x-50}{16} & ,if\ 50 \leq x \leq 66 \\ \frac{100-x}{34} & ,if\ 66 \leq x \leq 100 \\ 0 & ,if\ x \geq 100 \end{cases}$

Βιβλιογραφία

[1] Abe, N., Thomas, V. P., Kowalczyk, M., Melville, P.; Pendus, C., Reddy, C. K., Jensen, D. L., Bennett, J. J., Anderson, G. F., Cooley, B. R., Domick, M., Gardinier, T. (2010). Optimizing debt collections using constrained reinforcement learning. International Conference on Knowledge Discovery and Data Mining, July 25–28, Washington, DC, USA, 75–84.

[2] A. Bellotti, D. Brigo, P. Gambetti, F. Vrins (2021). Forecasting recovery rates on non-performing loans with machine learning. Int. J. Forecast. Vol. 37, No 1, pages 428–444.

[3] Aized Amin Soofi and Arshad Awan (2017). Classification Techniques in Machine Learning: Applications and Issues. Department of Computer Science, Allama Iqbal Open University, Islamabad, Pakistan. Journal of Basic & Applied Sciences, 2017, 13, 459-465

[4] Andrew Ng, Tengyu Ma (2007). CS229 Lecture Notes. Stanford Machine learning course lecture notes, Stanford University.

[5] Anthony M., Biggs N. (1997). Computational learning theory. Cambridge University Press. Department of Statistical and Mathematical Science London School of Economics.

[6] Ali Emrouznejad and William Ho (2018). Fuzzy Analytic Hierarchy Process. CRC Press is an imprint of Taylor & Francis Group, an Informa business.

[7] Batta Mahesh (2018). Machine Learning Algorithms - A Review. International Journal of Science and Research, ISSN: 2319-7064.

[8] Catalina Sanchez, Sebastian Maldonado, Carla Vairetti (2022). Improving debt collection via contact center information: A predictive analytics framework. Decision Support Systems, Volume 159, August 2022, 113812\

[9] Chen, S. C., Huang, M. Y. (2011). Constructing credit auditing and control & management model with data mining technique. Expert Systems with Applications Vol. 38, No 5, pages 5359–5365.

[10] European Payment Report (EPR) (2023). Intrum, 26th Annual Edition https://www.intrum.com/publications/european-payment-report/european-payment-report-2023/

[11] European Payment Report Greece (EPR Greece) (2023). Intrum, 26th Annual Edition

[12] Fishburn P.C. (1970). Utility theory of Decision Making. Wiley, New York.

[13] Georgopoulos, E. F., Giannaropoulos, S. M. (2007). Solving resource management optimization problems in contact centers with artificial neural networks. 19th IEEE International Conference on Tools with Artificial Intelligence, 29–31 October 2007, Patras, Greece, 405–412

[14] Fair Debt Collection Practices Act (2014). Consumer Financial Protection Bureau, Annual Report 2014.

[15] Fair Debt Collection Practices Act (FDCPA) 15 U.S.C. § 1692 et seq (2022). Consumer Financial Protection Bureau, March 2022

[16] Fei, C. I. (2010). Evaluate the performance of cardholders' repayment behaviors using artificial neural networks and data envelopment analysis. Sixth International Conference on Networked Computing and Advanced Information Management, 16–18 August 2010, Seoul, Korea, 478–483.

[17] Hsiu-Yu Wang, Chechen Liao, Cheng-Hsiung Kao (2013). A credit assessment mechanism for wireless telecommunication debt collection: an empirical study. Springer Link. Information Systems and e-Business Management volume 11, pages357–375 (2013)

[18] Huls N. (1992). American influences on European consumer bankruptcy law, Journal of Consumer Policy 15: 125–142. https://doi.org/10.1007/BF01352132

[19] IBISworld (2023). Debt collection agencies market research report [online], [cited 10 January 2023]. Available from Internet: http://www.ibisworld.com/industry/default.aspx?indid=1474

[20] J. Kim, P. Kang (2016). Late payment prediction models for fair allocation of customer contact lists to call center agents. Decision Support Systems, volume 85, pages 84–101.

[21] Kailan Shang, Zakir Hossen (2013). Applying Fuzzy Logic to Risk Assessment and Decision-Making. Casualty Actuarial Society, Canadian Institute of Actuaries, Society of Actuaries.

[22] Keeney R., Raifa H. (1993). Decision with Multiple Objectives: Performances and Value Tradeoffs. Cambridge University Press, Cambridge.

[23] Konar Amit (2006). Computational intelligence: principles, techniques and applications. Springer Science & Business Media.

[24] Jennifer S. Raj (2019). A Comprehensive Survey on the Computational Intelligence Techniques and Its Applications. Journal of IoT in Social, Mobile, Analytics, and Cloud, Vol.01, No. 03, pages 147-159. https://doi.org/10.36548/jismac.2019.3.002

[25] Lund, S. (2010). Soft debt collection – collecting money without alienating your customer, in J. Reuvid (Ed.). The Business guide to credit management: advice and solutions for cost control, financial risk management and capital protection. Kogan Page, Ltd., 119–125

[26] Maher Maalouf (2011). Logistic regression in data analysis: an overview. Int. J. Data Analysis Techniques and Strategies, Vol. 3, No. 3.

[27] Martin Del Vecchio, Shu Jin, Alana Mistretta, Hayden Rolando, Hope Tuck (2006). Designing a Search Mechanism for Debt Collection. 2006 IEEE Systems and Information Engineering Design Symposium. Doi: 10.1109/SIEDS.2006.278733

[28] Mastorokostas P. (2015). Εισαγωγή στην ασαφή λογική – ασαφή σύνολα – συναρτήσεις συμμετοχής [Chapter]. Kallipos, Open Academic Editions. https://hdl.handle.net/11419/5958

[29] M. Bahrami, B. Bozkaya, S. Balcisoy (2020). Using behavioral analytics to predict customer invoice payment. Big Data, Vol. 8, No 1, pages 25–37.

[30] M. Mitchner, R.P. Peterson (1957). An operations-research study of the collection of defaulted loans, Oper. Res. 5 (4), 522–545.

[31] Mohamed Ahmed Elnaggar, Mostafa Abed EL Azeem and Fahima A. Maghraby (2020). Machine Learning Model for Predicting Non-performing Agricultural Loans. Springer Nature Switzerland, pages 395–404.

[32] N. Chehrazi, P.W. Glynn, T.A. Weber (2019). Dynamic credit-collections optimization. Management Science, vol. 65, No 6, pages 2737–2769.

[33] Paul Komarek (2004). Logistic Regression for Data Mining and High-Dimensional Classification. Robotics Institute, Carnegie Mellon University.

[34] Paul Komarek, Andrew W. Moore (2003). Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, PMLR R4:163-170, 2003.

[35] P. Z. Lappas, S. Z. Xanthopoulos, A. N. Yannacopoulos (2023). Metaheuristic-Based Machine Learning Approach for Customer Segmentation. Springer Nature Singapore. Metaheuristics for Machine Learning, chapter 4, pages 112-144. https://doi.org/10.1007/978-981-19-3888-7

[36] R. van der Geer, Q. Wang, S. Bhulai (2018). Data-driven consumer debt collection via machine learning and approximate dynamic programming. SSRN Electron. J. 1–32.

[37] Roy, B. (1985). Méthodologie Multicritère d'Aide à la Décision, Economica, Paris.

[38] Sezi Cevik Onar, Basar Oztaysi, & Cengiz Kahraman (2015). A Fuzzy Rule Based Inference System For Early Debt Collection. Technological and Economic Development of Economy ISSN: 2029-4913/eISSN: 2029-4921, 2018 Volume 24 Issue 5: 1845–1865. https://doi.org/10.3846/20294913.2016.1266409

[39] Vecchio, M. D., Jin, S.; Mistretta, A., Rolando, H., Tuck, H. (2006). Designing a search mechanism for debt collection. Systems and Information Engineering Design Symposium IEEE, April 26–28, Charlottesville, VA, 168–173.

[40] Vladimir Nasteski (2017). An overview of the supervised machine learning methods. Faculty of Information and Communication Technologies, Partizanska bb, Bitola. DOI:10.20544/HORIZONS.B.04.1.17.P05

[41] Takahashi, M., Tsuda, K., (2013). Towards early detections of the bad debt customers among the mail order industry, in T. Matsuo, R. C. Palacios (Eds.). Electronic business and marketing: new trends on its process and applications, 167–176.

[42] Timothy J. Ross (2010). Fuzzy Logic with Engineering Applications. University of New Mexico, USA. John Wiley & Sons. ISBN: 978-0-470-74376-8

[43] Wang, H. Y., Liao, C., Kao, C. H. (2013). A credit assessment mechanism for wireless telecommunication debt collection: an empirical study. Information Systems and e-Business Management, Vol. 11, No 3, pages 357–375.

[44] Αθανασιάδης Ηλίας (2005). Μια κοινωνιολογική και ιστορική παρουσίαση της Στατιστικής. Παιδαγωγικά ρεύματα στο Αιγαίο, Τόμ. 3, Αρ. 1,Τεύχος 3. doi:10.12681/revmata.31026

[45] Γεωργούλη Α. (2015). Τεχνητή νοημοσύνη, κεφάλαιο 4 Μηχανική Μάθηση. . Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/3382

[46] Καλπινέλλη Ευαγγελία (2004). Το Πρόβλημα των Ελλειπόντων Δεδομένων και οι νέοι Τρόποι Αντιμετώπισης του (Μεταπτυχιακή Εργασία). Οικονομικό Πανεπιστήμιο Αθηνών, Αθήνα.

[47] Καράμπέλα Αικατερίνη (2011). Ασαφής λογική και εφαρμογές στην ασφάλιση (Μεταπτυχιακή εργασία). Πανεπιστήμιο Πειραιώς, Πειραιάς.

[48] Κόκκινος Γιάννης (2011). Παράλληλοι Αλγόριθμοι Εξόρυξης Γνώσης από Βάσεις Δεδομένων με Τεχνητά Νευρωνικά Δίκτυα και Μηχανές Διανυσμάτων Υποστήριξης (Μεταπτυχιακή Εργασία). Πανεπιστήμιο Μακεδονίας, Θεσσαλονίκη.

[49] Κουδέρη Χ. (2020). Κακόβουλη Μηχανική Μάθηση (Μεταπτυχιακή εργασία). Πανεπιστήμιο Αιγαίου, Σάμος.