ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΕΧΝΙΚΕΣ ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΗΣ ΜΑΘΗΣΗΣ ΣΤΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΟΥ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ

Διπλωματική Εργασία

της

Μητσάκα Σοφίας

Θεσσαλονίκη, Ιούνιος 2022

ΤΕΧΝΙΚΕΣ ΜΗ ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΑΘΗΣΗΣ ΣΤΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΟΥ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ

Σοφία Μητσάκα

Πτυχίο Μαθηματικών, Πανεπιστήμιο Πατρών 2019

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπουσα καθηγήτρια:
Κολωνιάρη Γεωργία

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28/06/2022:

Κ. Κολωνιάρη Γεωργία	Κ. Δασίλας Απόστολος	Κ. Ευαγγελίδης Γεώργιος
.....	.....	.....

Μητσάκα Σοφία

.....

Περίληψη

Η διαχείριση του πιστωτικού κινδύνου είναι ένα σημαντικό μέρος των δραστηριοτήτων των τραπεζών όπως και κάθε χρηματοπιστωτικού ιδρύματος. Σύμφωνα με τον επίσημο ορισμό, πιστωτικός κίνδυνος είναι ο κίνδυνος που προκύπτει από την μη ικανότητα των οφειλετών ενός χαρτοφυλακίου μιας τράπεζας ή ενός χρηματοπιστωτικού ιδρύματος να αποπληρώνουν τα δάνειά τους σε ορισμένο χρονικό διάστημα. Ένας τρόπος εκτίμησης/ποσοτικοποίησης του κινδύνου αυτού είναι μέσω της μοντελοποίησης της πιθανότητας αθέτησης (probability of default) των οφειλετών. Η συχνότερη και επικρατέστερη προσέγγιση είναι η μοντελοποίηση αυτή να γίνεται μέσω του στατιστικού μοντέλου λογιστικής παλινδρόμησης (logistic regression), χρησιμοποιώντας την πιστωτική ικανότητα (credit score) σε κάθε οφειλέτη, από την οποία και εξαρτάται η πιθανότητα αθέτησης. Στην εργασία αυτή θα μελετηθεί ένας εναλλακτικός τρόπος μοντελοποίησης της πιθανότητας αθέτησης μέσω τεχνικών μάθησης χωρίς επίβλεψη, και πιο συγκεκριμένα μέσω των αλγορίθμων Ανάλυσης Κύριων Συνιστωσών (Principal Components Analysis, PCA), την Multiple Correspondence Analysis (M.C.A.) και την Factor Analysis of Mixed Data (F.A.M.D) με σκοπό την μείωση των μεταβλητών του μοντέλου, και των αλγορίθμων ιεραρχικής συσταδοποίησης (hierarchical clustering) και k-Means. Στόχος της έρευνάς μας είναι να γίνει μια βαθύτερη ανάλυση των παραπάνω αλγορίθμων και του τρόπου εφαρμογής τους στην διαχείριση του πιστωτικού κινδύνου. Μέρος της εργασίας είναι και η υλοποίηση της εφαρμογής των προαναφερθέντων τεχνικών σε ένα επιλεγμένο δημόσια διαθέσιμο σύνολο δεδομένων με τη χρήση της γλώσσας προγραμματισμού Python.

Λέξεις Κλειδιά: τραπεζική, πιστωτικός κίνδυνος, μη εποπτευόμενη μάθηση, μέθοδοι μείωσης διαστάσεων, συσταδοποίηση, k-means, hierarchical, python

Abstract

Credit risk management is an important part of the activities of all banks as well as any financial institution. According to the official definition, credit risk is the risk arising from the inability of the borrowers of a to repay their loans within a certain period of time. One way to assess / quantify this risk is through modeling the probability of default of debtors. The most common and prevalent approach to perform this modeling is through the statistical model of logistic regression, using the credit score of each debtor, The probability of default depends on this score. This research will study an alternative way of modeling the probability of default through unsupervised learning techniques, and more specifically through Principal Components Analysis (PCA), Mulitple Correspondence Analysis (MCA) and Factor Analysis of Mixed Data (FAMD) algorithms, in order to reduce the dataset's dimensionality, and then through the implementation of hierarchical and k-Means clustering algorithms. The aim of the research is to make a deeper analysis of the above algorithms and how they could possibly be applied in credit risk management. Part of the work is to implement the aforementioned techniques in a selected publicly available benchmark data set using Python programming language.

Keywords: banking, credit risk, unsupervised learning techniques, dimensionality reduction, clustering, kmeans, hierarchical, python

Πρόλογος – Ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτριά μου κυρία Κολωνιάρη Γεωργία για την πολύτιμη καθοδήγηση και υποστήριξή της κατά τη διάρκεια της εκπόνησης της Μεταπτυχιακής διπλωματικής μου εργασίας.

Επίσης, θέλω να ευχαριστήσω την οικογένεια μου, τους φίλους και τα κοντινά μου πρόσωπα για την υποστήριξή τους όλο αυτό το διάστημα της φοίτησης μου στο πρόγραμμα μεταπτυχιακών σπουδών της Εφαρμοσμένης Πληροφορικής του Πανεπιστημίου Μακεδονίας.

Περιεχόμενα

Κατάλογος Εικόνων

1 Εισαγωγή

1.1 Πρόβλημα – Σημαντικότητα του θέματος

Η διαχείριση του πιστωτικού κινδύνου είναι μια από τις σημαντικότερες δραστηριότητες των τραπεζών και των χρηματοπιστωτικών ιδρυμάτων. Σύμφωνα με τον επίσημο ορισμό, πιστωτικός κίνδυνος είναι ο κίνδυνος που προκύπτει από την μη ικανότητα των οφειλετών ενός χαρτοφυλακίου μιας τράπεζας/χρηματοπιστωτικού ιδρύματος να αποπληρώνουν τα δάνειά τους σε ορισμένο χρονικό διάστημα. Η επιτυχής διαχείριση οδηγεί σε ορθή λήψη αποφάσεων και κατά συνέπεια σε κέρδη. Πρόκειται λοιπόν για ένα ζήτημα υψηλής σημασίας για τη χρηματοοικονομική δραστηριότητα των τραπεζών και χρηματοπιστωτικών ιδρυμάτων αλλά και συνολικά του κόσμου.

1.2 Σκοπός – Στόχοι

Ένας τρόπος εκτίμησης/ποσοτικοποίησης του κινδύνου αυτού είναι μέσω της μοντελοποίησης της πιθανότητας αθέτησης (probability of default) των οφειλετών. Η συχνότερη και επικρατέστερη προσέγγιση είναι η μοντελοποίηση αυτή να γίνεται μέσω του στατιστικού μοντέλου λογιστικής παλινδρόμησης (logistic regression). Στην εργασία αυτή θα μελετηθεί ένας εναλλακτικός τρόπος μοντελοποίησης της πιθανότητας αθέτησης μέσω τεχνικών μάθησης χωρίς επίβλεψη, και πιο συγκεκριμένα μέσω της εφαρμογής μεθόδων για την μείωση των μεταβλητών του μοντέλου, και των αλγορίθμων ιεραρχικής συσταδοποίησης (hierarchical clustering) και k-Means. Στόχος της έρευνας είναι να γίνει μια βαθύτερη ανάλυση των παραπάνω αλγορίθμων και του τρόπου εφαρμογής τους στην διαχείριση του πιστωτικού κινδύνου. Μέρος της εργασίας θα είναι και η υλοποίηση της εφαρμογής των προαναφερθέντων τεχνικών σε ένα επιλεγμένο δημόσια διαθέσιμο σύνολο δεδομένων σχετικά με την πληρωμή ή αθέτηση πληρωμών πιστωτικών καρτών στην Ταϊβάν, με τη χρήση της γλώσσας προγραμματισμού Python. Σκοπός είναι να γίνει σύγκριση των αποτελεσμάτων και να εξαχθούν συμπεράσματα για την απόδοση και καταλληλόλητα των προαναφερθέντων τεχνικών.

1.3 Διάρθρωση της μελέτης

Η διάρθρωση της παρούσας μελέτης αποτελείται από οκτώ κεφάλαια. Το πρώτο κεφάλαιο αποτελεί το παρόν κεφάλαιο, δηλαδή την εισαγωγή της εργασίας που περιγράφεται ο η σημαντικότητα του θέματος και ο σκοπός της εργασίας. Στο δεύτερο

κεφάλαιο αναλύονται οι εισαγωγικές έννοιες των τραπεζικών συστημάτων, στο τρίτο κεφάλαιο περιγράφεται το πρόβλημα της πιθανότητας αθέτησης των οφειλετών ενός χαρτοφυλακίου, στο τρίτο κεφάλαιο πραγματοποιείται μια εισαγωγή στην μηχανική μάθηση και στο ρόλο της στο τραπεζικό σύστημα, ενώ στο τέταρτο κεφάλαιο περιγράφονται οι μέθοδοι μη επιβλεπόμενης μάθησης που θα εφαρμοστούν στην υλοποίηση. Στο πέμπτο κεφάλαιο παρουσιάζεται μια ανασκόπηση της υπάρχουσας βιβλιογραφίας σχετικά με την μοντελοποίηση του πιστωτικού κινδύνου μέσω μεθόδων μηχανικής μάθησης, ενώ στο έκτο κεφάλαιο παρουσιάζεται η εμπειρική μελέτη της εργασίας. Τέλος, η εργασία κλείνει με το έβδομο κεφάλαιο που αποτελεί τα συμπεράσματα της μελέτης.

2 Εισαγωγικές Έννοιες σε Τραπεζικά Θέματα

2.1 Εισαγωγικές έννοιες στην τραπεζική και τον πιστωτικό κίνδυνο

Η τραπεζική ορίζεται ως η επιχειρηματική δραστηριότητα της αποδοχής και της διαφύλαξης χρημάτων που ανήκουν σε άτομα και οντότητες και στη συνέχεια ο δανεισμός αυτών των χρημάτων προκειμένου να διεξάγονται οικονομικές δραστηριότητες όπως η επίτευξη κέρδους ή απλώς η κάλυψη λειτουργικών εξόδων. Η τράπεζα είναι ένα χρηματοπιστωτικό ίδρυμα που έχει άδεια να λαμβάνει καταθέσεις και να χορηγεί δάνεια. Δύο από τους πιο συνηθισμένους τύπους τραπεζών είναι οι εμπορικές/λιανικές τράπεζες και οι επενδυτικές τράπεζες. Ανάλογα με τον τύπο, μια τράπεζα μπορεί επίσης να παρέχει διάφορες χρηματοοικονομικές υπηρεσίες που κυμαίνονται από την παροχή θυρίδων και ανταλλαγής συναλλάγματος έως τη συνταξιοδότηση και τη διαχείριση περιουσίας.

Οι κεντρικές τράπεζες κάθε χώρας, είναι υπεύθυνες για τη σταθερότητα του νομίσματος. Ελέγχουν τον πληθωρισμό, υπαγορεύουν/θέτουν νομισματικές πολιτικές και επιβλέπουν τη ζήτηση και την προσφορά χρήματος στην αγορά. Οι εμπορικές ή λιανικές τράπεζες προσφέρουν διάφορες υπηρεσίες, όπως, ενδεικτικά, διαχείριση καταθέσεων και αναλήψεων χρημάτων, παροχή βασικών λογαριασμών ελέγχου και ταμιευτηρίου, πιστοποιητικών καταθέσεων, έκδοση χρεωστικών και πιστωτικών καρτών σε ειδικευμένους πελάτες, παροχή βραχυπρόθεσμων και μακροπρόθεσμων δανείων όπως δάνεια αυτοκινήτου, στεγαστικά δάνεια ή πιστωτική γραμμή μετοχικού κεφαλαίου. Οι τράπεζες επενδύσεων προσανατολίζουν τις υπηρεσίες τους προς εταιρικούς πελάτες. Παρέχουν υπηρεσίες όπως συγχωνεύσεις και εξαγορές και αναδοχές μεταξύ άλλων επενδυτικών υπηρεσιών.

Δεδομένης της δραστηριότητας των τραπεζών, είναι επόμενο ότι η λειτουργία τους συνεπάγεται την ανάληψη ρίσκου/κινδύνου. Τα χρηματοπιστωτικά ιδρύματα που λειτουργούν αποφεύγοντας όλους τους κινδύνους θα είναι στάσιμα και δεν θα εξυπηρετούν επαρκώς τις πιστωτικές ανάγκες της κοινότητας, ενώ από την άλλη, μια τράπεζα που παίρνει υπερβολικά ρίσκα είναι πιθανόν να αντιμετωπίσει δυσκολίες. Η διαχείριση κινδύνου περιλαμβάνει τον εντοπισμό, την ανάλυση και την απόκριση σε παράγοντες κινδύνου που αποτελούν μέρος της ζωής μιας επιχείρησης. Αποτελεσματική

διαχείριση κινδύνου σημαίνει προσπάθεια ελέγχου, όσο το δυνατόν περισσότερο, των μελλοντικών αποτελεσμάτων ενεργώντας προληπτικά και όχι αντιδραστικά. Ως εκ τούτου, η αποτελεσματική διαχείριση κινδύνου προσφέρει τη δυνατότητα μείωσης τόσο της πιθανότητας εμφάνισης ενός κινδύνου όσο και της πιθανής επίδρασής του. Όταν μια επιχείρηση αξιολογεί το σχέδιό της για τον χειρισμό πιθανών απειλών και στη συνέχεια αναπτύσσει δομές για την αντιμετώπισή τους, βελτιώνει τις πιθανότητες να γίνει επιτυχημένη οντότητα.

Λόγω του μεγάλου μεγέθους ορισμένων τραπεζών, η υπερβολική έκθεση σε κίνδυνο μπορεί να προκαλέσει χρεοκοπία τραπεζών και να επηρεάσει εκατομμύρια ανθρώπους. Κατανοώντας τους κινδύνους που ενέχουν οι τράπεζες, οι κυβερνήσεις μπορούν να θεσπίσουν καλύτερους κανονισμούς για να ενθαρρύνουν τη συνετή διαχείριση και λήψη αποφάσεων. Η ικανότητα μιας τράπεζας να διαχειρίζεται τον κίνδυνο επηρεάζει επίσης τις αποφάσεις των επενδυτών. Ακόμα κι αν μια τράπεζα μπορεί να δημιουργήσει μεγάλα έσοδα, η έλλειψη διαχείρισης κινδύνου μπορεί να μειώσει τα κέρδη λόγω ζημιών από δάνεια. Οι επενδυτές είναι πιο πιθανό να επενδύσουν σε μια τράπεζα που είναι σε θέση να παρέχει κέρδη και δεν διατρέχει υπερβολικό κίνδυνο να χάσει χρήματα. Οι κύριοι κίνδυνοι που αντιμετωπίζουν οι τράπεζες περιλαμβάνουν πιστωτικούς, λειτουργικούς κινδύνους, κινδύνους αγοράς και ρευστότητας.

- Πιστωτικός κίνδυνος

Ο πιστωτικός κίνδυνος είναι ο μεγαλύτερος κίνδυνος για τις τράπεζες. Συμβαίνει όταν οι δανειολήπτες ή οι αντισυμβαλλόμενοι αδυνατούν να εκπληρώσουν τις συμβατικές τους υποχρεώσεις. Ένα παράδειγμα είναι όταν οι δανειολήπτες δεν πληρώνουν το κεφάλαιο ή τους τόκους ενός δανείου. Οι αθετήσεις μπορεί να προκύψουν σε στεγαστικά δάνεια, πιστωτικές κάρτες και τίτλους σταθερού εισοδήματος. Η μη τήρηση των υποχρεωτικών συμβάσεων μπορεί επίσης να συμβεί σε τομείς όπως τα παράγωγα και οι παρεχόμενες εγγυήσεις.

Ενώ οι τράπεζες δεν μπορούν να προστατευθούν πλήρως από τον πιστωτικό κίνδυνο λόγω της φύσης του επιχειρηματικού τους μοντέλου, μπορούν να μειώσουν την έκθεσή τους με διάφορους τρόπους. Δεδομένου ότι η επιδείνωση σε έναν κλάδο ή έναν εκδότη

είναι συχνά απρόβλεπτη, οι τράπεζες μειώνουν την έκθεσή τους μέσω της διαφοροποίησης (diversification).

Με αυτόν τον τρόπο, κατά τη διάρκεια μιας πιστωτικής ύφεσης, οι τράπεζες είναι λιγότερο πιθανό να υπερεκτεθούν σε μια κατηγορία με μεγάλες ζημίες. Για να μειώσουν την έκθεσή τους στον κίνδυνο, μπορούν να δανείσουν χρήματα σε άτομα με καλό πιστωτικό ιστορικό, να συναλλάσσονται με αντισυμβαλλομένους υψηλής ποιότητας ή να έχουν εξασφαλίσεις για τη δημιουργία αντιγράφων ασφαλείας των δανείων.

- Λειτουργικός κίνδυνος

Λειτουργικός κίνδυνος είναι ο κίνδυνος απώλειας λόγω σφαλμάτων, διακοπών ή ζημιών που προκαλούνται από άτομα, συστήματα ή διαδικασίες. Ο λειτουργικός τύπος κινδύνου είναι χαμηλός για απλές επιχειρηματικές δραστηριότητες, όπως η λιανική τραπεζική και η διαχείριση περιουσιακών στοιχείων, και υψηλότερος για δραστηριότητες όπως οι πωλήσεις και οι συναλλαγές. Οι απώλειες που προκύπτουν λόγω ανθρώπινου λάθους περιλαμβάνουν εσωτερική απάτη ή λάθη που έγιναν κατά τη διάρκεια των συναλλαγών. Ένα παράδειγμα είναι όταν ένας ταμίας δίνει κατά λάθος έναν επιπλέον λογαριασμό 50 $ σε έναν πελάτη.

Σε μεγαλύτερη κλίμακα, η απάτη μπορεί να συμβεί μέσω παραβίασης της κυβερνοασφάλειας μιας τράπεζας. Επιτρέπει στους χάκερ να κλέβουν πληροφορίες πελατών και χρήματα από την τράπεζα και να εκβιάζουν τα ιδρύματα για επιπλέον χρήματα. Σε μια τέτοια κατάσταση, οι τράπεζες χάνουν κεφάλαια και εμπιστοσύνη από τους πελάτες. Η ζημιά στη φήμη της τράπεζας μπορεί να καταστήσει πιο δύσκολη την προσέλκυση καταθέσεων ή επιχειρήσεων στο μέλλον.

- Κίνδυνος αγοράς

Ο κίνδυνος αγοράς προκύπτει κυρίως από τις δραστηριότητες μιας τράπεζας στις κεφαλαιαγορές. Οφείλεται στο απρόβλεπτο των αγορών μετοχών, των τιμών των εμπορευμάτων, των επιτοκίων και των πιστωτικών περιθωρίων. Οι τράπεζες είναι πιο εκτεθειμένες εάν συμμετέχουν σε μεγάλο βαθμό σε επενδύσεις σε κεφαλαιαγορές ή πωλήσεις και συναλλαγές.

Οι τιμές των εμπορευμάτων παίζουν επίσης ρόλο, επειδή μια τράπεζα μπορεί να επενδύσει σε εταιρείες που παράγουν εμπορεύματα. Καθώς η αξία του εμπορεύματος αλλάζει, αλλάζει και η αξία της εταιρείας και η αξία της επένδυσης. Οι αλλαγές στις τιμές των εμπορευμάτων προκαλούνται από μετατοπίσεις της προσφοράς και της ζήτησης που συχνά είναι δύσκολο να προβλεφθούν. Επομένως, για τη μείωση του κινδύνου αγοράς, η διαφοροποίηση των επενδύσεων είναι σημαντική. Άλλοι τρόποι με τους οποίους οι τράπεζες μειώνουν τις επενδύσεις τους περιλαμβάνουν την αντιστάθμιση των επενδύσεών τους με άλλες, αντιστρόφως σχετικές επενδύσεις.

- Κίνδυνος ρευστότητας

Ο κίνδυνος ρευστότητας αναφέρεται στην ικανότητα μιας τράπεζας να έχει πρόσβαση σε μετρητά για να ανταποκριθεί στις υποχρεώσεις χρηματοδότησης. Στις υποχρεώσεις περιλαμβάνεται η δυνατότητα στους πελάτες να λάβουν τις καταθέσεις τους. Η αδυναμία έγκαιρης παροχής μετρητών στους πελάτες μπορεί να οδηγήσει σε ένα φαινόμενο χιονοστιβάδας. Εάν μια τράπεζα καθυστερήσει να παράσχει μετρητά για μερικούς από τους πελάτες της για μια ημέρα, άλλοι καταθέτες μπορεί να βιαστούν να αφαιρέσουν τις καταθέσεις τους καθώς χάνουν την εμπιστοσύνη τους στην τράπεζα. Αυτό μειώνει περαιτέρω την ικανότητα της τράπεζας να παρέχει κεφάλαια και οδηγεί σε bank run.

Οι λόγοι για τους οποίους οι τράπεζες αντιμετωπίζουν προβλήματα ρευστότητας περιλαμβάνουν την υπερβολική εξάρτηση από βραχυπρόθεσμες πηγές κεφαλαίων, τον ισολογισμό επικεντρωμένο σε μη ρευστοποιήσιμα στοιχεία ενεργητικού και την απώλεια εμπιστοσύνης προς την τράπεζα από την πλευρά των πελατών. Η κακή διαχείριση της διάρκειας του ενεργητικού-παθητικού μπορεί επίσης να προκαλέσει δυσκολίες χρηματοδότησης. Αυτό συμβαίνει όταν μια τράπεζα έχει πολλές βραχυπρόθεσμες υποχρεώσεις και όχι αρκετά βραχυπρόθεσμα περιουσιακά στοιχεία.

Οι βραχυπρόθεσμες υποχρεώσεις είναι καταθέσεις πελατών ή βραχυπρόθεσμα εγγυημένα επενδυτικά συμβόλαια (GIC) που η τράπεζα πρέπει να πληρώσει στους πελάτες. Εάν το σύνολο ή το μεγαλύτερο μέρος των περιουσιακών στοιχείων μιας τράπεζας είναι συνδεδεμένα με μακροπρόθεσμα δάνεια ή επενδύσεις, η τράπεζα μπορεί να αντιμετωπίσει αναντιστοιχία στη διάρκεια του ενεργητικού-παθητικού.

Υπάρχουν κανονισμοί για τη μείωση των προβλημάτων ρευστότητας. Περιλαμβάνουν την απαίτηση για τις τράπεζες να διατηρούν αρκετά ρευστά περιουσιακά στοιχεία για να επιβιώσουν για ένα χρονικό διάστημα ακόμη και χωρίς εισροή εξωτερικών κεφαλαίων.

Συμπερασματικά, η ανάλυση κινδύνου είναι μια ποιοτική προσέγγιση επίλυσης προβλημάτων που χρησιμοποιεί διάφορα εργαλεία για την επεξεργασία και την ταξινόμηση των κινδύνων με σκοπό την αξιολόγηση και την επίλυσή τους.

2.2 Ο πιστωτικός κίνδυνος και το πρόβλημα της πιθανότητας αθέτησης των δανειοληπτών

Ο πιστωτικός κίνδυνος μπορεί, εν συντομία, να οριστεί ως «η πιθανότητα ένα συμβαλλόμενο πρόσωπο να μην εκπληρώσει τις υποχρεώσεις του σύμφωνα με τους συμφωνημένους όρους». Ο πιστωτικός κίνδυνος αναφέρεται, επίσης ως κίνδυνος αθέτησης, κίνδυνος απόδοσης ή κίνδυνος αντισυμβαλλομένου. Υπάρχουν τρία χαρακτηριστικά που καθορίζουν τον πιστωτικό κίνδυνο:

1. Η έκθεση (σε συμβαλλόμενο πρόσωπο που ενδέχεται να χρεοκοπήσει ή να υποστεί δυσμενή αλλαγή στην ικανότητά του να αποδώσει).

2. Η πιθανότητα αυτό το συμβαλλόμενο πρόσωπο να αθετήσει τις υποχρεώσεις του (η πιθανότητα αθέτησης).

3. Το ποσοστό ανάκτησης (δηλαδή, το ποσό που μπορεί να ανακτηθεί εάν πραγματοποιηθεί χρεοκοπία).

Αξίζει να σημειωθεί ότι όσο μεγαλύτερα είναι τα δύο πρώτα στοιχεία, τόσο μικρότερο είναι το ποσοστό ανάκτησης και συνεπώς μεγαλύτερος ο κίνδυνος, ενώ, από την άλλη πλευρά, όσο υψηλότερο είναι το ποσό που μπορεί να ανακτηθεί, τόσο χαμηλότερος είναι ο κίνδυνος. Ο κίνδυνος μπορεί να εκφραστεί ως εξής:

Πιστωτικός κίνδυνος = Έκθεση x Πιθανότητα χρεοκοπίας x (1 – Ποσοστό ανάκτησης)

Με βάση τα παραπάνω, η διαχείριση πιστωτικού κινδύνου είναι η ουσιαστικά η διαδικασία ελέγχου των πιθανών συνεπειών του πιστωτικού κινδύνου.

2.3 Η απόφαση πίστωσης

Η αξιολόγηση του πιστωτικού κινδύνου απαιτεί την μοντελοποίηση της πιθανότητας αθέτησης των υποχρεώσεων από τη μεριά του δανειολήπτη, πλήρως ή εν μέρει. Η πιστωτική απόφαση μπορεί να απεικονισθεί με βάση το βασικό μοντέλο διαχείρισης κινδύνου. Αυτό περιλαμβάνει μια απόφαση είτε (Α) για αποδοχή της πίστωσης, η οποία παρέχει ανταμοιβή αλλά συνεπάγεται κίνδυνο, είτε (Β) για άρνηση πίστωσης. Η κατάσταση που αντιμετωπίζει η τράπεζα/το χρηματοπιστωτικό ίδρυμα παρουσιάζεται ως πρόβλημα απόφασης. Η απαίτηση είναι να εξισορροπηθεί το κέρδος από την ανάληψη του πιστωτικού κινδύνου/ρίσκου με την αποδοχή της πίστωσης έναντι της πιθανής ζημίας. Στο πρόβλημα της απόφασης, η εναλλακτική είναι να αρνηθεί κανείς την πίστωση και συνεπώς να μην λάβει καμία πιθανή ανταμοιβή.

Στην περίπτωση αποδοχής της πίστωσης, υπάρχουν δύο πιθανά σενάρια: η πίστωση λειτουργεί σύμφωνα με τις προσδοκίες ή υπάρχουν αδυναμίες/δυσκολίες στην αποπλήρωση. Εάν ο δανειολήπτη αθετήσει την πληρωμή, το κόστος για την τράπεζα/χρηματοπιστωτικό ίδρυμα θα είναι το κόστος ή η αξία αντικατάστασης για ό,τι δεν έχει αποπληρωθεί. Για παράδειγμα, όταν αποφασίζει εάν θα παράσχει εμπορική πίστωση, μια επιχείρηση αντιμετωπίζει μια απόφαση σχετικά με τις αιτήσεις που θα προχωρήσει· ποιο όριο πρέπει να τεθεί στο ποσό της πίστωσης που παρατείνεται και εάν αυτό πρέπει να τροποποιηθεί με την πάροδο του χρόνου και τι μέτρα πρέπει να ληφθούν εάν υπάρχει καθυστέρηση στην αποπληρωμή.

Αν και η φύση της απόφασης της πιστωτικής ανάλυσης μπορεί να περιγραφεί εύκολα, τα βήματα που απαιτούνται για την αποτελεσματική διαχείριση της διαδικασίας είναι πιο περίπλοκα. Ουσιαστικά, το πρόβλημα σχετίζεται με τον κίνδυνο οι αντισυμβαλλόμενοι να μην τηρήσουν τις υποχρεώσεις τους όταν έρθει η στιγμή να εκπληρώσουν το συμβόλαιό τους. Ο προσδιορισμός του ποιος αντισυμβαλλόμενος μπορεί να αθετήσει είναι η τέχνη της διαχείρισης πιστωτικού κινδύνου. Διαφορετικές προσεγγίσεις χρησιμοποιούν μοντέλα κρίσης, ντετερμινιστικά ή σχέσεων, ή κάνουν χρήση στατιστικών μοντέλων, προκειμένου να ταξινομήσουν την πιστωτική ποιότητα και να

προβλέψουν την πιθανή συχνότητα αθέτησης υποχρεώσεων. Μόλις ολοκληρωθεί η διαδικασία αξιολόγησης της πιστοληπτικής ικανότητας, μπορεί να προσδιοριστεί το ύψος του κινδύνου που πρέπει να αναληφθεί. Κατά την εφαρμογή του μοντέλου απόφασης, η επίδραση της πιστωτικής έκθεσης μετριέται από το κόστος αντικατάστασης των ταμειακών ροών, εάν το άλλο μέρος αθετήσει την υποχρέωση. Όμως οι απώλειες δεν προκύπτουν μόνο από αδυναμία πληρωμής. Απώλειες προκύπτουν, επίσης, από πιστωτικό κίνδυνο όταν οι οργανισμοί αξιολόγησης πιστοληπτικής ικανότητας (βλ. κεφάλαιο 2.5) υποβαθμίζουν εταιρείες ή επιχειρήσεις. Όπου αυτές οι υποχρεώσεις είναι διαπραγματεύσιμες κινητές αξίες, υπάρχει συχνά μείωση της αγοραίας αξίας τους. Επιπλέον, όταν η συναλλαγή είναι διασυνοριακή, ο κίνδυνος χώρας πρέπει να περιλαμβάνεται στην αξιολόγηση κινδύνου.

2.4 Αξιολόγηση της απόφασης πίστωσης

Σε αυτό το σημείο θα πρέπει να σημειωθεί ότι η σύγχρονη χρηματοοικονομική θεωρία προτείνει ότι η απόρριψη της πίστωσης δεν είναι απαραίτητα η κατάλληλη απάντηση στην κακή πιστωτική ποιότητα. Οι αρχές της σύγχρονης θεωρίας σχετικά με τον κίνδυνο υποθέτουν ότι η απαιτούμενη απόδοση πρέπει να προσαρμοστεί για τον κίνδυνο που αναλαμβάνεται. Εάν ο κίνδυνος έχει εκτιμηθεί σωστά, τότε, για μεγάλους οργανισμούς μεσοπρόθεσμα και όπου ισχύουν αποτελέσματα διαφοροποίησης χαρτοφυλακίου, οι ζημίες θα αντισταθμιστούν από κέρδη αλλού. Η χρηματοοικονομική θεωρία θα πρότεινε, επιπλέον, ότι μόνο η συνιστώσα του συστηματικού κινδύνου χρειάζεται να τιμολογηθεί. Ένας στόχος της διαδικασίας πιστωτικής μοντελοποίησης θα πρέπει να είναι η παροχή εκτιμήσεων για τον πιθανό κίνδυνο. Στη συνέχεια, μπορεί να ληφθεί η απόφαση για το εάν θα παρασχεθεί ένα πιστωτικό όριο σε μια κατάλληλη τιμή προσαρμοσμένη στον κίνδυνο για την αντιστάθμιση του κινδύνου ή για την εξεύρεση τρόπων για τη μείωση του βαθμού έκθεσης, αλλά και τη συνέχιση της συναλλαγής. Αυτή είναι μια πιο περίπλοκη προσέγγιση από αυτό που χρησιμοποιείται στους περισσότερους οργανισμούς, οι οποίοι τείνουν να υιοθετούν μια άποψη «ναι» ή «όχι» για την επέκταση της πίστωσης και επίσης επιδιώκουν να ελέγξουν την έκθεσή τους μέσω ορίων στα ποσά που διατρέχουν κίνδυνο.

2.5 Ταξινόμηση κινδύνου, ταξινόμηση πιστωτικής ποιότητας, και πρόβλεψη αθέτησης

Για την μοντελοποίηση του πιστωτικού κινδύνου, συχνά χρησιμοποιείται η πιστωτική ικανότητα (credit score) του κάθε δανειολήπτη, από την οποία και εξαρτάται η πιθανότητα αθέτησης. Αυτό αφορά κυρίως επιχειρήσεις που αιτούνται ένα δάνειο. Η αξιολόγηση της πιστοληπτικής ικανότητας είναι ουσιαστικά η ταξινόμηση μιας συγκεκριμένης επιχείρησης σε μια δεδομένη πιστωτική κατηγορία. Οι διάφοροι εμπορικοί αξιολογητές πιστοληπτικής ικανότητας και εταιρείες που χρησιμοποιούν τις δικές τους αξιολογήσεις πιστοληπτικής ικανότητας χρησιμοποιούν διαφορετικά συστήματα αξιολόγησης. Το σύστημα αξιολόγησης που χρησιμοποιείται από οίκους αξιολόγησης όπως η Standard & Poor's, η Moodys και η Fitch έχει τέσσερις κατηγορίες επενδυτικής πιστωτικής ποιότητας και τρεις κατηγορίες κερδοσκοπικής πιστωτικής ποιότητας. Η πρόθεση είναι να ομαδοποιηθούν οι αξιολογήσεις/υποθέσεις με συνεπή τρόπο, έτσι ώστε, για σκοπούς λήψης αποφάσεων, όλες οι εταιρείες ενός συγκεκριμένου ομίλου να αντιμετωπίζονται ως ισοδύναμες. Δεδομένου ότι όλες οι εταιρείες μιας συγκεκριμένης κατηγορίας μπορούν να θεωρηθούν ότι έχουν τον ίδιο βαθμό πιστοληπτικής ικανότητας, η αντιμετώπιση που ακολουθείται για κάθε κατηγορία μπορεί να εφαρμοστεί σε κάθε νέα περίπτωση πίστωσης που αναλύεται.

Σε ορισμένες περιπτώσεις, χρησιμοποιείται τόσο μια επίσημη ποσοτική προσέγγιση όσο και μια πιο κρίσιμη ποιοτική προσέγγιση προκειμένου να προσδιοριστεί η πιστωτική της κατηγορία. Με αυτόν τον τρόπο, η επιχείρηση Κ συγκρίνεται με παρόμοιες εταιρείες των οποίων η πιστοληπτική ποιότητα έχει ήδη προσδιοριστεί, και ως εκ τούτου η εταιρεία θεωρείται ότι είναι σαν ένα συγκεκριμένο είδος πίστωσης για σκοπούς πρόβλεψης αθέτησης πληρωμών. Δεδομένων των διαφορετικών ποσοστών αθέτησης υποχρεώσεων και τύπων εταιρειών, ο αριθμός των κατηγοριών πιστωτικής ποιότητας μπορεί να είναι μεγαλύτερος ή μικρότερος, ανάλογα με την ευαισθησία του μοντέλου. Για παράδειγμα, πολλές τράπεζες χρησιμοποιούν μια κλίμακα 10 πιστωτικών κατηγοριών, με την υψηλότερη πιστωτική ποιότητα να είναι αυτή της κατάστασης της χώρας ίδρυσης και η χαμηλότερη να είναι η χρεοκοπία.

3 Μηχανική Μάθηση και Τεχνητή Νοημοσύνη και ο ρόλος τους στον χρηματοοικονομικό τομέα

Αν και η τεχνητή νοημοσύνη και η μηχανική μάθηση χρησιμοποιούνται μερικές φορές εναλλακτικά, σημαίνουν δύο διαφορετικά πράγματα. Η τεχνητή νοημοσύνη είναι ουσιαστικά η διαδικασία ανάπτυξης ευφυούς λογισμικού και συστημάτων υπολογιστών που μιμούνται τους ανθρώπους, μελετώντας πώς σκέφτονται οι άνθρωποι, πώς μαθαίνουν και τη νοητική τους ικανότητα στην επίλυση ενός προβλήματος. Με άλλα λόγια, η τεχνητή νοημοσύνη δημιουργεί δείκτη νοημοσύνης (intelligent quotient-I.Q.) και δείκτη συναισθηματικής νοημοσύνης (emotional quotient-E.Q.) σε υπολογιστές. Η μηχανική μάθηση από την άλλη είναι η υλοποίηση μεθόδων που 'μαθαίνουν', αξιοποιούν δηλαδή δεδομένα και πληροφορία για την βελτίωση της απόδοσης σε κάποια διαδικασία. Οι αλγόριθμοι μηχανικής μάθησης χτίζουν ένα μοντέλο που βασίζεται σε δείγματα δεδομένων, γνωστά ως δεδομένα εκπαίδευσης, προκειμένου να κάνουν προβλέψεις ή αποφάσεις χωρίς να είναι ρητά προγραμματισμένοι να το κάνουν.

Οι εφαρμογές μηχανικής μάθησης και τεχνητής νοημοσύνης στον χρηματοοικονομικό τομέα έχουν ακμάσει τα τελευταία χρόνια. Η τεράστια δύναμή τους έχει αξιοποιηθεί σε αυτά στα χρηματοοικονομικά ιδρύματα για να προσφέρουν επιχειρηματικές λύσεις σε διεργασίες front end και back end για τη δημιουργία αποτελεσματικότητας και τη βελτίωση της εμπειρίας των πελατών. Στον τραπεζικό κλάδο οι τεχνολογίες αυτές χρησιμοποιούνται κυρίως για αυτοματοποίηση, ανάλυση και λήψη αποφάσεων, δημιουργώντας έτσι νέα επιχειρηματικά μοντέλα. Τον τελευταίο καιρό, η υπολογιστική νοημοσύνη είναι πολύτιμος παράγοντας για την επίτευξη ανταγωνιστικού πλεονεκτήματος αξιοποιώντας τις ικανότητές της και στη λήψη αποφάσεων. Ως εκ τούτου, οι χρηματοοικονομικές και τραπεζικές υπηρεσίες στον κόσμο έχουν περάσει από μια 'αλλαγή εποχης', χάρη σε αυτές τις δύο τεχνολογίες. Η ανάπτυξη των οργανισμών fintech διαδραματίζει πρωταγωνιστικό ρόλο στη 'μεταμόρφωση' που εκτυλίσσεται. Οι εταιρείες fintech, οι οποίες φαίνεται να έχουν ενσωματώσει την τεχνητή νοημοσύνη εδώ και πολύ καιρό και διαδραματίζουν κρίσιμο ρόλο μέσω της καινοτομίας τους συμβάλλοντας ουσιαστικά στην οικονομική ευφυΐα (Das et al., 2015). Η εργασία του Donepudi (2017) περιγράφει τις εφαρμογές της μηχανικής μάθησης και της τεχνητής νοημοσύνης και αξιολογεί τη χρησιμότητά τους σε διαφορετικούς λειτουργικούς τομείς

του τραπεζικού κλάδου. Πλαισιώνει τον τρόπο με τον οποίο οι τράπεζες χρησιμοποιούν αποτελεσματικά την υπολογιστική νοημοσύνη για να βελτιώσουν την επιχείρησή τους. Αν και τα περισσότερα χρηματοπιστωτικά ιδρύματα δεν έχουν ακόμη καθολική υιοθέτηση των τεχνολογιών υπολογιστικής νοημοσύνης, φαίνεται ότι στο μέλλον αυτές οι τεχνολογίες θα παραμείνουν και θα κυριαρχήσουν στην χρηματοοικονομική βιομηχανία (Castelli et al., 2016; Donepudi, 2016).

Το φάσμα εφαρμογών αυτών των δύο τεχνολογιών στα χρηματοοικονομικά ιδρύματα αυξάνεται καθημερινά και η συνολική τραπεζική εμπειρία έχει βελτιωθεί με πολλούς τρόπους, από την επικοινωνία, την ταχύτητα συναλλαγών μέχρι και την ασφάλεια. Παράλληλα όμως, οι πιθανοί κίνδυνοι που ενέχουν αυτά τα δύο αυξάνονται επίσης. Η τεχνολογία υπολογιστικής νοημοσύνης δεν καταλαμβάνει μόνο τον τραπεζικό κλάδο αλλά και άλλους τομείς όπως οι ασφαλιστικές εταιρείες και οι κεφαλαιαγορές. Μια μελέτη που έγινε από τους Purdy & Daugherty (2016) αποκάλυψε ότι η τεχνητή νοημοσύνη και οι εφαρμογές μηχανικής μάθησης θα μπορούσαν πιθανών να «υπαγορεύουν» τον τρόπο με τον οποίο οι τράπεζες θα αλληλεπιδρούν με τους πελάτες τους στο μέλλον.

4 Μέθοδοι μη επιβλεπόμενης μάθησης

Η μάθηση χωρίς επίβλεψη αναφέρεται στη χρήση αλγορίθμων τεχνητής νοημοσύνης για τον εντοπισμό προτύπων σε σύνολα δεδομένων που περιέχουν σημεία δεδομένων που δεν είναι ούτε ταξινομημένα ούτε επισημασμένα (labelled). Με τον τρόπο αυτό, επιτρέπεται στους αλγόριθμους να ταξινομούν, να επισημαίνουν και/ή να ομαδοποιούν τα σημεία δεδομένων που περιέχονται στα σύνολα δεδομένων χωρίς να έχουν εξωτερική καθοδήγηση κατά την εκτέλεση αυτής της εργασίας. Με άλλα λόγια, η μάθηση χωρίς επίβλεψη επιτρέπει στο σύστημα να αναγνωρίζει από μόνο του μοτίβα μέσα σε σύνολα δεδομένων.

Συγκρίνοντας την εποπτευόμενη με τη μη εποπτευόμενη μάθηση, η εποπτευόμενη μάθηση χρησιμοποιεί σύνολα δεδομένων με ετικέτα για να εκπαιδεύσει αλγόριθμους ώστε να αναγνωρίζουν και να ταξινομούν με βάση τις παρεχόμενες ετικέτες. Το αντικείμενο εισόδου, ή το δείγμα, έχει μια αντίστοιχη ετικέτα, έτσι ώστε οι αλγόριθμοι

να μαθαίνουν να αναγνωρίζουν και να ταξινομούν εκείνα τα αντικείμενα εισόδου που ταιριάζουν με την ίδια ετικέτα. Με άλλα λόγια, οι αλγόριθμοι δημιουργούν ‘χάρτες’ από δεδομένες εισόδους σε συγκεκριμένα αποτελέσματα με βάση αυτά που μαθαίνουν από δεδομένα εκπαίδευσης που έχουν επισημανθεί από μηχανικούς ή επιστήμονες δεδομένων.

Επιπλέον, η εποπτευόμενη μάθηση χρησιμοποιεί τόσο δεδομένα εκπαίδευσης (training) με ετικέτα όσο και δεδομένα επικύρωσης/ελέγχου (validation) με ετικέτα. Αυτό επιτρέπει τον έλεγχο της ακρίβειας των εποπτευόμενων μαθησιακών αποτελεσμάτων, κάτι που δεν είναι εφικ΄το στη μη εποπτευόμενη μάθηση. Οι μηχανικοί ή οι επιστήμονες δεδομένων ενδέχεται να επιλέξουν να χρησιμοποιήσουν έναν συνδυασμό δεδομένων με ετικέτα και χωρίς ετικέτα για να εκπαιδεύσουν τους αλγόριθμούς τους. Αυτή η ενδιάμεση τεχνική ονομάζεται κατάλληλα ημι-εποπτευόμενη μάθηση.

Στην μάθηση χωρίς επίβλεψη, ένα σύστημα/αλγόριθμος τεχνητής νοημοσύνης ομαδοποιεί μη ταξινομημένες πληροφορίες σύμφωνα με ομοιότητες και διαφορές, παρόλο που δεν παρέχονται κατηγορίες. Οι αλγόριθμοι μάθησης χωρίς επίβλεψη μπορούν να εκτελέσουν πιο σύνθετες εργασίες επεξεργασίας από τα εποπτευόμενα συστήματα μάθησης. Επιπλέον, η υποβολή ενός συστήματος σε μάθηση χωρίς επίβλεψη είναι και ένας τρόπος δοκιμής της τεχνητής νοημοσύνης.

Η μάθηση χωρίς επίβλεψη ξεκινά με την είσοδο δεδομένων μέσω αλγορίθμων για να τα εκπαιδεύσουν. Όπως αναφέρθηκε προηγουμένως, δεν υπάρχουν ετικέτες ή κατηγορίες που περιέχονται στα σύνολα δεδομένων που χρησιμοποιούνται για την εκπαίδευση τέτοιων συστημάτων. Κάθε μέρος δεδομένων που περνά μέσα από τους αλγόριθμους κατά τη διάρκεια της εκπαίδευσης είναι ένα αντικείμενο ή δείγμα εισόδου χωρίς ετικέτα. Ο στόχος με την μάθηση χωρίς επίβλεψη είναι οι αλγόριθμοι να αναγνωρίζουν μοτίβα μέσα στα σύνολα δεδομένων εκπαίδευσης και να κατηγοριοποιούν τα αντικείμενα εισόδου με βάση τα μοτίβα που προσδιορίζει το ίδιο το σύστημα. Οι αλγόριθμοι αναλύουν την υποκείμενη δομή των συνόλων δεδομένων εξάγοντας χρήσιμες πληροφορίες ή χαρακτηριστικά από αυτά. Έτσι, αυτοί οι αλγόριθμοι αναμένεται να αναπτύξουν συγκεκριμένες εξόδους από τις μη δομημένες εισόδους αναζητώντας σχέσεις μεταξύ κάθε δείγματος ή αντικειμένου εισόδου. Χρησιμοποιώντας ζώα ως

παράδειγμα, μπορεί να δοθούν στους αλγόριθμους σύνολα δεδομένων που περιέχουν εικόνες ζώων. Οι αλγόριθμοι μπορούν στη συνέχεια να ταξινομήσουν τα ζώα σε κατηγορίες όπως αυτά με γούνα, αυτά με λέπια και αυτά με φτερά. Στη συνέχεια, μπορούν να ομαδοποιήσουν τις εικόνες σε όλο και πιο συγκεκριμένες υποομάδες, καθώς μαθαίνουν να εντοπίζουν διαφορές σε κάθε κατηγορία.

Στα πλεονεκτήματα, η μη εποπτευόμενη μηχανική εκμάθηση μπορεί να αναγνωρίσει προηγουμένως άγνωστα μοτίβα στα δεδομένα. Μπορεί να είναι ευκολότερο, ταχύτερο και λιγότερο δαπανηρό στη χρήση από την εποπτευόμενη μάθηση, καθώς η μάθηση χωρίς επίβλεψη δεν απαιτεί τη χειρωνακτική εργασία που σχετίζεται με την επισήμανση δεδομένων που απαιτεί η εποπτευόμενη μάθηση. Επίσης, η μάθηση χωρίς επίβλεψη μπορεί να λειτουργήσει με δεδομένα σε πραγματικό χρόνο για τον εντοπισμό προτύπων.

Αν και είναι πολύ σημαντικά αυτά τα χαρακτηριστικά της μάθησης χωρίς επίβλεψη, υπάρχουν ορισμένα μειονεκτήματα. Αρχικά, είναι σημαντική η αβεβαιότητα που προκύπτει σχετικά με την ακρίβεια των μη εποπτευόμενων μαθησιακών αποτελεσμάτων. Επίσης, υπάρχει δυσκολία στον έλεγχο της ακρίβειας των μη εποπτευόμενων μαθησιακών αποτελεσμάτων, καθώς δεν υπάρχουν σύνολα δεδομένων με ετικέτα για την επαλήθευση των αποτελεσμάτων. Επιπλέον, δεν υπάρχει πλήρης εικόνα και γνώση σχετικά με το πώς ή γιατί ένα σύστημα χωρίς επίβλεψη φτάνει στα αποτελέσματά του. Όλα αυτά συνεπάγονται την ανάγκη για μηχανικούς και επιστήμονες δεδομένων να αφιερώνουν περισσότερο χρόνο στην ερμηνεία και την επισήμανση (labelling) των αποτελεσμάτων με μάθηση χωρίς επίβλεψη από ό,τι με την εποπτευόμενη μάθηση.

Εκτός από τη ομαδοποίηση, η μάθηση χωρίς επίβλεψη μπορεί να χρησιμοποιηθεί για τον προσδιορισμό του τρόπου με τον οποίο τα δεδομένα κατανέμονται στο χώρο (εκτίμηση πυκνότητας).

Σε ένα γενικότερο πλαίσιο, η διερευνητική ανάλυση (exploratory analysis), η μείωση διαστάσεων και η συσταδοποίηση είναι οι τρεις πιο κοινές χρήσεις για μάθηση χωρίς επίβλεψη. Η διερευνητική ανάλυση, στην οποία οι αλγόριθμοι χρησιμοποιούνται για την ανίχνευση προτύπων που ήταν προηγουμένως άγνωστα, έχει αρκετές εταιρικές εφαρμογές. Για παράδειγμα, οι επιχειρήσεις μπορούν να χρησιμοποιήσουν τη

διερευνητική ανάλυση ως σημείο εκκίνησης για τις προσπάθειες τμηματοποίησης/κατάταξης των πελατών τους. Στη μείωση διαστάσεων, οι αλγόριθμοι μειώνουν τον αριθμό των μεταβλητών ή των χαρακτηριστικών (δηλαδή πρακτικά των διαστάσεων) εντός των συνόλων δεδομένων, έτσι ώστε η εστίαση να μπορεί να δοθεί στα σχετικά χαρακτηριστικά για διάφορους στόχους. Ορισμένοι ειδικοί το εξηγούν λέγοντας ότι η μείωση διαστάσεων αφαιρεί τα θορυβώδη δεδομένα. (Οι μηχανικοί μηχανικής εκμάθησης χρησιμοποιούν συχνά αλγόριθμους λανθάνουσας μεταβλητής που βασίζονται σε μοντέλα για να κάνουν αυτή τη δουλειά.)

Επιπλέον, οι οργανισμοί μπορούν να χρησιμοποιούν μάθηση χωρίς επίβλεψη για τις ακόλουθες εφαρμογές:

- ανίχνευση ανωμαλιών ομαδοποίησης, όπου οι αλγόριθμοι μπορούν να εντοπίσουν ασυνήθιστα σημεία δεδομένων σε σύνολα δεδομένων, μια ικανότητα ιδιαίτερα χρήσιμη για την ταυτοποίηση δόλιας (fraudulent) δραστηριότητας ή ανθρώπινων λαθών ή ελαττωματικών προϊόντων

- συσχέτιση εξόρυξης, όπου οι αλγόριθμοι βρίσκουν συσχετίσεις μεταξύ σημείων δεδομένων, μια δυνατότητα που οι έμποροι λιανικής, για παράδειγμα, μπορούν να χρησιμοποιήσουν για να προσδιορίσουν ποια προϊόντα αγοράζονται συχνά μαζί.

4.1 Αλγόριθμοι Συσταδοποίησης (Ομαδοποίησης)

Όπως προαναφέρθηκε, μία βασική μέθοδος της μη επιβλεπόμενης μάθησης είναι η ομαδοποίησης δεδομένων (Zaki et al., 2014). Ομαδοποίηση ή Συσταδοποίηση είναι η διαδικασία της συσταδοποίησης των δεδομένων σε σύνολα ομοειδών αντικειμένων καλούμενα ομάδες (clusters). Στόχος είναι ναα παράγει ένα σύνολο από ομάδες με υψηλή εντός των ομάδων ομοιότητα (intra-cluster similarity), ενώ παράλληλα να διατηρείται χαμηλή η ομοιότητα μεταξύ των διαφόρων ομάδων (inter-cluster similarity). Η ομαδοποίηση έχει διάφορες κατηγορίες, όπως:

- Well Separated: μία συστάδα αποτελείται από το σύνολο των αντικειμένων όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της συστάδας, από ότι σε κάποιο άλλο αντικείμενο.
- Prototype Based: μία συστάδα αποτελείται από τα αντικείμενα που είναι πιο κοντά σε ένα πρωτότυπο (prototype) από ότι κάποιο άλλο αντικείμενο. Συνήθως σαν πρωτότυπο επιλέγεται το μέσο των σημείων μίας συστάδας.
- Graph Based: μία συνεκτική συνιστώσα ή μία κλίκα του γραφήματος.
- Density Based: μία πυκνή περιοχή αντικειμένων που περιβάλλεται από μία αραιή.
- Shared Property (conceptual clusters): σύνολο αντικειμένων που μοιράζονται μία ιδιότητα – έχει εφαρμογή κυρίως σε κατηγορικά αντικείμενα.

Οι 2 βασικοί πυλώνες της ομαδοποίησης είναι οι εξής:

- Διαχωριστική Συσταδοποίηση (Partitional Clustering): Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα -non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο να ανήκει σε ακριβώς ένα υποσύνολο
- Ιεραρχική Συσταδοποίηση (Hierarchical clustering): Ένα σύνολο από εμφωλευμένες (nested) ομάδες. Επιτρέπεται σε μια συστάδα να έχει υπο-συστάδες οργανωμένες ιεραρχικά (σε ένα ιεραρχικό δέντρο).

4.1.1 Ο αλγόριθμος k-means

Ο αλγόριθμος και k-means είναι αρκετά απλός και έχει ευρεία εφαρμογή ακόμα και σήμερα σε πάρα πολλά προβλήματα. Στηρίζεται στις αποστάσεις μεταξύ των σημείων (δηλαδή των δειγμάτων) και προσπαθεί να βρει ομάδες οι οποίες περιέχουν σημεία τα οποία είναι αρκετά κοντά μεταξύ τους (στο χώρο αναζήτησης που γίνεται αναζήτηση) και στο κέντρο της ομάδας, το οποίο αντιπροσωπεύει την καθεμία. Πιο συγκεκριμένα, αν δίνεται ένα σύνολο δεδομένων με n σημεία σε έναν d-διάστατο χώρο $D = \{x_i\}_{i=1}^{n}$ καθώς και το πλήθος των επιθυμητών συστάδων k, ο στόχος της βασισμένης σε αντιπροσώπους συσταδοποίησης είναι ο διαμερισμός του συνόλου δεδομένων σε k ομάδες ή συστάδες. Για κάθε συστάδα C_i υπάρχει ένα σημείο που τη αντιπροσωπεύει. Μια δημοφιλής επιλογή γι' αυτό το σημείο είναι ο μέσος μ_i όλων των σημείων της συστάδας, που ονομάζεται επίσης κέντρο βάρους:

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$$

όπου $n_i = |C_i|$ είναι το πλήθος των σημείων που ανήκουν στη συστάδα C_i.

Υπάρχει ένας απλοϊκός αλγόριθμος εξαντλητικής αναζήτησης για την εύρεση μιας καλής συσταδοποίησης:

- επιλογή όλων των πιθανών διαμερισμών των n σημείων σε k συστάδες,
- αξιολόγηση με κάποια μετρική βελτιστοποίησης ώστε να προκύψει μια «βαθμολογία» για καθεμία από τις συστάδες,
- και επιλογή της συσταδοποίησης με την καλύτερη βαθμολογία.

Όμως, αυτό είναι πρακτικά ανέφικτο επειδή υπάρχουν $O(k^n/k!)$ συσταδοποιήσεις των n σημείων σε k ομάδες.

Η συνάρτηση που βασίζεται στο άθροισμα των τετραγώνων των σφαλμάτων (SSE) ορίζεται ως:

$$\mathrm{SSE}(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

Στόχος είναι να βρεθεί εκείνη η συσταδοποίηση που ελαχιστοποιεί τη SSE:

$$C^* = \arg\min_{\mathcal{C}} \{\mathrm{SSE}(\mathcal{C})\}$$

Ο αλγόριθμος k-means χρησιμοποιεί μια άπληστη επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την αντικειμενική συνάρτηση SSE. Κατά συνέπεια, μπορεί να συγκλίνει σε τοπικά βέλτιστα και όχι σε μια καθολικά βέλτιστη συσταδοποίηση.

Οι αρχικές τιμές των μέσων για τις συστάδες καθορίζονται επιλέγοντας με τυχαίο τρόπο k σημεία στον χώρο δεδομένων. Κάθε επανάληψη του αλγορίθμου k-means αποτελείται από δύο βήματα:

1. την αντιστοίχιση σε συστάδες και
2. την ενημέρωση των κέντρων βάρους.

Με την προϋπόθεση ότι δίνονται οι μέσοι των k συστάδων, κάθε σημείο $x_j \in D$ αντιστοιχίζεται στον πλησιέστερο μέσο κατά τη διάρκεια του πρώτου βήματος του αλγορίθμου. Έτσι δημιουργείται μια συσταδοποίηση, με κάθε συστάδα C_i να περιλαμβάνει σημεία που βρίσκονται πιο κοντά στον μέσο μ_i σε σύγκριση με τον μέσο οποιασδήποτε άλλης συστάδας. Δηλαδή, κάθε σημείο x_j αντιστοιχίζεται στη συστάδα C_{j^*}, όπου

$$j^* = \arg\min_{i=1} \{||x_j - \mu_i||^2\}$$

Για ένα καθορισμένο σύνολο συστάδων C_i, i = 1,…, k, στο δεύτερο βήμα του αλγορίθμου (ενημέρωση των κέντρων βάρους) υπολογίζονται νέες μέσες τιμές για κάθε συστάδα από τα σημεία του συνόλου C_i. Τα βήματα της αντιστοίχισης σε συστάδες και της ενημέρωσης των κέντρων βάρους εκτελούνται επαναληπτικά μέχρι να καταλήξουμε σε ένα σταθερό σημείο ή σε τοπικά ελάχιστα.

Συγκεντρωτικά, τα βήματα του αλγορίθμου είναι τα εξής

1. Ορισμός K κέντρων συστάδων με τυχαίο τρόπο (το k το ορίζει ο χρήστης)
2. Εισαγωγή αντικειμένου στη συστάδα με το πιο κοντινό κέντρο (με κριτήριο απόστασης που ορίζεται από τον χρήστη (πχ Ευκλείδεια)
3. Ανανέωση του κέντρου της συστάδας (Μέσος όρος των σημείων κάθε cluster)
4. Επανάληψη των βημάτων 2,3 μέχρι τη σύγκλιση (αλλαγή στις συστάδες μικρότερη από ένα threshold)

Ουσιαστικά, ο αλγόριθμος προσπαθεί επαναληπτικά να μειώσει την απόσταση όλων των σημείων από ένα σημείο της συστάδας. Σχετικά με τη σύγκλιση του αλγόριθμου, συμβαίνει και σε άλλες περιπτώσεις όπως όταν ξεπεραστούν οι επαναλήψεις που έχει ορίσει ο χρήστης ή όταν το σφάλμα μετά από κάθε επανάληψη δεν μειώνεται πολύ (μικρότερο από ένα δεδομένο threshold) ή όταν κατά την ανανέωση του κέντρου δεν μετακινούνται πολύ τα κέντρα και αλλάζουν θέση με διαφορά μικρότερη από ένα δοσμένο threshold. Οι τελευταίες δύο περιπτώσεις σχετίζονται μεταξύ τους γιατί όταν ικανοποιούνται αυτές οι συνθήκες σημαίνει ότι οι συστάδες στην επόμενη επανάληψη δεν είχανε έντονες αλλαγές, δηλαδή δεν μετακινήθηκαν πολλά σημεία από το ένα cluster στο άλλο. Για παράδειγμα αν υπάρχουν 1.000 σημεία στο σύνολο δεδομένων και σε μία επανάληψη έχουν αλλάξει cluster τρία από αυτά τα σημεία, είναι μία πάρα πολύ μικρή

αλλαγή η οποία ίσως δεν είναι ικανή να αλλάξεις το αποτέλεσμα. Συνεπώς ο αλγόριθμος μπορεί σε αυτό το σημείο να τερματιστεί και να μην εκτελέσει άλλες επαναλήψεις.

Στην ίδια κατηγορία ανήκει και ο αλγόριθμος k-medoid, ο οποίος εφαρμόζει την επαναληπτική διαδικασία του k-means, με τη διαφορά ότι για τον εντοπισμό του κέντρου της ομάδας επιλέγει το πιο κεντρικό σημείο της συστάδας (αντί να χρησιμοποιεί το μέσο όρο). Δηλαδή το κέντρο του cluster είναι σημείο του συνόλου δεδομένων και όχι ένα σημείο στον χώρο. Ο αλγόριθμος αυτός μειώνει την ευαισθησία σε outliers και μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα). Το πρώτο γίνεται κατανοητό μιας και σε ένα cluster αν ένα σημείο είναι πολύ απομακρυσμένο θα επηρεάζει το κέντρο της ομάδας και θα το απομακρύνει από την ρεαλιστική του θέση.

4.1.2 Ιεραρχική συσταδοποίηση

Η ιεραρχική ομαδοποίηση, γνωστή και ως ανάλυση ιεραρχικής συστάδας (Hierarchical Cluster Analysis), είναι ένας αλγόριθμος ομαδοποίησης χωρίς επίβλεψη που μπορεί να κατηγοριοποιηθεί με δύο τρόπους. Μπορεί να είναι συγκεντρωτικός/συσσωματωτικός (agglomerative) ή διασπαστικός (divisive) (Zaki et al., 2014).

Η συγκεντρωτική ομαδοποίηση θεωρείται μια «προσέγγιση από κάτω προς τα πάνω» και λειτουργεί συνθετικά. Δηλαδή, ξεκινά με μία ξεχωριστή συστάδα για καθένα από τα n σημεία και κατόπιν συγχωνεύει επανειλημμένα το ζεύγος των συστάδων οι οποίες εμφανίζουν τη μεγαλύτερη ομοιότητα (μικρότερη απόσταση). Πιο αναλυτικά, αν δίνεται ένα σύνολο συστάδων $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, βρίσκουμε το ζεύγος των πλησιέστερων/πιο όμοιων συστάδων C_i και C_j και τις συγχωνεύουμε σε μια νέα συστάδα $C_{ij} = C_i \cup C_j$. Στη συνέχεια ενημερώνουμε το σύνολο των συστάδων διαγράφοντας τις συστάδες C_i και C_j, και προσθέτοντας τη συστάδα C_{ij}, ως εξής:

$$C = \left(C \backslash \{C_i, C_j\}\right) \cup \{C_{ij}\} \quad .$$

Η διαδικασία επαναλαμβάνεται έως ότου το σύνολο C να περιέχει μόνο μία συστάδα και μπορεί να σταματήσει όταν υπάρχουν ακριβώς k εναπομείναντες συστάδες, αν αυτό καθορίζεται.

Τέσσερις διαφορετικές μέθοδοι χρησιμοποιούνται συνήθως για τη μέτρηση της ομοιότητας:

- Σύνδεση Ward (Ward's linkage): Αυτή η μέθοδος δηλώνει ότι η απόσταση μεταξύ δύο συστάδων ορίζεται από την αύξηση του αθροίσματος του τετραγώνου της απόστασης μετά τη συγχώνευση των συστάδων.

- Μέση σύνδεση (Average linkage): Αυτή η μέθοδος ορίζεται από τον μέσο όρο της απόστασης ανά ζεύγη μεταξύ σημείων της συστάδας Ci και της συστάδας Cj:

$$\delta(C_i, C_j) = \frac{\sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}, \mathbf{y})}{n_i \cdot n_j}$$

- Πλήρης (ή μέγιστη) σύνδεση (Complete linkage): Αυτή η μέθοδος ορίζεται από τη μέγιστη απόσταση μεταξύ δύο σημείων σε κάθε σύμπλεγμα, για παράδειγμα ενός σημείου της συστάδας C_i και ενός σημείου της συστάδας C_j:

$$\delta(C_i, C_j) = \max\{\delta(x, y) \mid x \in C_i, y \in C_j\}$$

- Ενιαία (ή ελάχιστη) σύνδεση (Single linkage): Αυτή η μέθοδος ορίζεται από την ελάχιστη απόσταση μεταξύ δύο σημείων σε κάθε σύμπλεγμα, για παραδειγμα ενός σημείου της συστάδας C_i από ένα σημείο της συστάδας C_j:

$$\delta(C_i, C_j) = \min\{\delta(x, y) \mid x \in C_i, y \in C_j\}$$

Η απόσταση δύο σημείων υπολογίζεται συνήθως με χρήση της Ευκλείδειας απόστασης ή L_2-νόρμας, που ορίζεται ως:

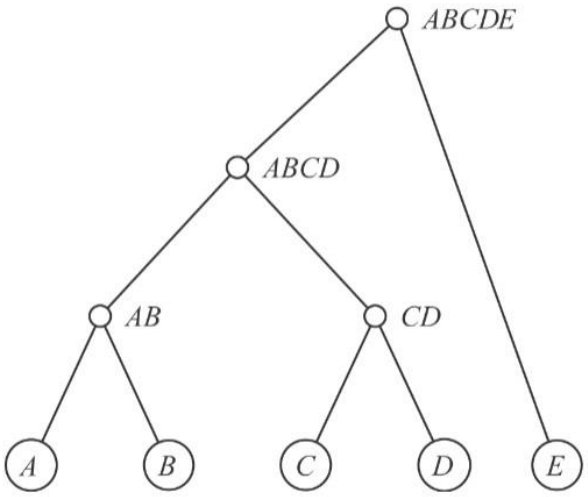
$$\delta(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_2 = \left(\sum_{i=1}^{d} (x_i - y_i)^2 \right)^{1/2}$$

Ωστόσο, άλλες μετρήσεις, όπως η απόσταση του Μανχάταν, αναφέρονται επίσης στη βιβλιογραφία ομαδοποίησης.

Η διαιρετική ομαδοποίηση μπορεί να οριστεί ως το αντίθετο της συσσωρευτικής ομαδοποίησης. Αντίθετα, ακολουθεί μια προσέγγιση «από πάνω προς τα κάτω». Σε αυτήν την περίπτωση, ένα μεμονωμένο σύμπλεγμα δεδομένων χωρίζεται με βάση τις διαφορές μεταξύ των σημείων δεδομένων. Η διαιρετική ομαδοποίηση δεν χρησιμοποιείται συνήθως, αλλά αξίζει να σημειωθεί στο πλαίσιο της ιεραρχικής ομαδοποίησης.

Αυτές οι διαδικασίες ομαδοποίησης συνήθως οπτικοποιούνται χρησιμοποιώντας ένα δενδρογραμμα, ένα διάγραμμα που μοιάζει με δέντρο που τεκμηριώνει τη συγχώνευση ή τον διαχωρισμό σημείων δεδομένων σε κάθε επανάληψη. Για παράδειγμα, το παρακάτω δενδρόγραμμα αναπαριστά την παρακάτω ακολουθία διαιρετικής συσταδοποίησης, με $C_{t-1} \subset C_t$ για t = 2,…,5. Υποθέτουμε ότι οι συστάδες A και B συγχωνεύονται πριν από τις συστάδες C και D.

Συσταδοποίηση	Συστάδες
C_1	{A}, {B}, {C}, {D}, {E}
C_2	{AB}, {C}, {D}, {E}
C_3	{AB}, {CD}, {E}
C_4	{ABCD}, {E}
C_5	{ABCDE}

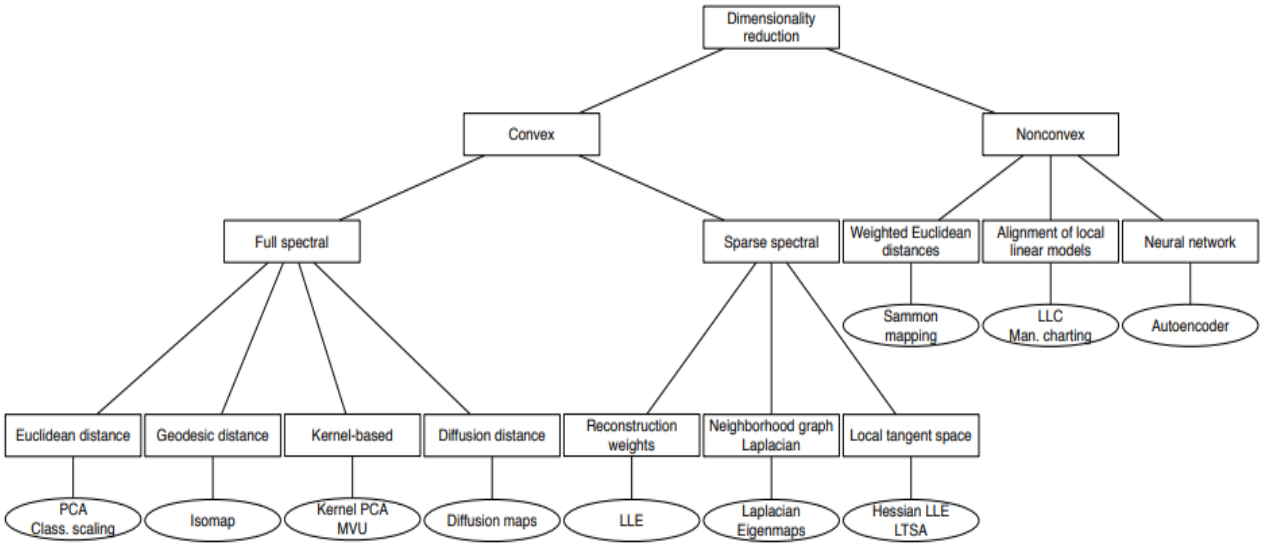


Εικόνα 1. .Παράδειγμα ιεραρχικής διαιρετικής συσταδοποίησης με δενδρόγραμμα

4.2 Μέθοδοι Μείωσης Διαστάσεων

Το πρόβλημα της (μη γραμμικής) μείωσης διαστάσεων μπορεί να οριστεί ως εξής. Έστω ότι δίνεται ένα σύνολο δεδομένων που αντιπροσωπεύεται σε έναν πίνακα $\mathbf{X}$ $(n \times D)$, ο οποίος αποτελείται από n διανύσματα δεδομένων $\mathbf{x}_i$ $(i \in \{1,2,\dots,n\})$, με διαστάσεις D. Επιπλέον, έστω ότι αυτό το σύνολο δεδομένων έχει εγγενή διάσταση d (όπου $d < D$, και συχνά $d \ll D$). Στη προκειμένη περίπτωση, με μαθηματικούς όρους, η εγγενής διάσταση σημαίνει ότι τα σημεία του συνόλου δεδομένων **X,** βρίσκονται πάνω ή κοντά σε μια πολλαπλότητα (manifold) με διαστάσεις d, η οποία είναι ενσωματωμένη στον διαστατικό χώρο D. Ωστόσο, πρέπει να σημειωθεί ότι δεν γίνονται υποθέσεις σχετικά με τη δομή αυτής της πολλαπλότητας. Πιο συγκεκριμένα και εξαιτίας πιθανών ασυνεχειών (δηλαδή, η πολλαπλότητα μπορεί να αποτελείται από έναν αριθμό αποσυνδεδεμένων υποπλασιών), η πολλαπλότητα μπορεί να είναι ψευδο- Riemannian (pseudo-Riemannian manifold). Επίσης, οι τεχνικές μείωσης των διαστάσεων, μετατρέπουν το σύνολο δεδομένων $\mathbf{X}$ (με διαστάσεις D), σε ένα νέο σύνολο δεδομένων $\mathbf{Y}$ με διαστάσεις d. Παράλληλα, διατηρούν όσο το δυνατόν περισσότερο, τη γεωμετρία των δεδομένων. Γενικά, δεν είναι γνωστές ούτε η γεωμετρία της πολλαπλότητας δεδομένων αλλά ούτε και η εγγενής διάσταση d του συνόλου δεδομένων $\mathbf{X}$. Επομένως, η μείωση των διαστάσεων είναι ένα δύσκολο πρόβλημα που μπορεί να επιλυθεί μόνο αν γίνει κάποια υπόθεση για ορισμένες ιδιότητες των δεδομένων (π.χ. για την εγγενή τους διάσταση). Για το λόγο αυτό, υποδηλώνεται ένα σημείο δεδομένων (υψηλής διάστασης) $\mathbf{x}_i$, το οποίο αποτελεί την αντίστοιχη i σειρά του πίνακα δεδομένων $\mathbf{X}$, με διαστάσεις D. Το αντίστοιχο σημείο (χαμηλής διάστασης) του $\mathbf{x}_i$, συμβολίζεται με $\mathbf{y}_i$, όπου $\mathbf{y}_i$ είναι αντίστοιχα η i σειρά του πίνακα δεδομένων Y (d διαστάσεων). Τελικώς, υιοθετώντας τις παραπάνω καταγραφές, εκλαμβάνεται ως υπόθεση ότι το σύνολο δεδομένων **X,** έχει μηδενικό μέσο όρο Maaten & Postma (2009).

Η παρούσα εργασία επικεντρώνεται στην Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis – PCA) η οποία και αναλύεται με λεπτομέρειες στο επόμενο κεφάλαιο. Περιγράφονται επίσης σύντομα οι Multiple Correspondence Analysis (M.C.A.) και Factor Analysis of Mixed Data (F.A.M.D.) που χρησιμοποιούνται στην υλοποίηση. Στο παρακάτω διάγραμμα παρουσιάζεται μια επισκόπηση και ταξινόμηση τεχνικών για μείωση διαστάσεων.



Εικόνα 2. Ταξινόμηση τεχνικών για μείωση διαστάσεων.
Πηγή: (Van Der Maaten et al., 2009).

4.2.1 Principal Components Analysis (P.C.A.)

Η ανάλυση κυρίων συνιστωσών (Principal Components Analysis - PCA) είναι μια γραμμική τεχνική που χρησιμοποιείται για τη μείωση διαστάσεων, ενσωματώνοντας τα δεδομένα σε έναν γραμμικό υποχώρο χαμηλότερης διαστατικότητας. Αν και υπάρχουν διάφορες τεχνικές για αυτό το σκοπό, η PCA είναι μακράν η πιο δημοφιλής γραμμική τεχνική (χωρίς επίβλεψη). Μπορεί ουσιαστικά να προβλέψει ποιες μεταβλητές είναι περισσότερο σημαντικές, οδηγώντας έτσι σε ένα μικρότερο αριθμό από ομαδοποιημένες μεταβλτές. Ο λόγος που προτιμάται σε σχέση με τις υπόλοιπες τεχνικές, είναι ότι κατασκευάζει μια αναπαράσταση (χαμηλών διαστάσεων) των δεδομένων, η οποία περιγράφει με όσο το δυνατόν μεγαλύτερη ακρίβεια, την απόκλισή τους. Αυτό γίνεται με την εύρεση μιας γραμμικής βάσης (μειωμένης διαστατικότητας/διάστασης), στην οποία

η ποσότητα της διακύμανσης των δεδομένων μπορεί να φτάσει ευκολότερα, στη μέγιστη δυνατή τιμή.

Η διαδικασία αποτελείται από ορισμένα βήματα¨

1. Κανονικοποίηση

Ο στόχος αυτού του βήματος είναι να 'κανονικοποιήσει' το εύρος των συνεχών αρχικών μεταβλητών έτσι ώστε κάθε μία από αυτές να συνεισφέρει εξίσου στην ανάλυση.

Πιο συγκεκριμένα, ο λόγος για τον οποίο είναι κρίσιμο να πραγματοποιηθεί η τυποποίηση πριν από την PCA, είναι ότι η τελευταία είναι αρκετά ευαίσθητη ως προς τις διακυμάνσεις των αρχικών μεταβλητών. Δηλαδή, εάν υπάρχουν μεγάλες διαφορές μεταξύ των περιοχών των αρχικών μεταβλητών, αυτές οι μεταβλητές με μεγαλύτερα εύρη θα κυριαρχούν σε αυτές με μικρές περιοχές (Για παράδειγμα, μια μεταβλητή που κυμαίνεται μεταξύ 0 και 100 θα κυριαρχεί σε μια μεταβλητή που κυμαίνεται μεταξύ 0 και 1), το οποίο θα οδηγήσει σε μεροληπτικά αποτελέσματα. Έτσι, η μετατροπή των δεδομένων σε συγκρίσιμες κλίμακες μπορεί να αποτρέψει αυτό το πρόβλημα.Αυτό γίνεται με την σχέση:

z=values-mean/standard deviation

Μόλις γίνει η κανονικοποίηση, όλες οι μεταβλητές θα μετατραπούν στην ίδια κλίμακα.

2. Υπολογισμός πίνακα συνδιακύμανσης (covariance)

Ο στόχος αυτού του βήματος είναι να γίνει κατανοητό πώς οι μεταβλητές του συνόλου δεδομένων εισόδου ποικίλλουν/απέχουν από τη μέση τιμή μεταξύ τους, ή με άλλα λόγια, αν υπάρχει κάποια σχέση μεταξύ τους. Επειδή μερικές φορές, οι μεταβλητές συσχετίζονται σε μεγάλο βαθμό με τέτοιο τρόπο που περιέχουν περιττές πληροφορίες. Έτσι, για να προσδιορίσουμε αυτές τις συσχετίσεις, υπολογίζουμε τον πίνακα συνδιακύμανσης.

Ο πίνακας συνδιακύμανσης είναι ένας συμμετρικός πίνακας p × p (όπου p είναι ο αριθμός των διαστάσεων) που έχει ως καταχωρήσεις τις συνδιακυμάνσεις που σχετίζονται με όλα τα πιθανά ζεύγη των αρχικών μεταβλητών. Για παράδειγμα, για ένα

τρισδιάστατο σύνολο δεδομένων με 3 μεταβλητές x, y και z, ο πίνακας συνδιακύμανσης είναι ένας πίνακας 3×3 αυτής της μορφής:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Εικόνα 3. Πίνακας συνδιακύμανσης

Εφόσον η συνδιακύμανση μιας μεταβλητής με τον εαυτό της είναι η διακύμανσή της (Cov(a,a)=Var(a)), στην κύρια διαγώνιο, έχουμε στην πραγματικότητα τις διακυμάνσεις κάθε αρχικής μεταβλητής. Και επειδή η συνδιακύμανση είναι αντιμεταθετική (Cov(a,b)=Cov(b,a)), οι εγγραφές του πίνακα συνδιακύμανσης είναι συμμετρικές ως προς την κύρια διαγώνιο, πράγμα που σημαίνει ότι το επάνω και το κάτω τριγωνικό τμήμα είναι ίσα.

Η συνδιακύμανση πρακτικά είναι σημαντική για το πρόσιμο που έχει. Αν είναι θετικό τότε οι δύο μεταβλητές αυξάνονται ή μειώνονται μαζί (συσχετίζονται) ενώ αν είναι αρνητικό τότε το ένα αυξάνεται όταν το άλλο μειώνεται (Αντιστρόφως συσχετισμένο)

3. Υπολογισμός ιδιοδιανυσμάτων και ιδιοτιμών για να προσδιοριστούν οι κύριες συνιστώσες.

Οι κύριες συνιστώσες είναι νέες μεταβλητές που κατασκευάζονται ως γραμμικοί συνδυασμοί ή μείξεις των αρχικών μεταβλητών. Αυτοί οι συνδυασμοί γίνονται με τέτοιο τρόπο ώστε οι νέες μεταβλητές (δηλαδή οι κύριες συνιστώσες) να μην συσχετίζονται και οι περισσότερες πληροφορίες εντός των αρχικών μεταβλητών συμπιέζονται ή συμπιέζονται στα πρώτα στοιχεία. Έτσι, η ιδέα είναι ότι τα 10-διάστατα δεδομένα δίνουν 10 κύρια στοιχεία, αλλά το PCA προσπαθεί να βάλει τις μέγιστες δυνατές πληροφορίες στο πρώτο στοιχείο, μετά τις μέγιστες υπολειπόμενες πληροφορίες στο δεύτερο και ούτω καθεξής.

Καθώς υπάρχουν τόσα κύρια στοιχεία όσες και οι μεταβλητές στα δεδομένα, τα κύρια στοιχεία κατασκευάζονται με τέτοιο τρόπο ώστε το πρώτο κύριο στοιχείο να αντιπροσωπεύει τη μεγαλύτερη δυνατή απόκλιση στο σύνολο δεδομένων.

Τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης είναι στην πραγματικότητα οι κατευθύνσεις των αξόνων όπου υπάρχει η μεγαλύτερη διακύμανση (περισσότερες πληροφορίες) και που ονομάζουμε κύριες συνιστώσες . Οι ιδιοτιμές είναι απλώς οι συντελεστές που συνδέονται με τα ιδιοδιανύσματα, οι οποίες δίνουν το ποσό της διακύμανσης που μεταφέρεται σε κάθε κύρια συνιστώσα.

4. Feature Vector

Ο υπολογισμός των ιδιοδιανυσμάτων και η ταξινόμηση τους με βάση τις ιδιοτιμές τους σε φθίνουσα σειρά, μας επιτρέπουν να βρούμε τα κύρια συστατικά κατά σειρά σημασίας. Σε αυτό το βήμα, γίνεται επιλογή για το αν θα διατηρηθούν όλα αυτά τα συστατικά ή θα απορρίφθούν εκείνα μικρότερης σημασίας (χαμηλών ιδιοτιμών) ώστε να σχηματιστεί με τα υπόλοιπα ένας πίνακα διανυσμάτων που ονομάζουμε Feature vector.

5. Τοποθέτηση των αρχικών δεδομένων κατά τους άξονες των κύριων συνιστωσών

Στα προηγούμενα βήματα, εκτός από την κανονικοποίηση, δεν έγινε καμία αλλαγή στα δεδομένα, απλώς επιλέγονται τα κύρια στοιχεία και σχηματίζεται το feature vector, αλλά το σύνολο δεδομένων εισόδου παραμένει πάντα ως προς τους αρχικούς άξονες (δηλ. οι αρχικές μεταβλητές).

Σε αυτό το βήμα, που είναι το τελευταίο, ο στόχος είναι να χρησιμοποιηθεί το feature vetor για να επαναπροσανατολίσει τα δεδομένα από τους αρχικούς άξονες σε αυτούς που αντιπροσωπεύονται από τις κύριες συνιστώσες (εξ ου και η ονομασία Principal Components Analysis). Αυτό μπορεί να γίνει πολλαπλασιάζοντας τη μεταφορά (transpose) του αρχικού συνόλου δεδομένων με τη μεταφορά του feature vector.

Η τεχνική PCA είναι πανομοιότυπη με την παραδοσιακή τεχνική της πολυδιάστατης κλιμάκωσης, που ονομάζεται κλασική κλιμάκωση. Όπως και στις περισσότερες πολυδιάστατες τεχνικές κλιμάκωσης, έτσι και στην κλασική κλιμάκωση, καταχωρείται αρχικά ένας κατά ζεύγη, Ευκλείδειος πίνακας αποστάσεων **D**. Κατ' επέκταση, οι

καταχωρήσεις του πίνακα (d_{ij}), αντιπροσωπεύουν την Ευκλείδεια απόσταση των σημείων, όσον αφορά τα δεδομένα υψηλών διαστάσεων ($\mathbf{x}_i$ και $\mathbf{x}_j$). Έπειτα, η κλασική κλιμάκωση εντοπίζει τη γραμμική αντιστοίχιση **M,** η οποία ελαχιστοποιεί τη συνάρτηση 'κόστους':

$$\phi(\mathbf{Y}) = \sum_{ij} \left(d_{ij}^2 - \|\mathbf{y}_i - \mathbf{y}_j\|^2 \right)$$

στην οποία, η $\|\mathbf{y}_i - \mathbf{y}_j\|^2$ είναι η τετραγωνισμένη Ευκλείδεια απόσταση μεταξύ των σημείων $\mathbf{y}_i$ και $\mathbf{y}_j$ και το $\mathbf{y}_i$ περιορίζεται να είναι xiM, και ||mj||2=1 για το ∀j. Σύμφωνα με τη μελέτη του (Williams, 2002), η ελάχιστη τιμή της συνάρτησης κόστους, δίνεται από την ιδιοδιάσπαση (eigendecomposition) των δεδομένων υψηλών διαστάσεων, του πίνακα Gram ($\mathbf{K} = \mathbf{X}\mathbf{X}^T$). Οι καταχωρίσεις του πίνακα μπορούν να ληφθούν με διπλή κεντροθέτηση του (κατά ζεύγους τετραγωνισμένου), Ευκλείδειου πίνακα αποστάσεων. Δηλαδή, μέσω του παρακάτω υπολογισμού

$$k_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_l d_{il}^2 - \frac{1}{n}\sum_l d_{jl}^2 + \frac{1}{n^2}\sum_{lm} d_{lm}^2 \right)$$

Επομένως, η ελάχιστη τιμή της συνάρτησης (μπορεί να ληφθεί πολλαπλασιάζοντας τα κύρια ιδιοδιανύσματα του διπλά κεντραρισμένου τετραγωνικου Ευκλείδειου πίνακα αποστάσεων (δηλαδή τα κύρια ιδιοδιανύσματα του πίνακα Gram), με την τετραγωνική ρίζα των αντίστοιχων ιδιοτιμών τους.

Η ομοιότητα της κλασικής κλιμάκωσης με αυτή της PCA, οφείλεται σε μια σχέση μεταξύ των ιδιοδιανυσμάτων τόσο του πίνακα συνδιακύμανσης όσο και του πίνακα Gram. Δηλαδή, τα ιδιοδιανύσματα $\mathbf{u}_i$ και $\mathbf{v}_i$ των πινάκων $\mathbf{X}^T\mathbf{X}$ και $\mathbf{X}\mathbf{X}^T$, σχετίζονται μέσω της $\sqrt{\lambda_i}\mathbf{v}_i = \mathbf{X}\mathbf{u}_i$. Τέλος, η σύνδεση μεταξύ της PCA και της κλασικής κλιμάκωσης, περιγράφεται με περισσότερες λεπτομέρειες σε διάφορες εμπειρικές μελέτες, όπως για παράδειγμα σε αυτή του (Platt, 2005).

4.2.2 Multiple Correspondence Analysis (M.C.A.)

Η Multiple Correspondence Analysis (M.C.A.) είναι ένα εργαλείο επιστήμης δεδομένων για τη ‘σύνοψη’ πινάκων. Ουσιαστικά το dataset αντιμετωπίζεται σαν ένας πίνακας με γραμμές και στήλες. Είναι μια τεχνική ανάλυσης δεδομένων για κατηγορικά δεδομένα, που χρησιμοποιείται για τον εντοπισμό και την αναπαράσταση υποκείμενων δομών σε ένα σύνολο δεδομένων. Αυτό το κάνει αναπαριστώντας δεδομένα ως σημεία σε έναν ευκλείδειο χώρο χαμηλής διάστασης. Επομένως, η διαδικασία φαίνεται να είναι το αντίστοιχο της PCA που αναλύθηκε προηγουμένως αλλά για κατηγορικά δεδομένα.

Η MCA εκτελείται εφαρμόζοντας τον αλγόριθμο Correspondence Analysis (C.A.) είτε σε έναν πίνακα δεικτών (ονομάζεται επίσης πλήρης διαχωριστικός πίνακας – CDT) είτε σε έναν πίνακα Burt που σχηματίζεται από αυτές τις μεταβλητές. Ένας πίνακας δεικτών είναι ένας πίνακας εγγραφών × μεταβλητών, όπου οι σειρές αντιπροσωπεύουν εγγραφές και οι στήλες είναι εικονικές μεταβλητές που αντιπροσωπεύουν κατηγορίες των μεταβλητών. Η ανάλυση του πίνακα δεικτών επιτρέπει την άμεση αναπαράσταση των εγγραφών ως σημείων στο γεωμετρικό χώρο. Ο πίνακας Burt είναι ο συμμετρικός πίνακας όλων των αμφίδρομων διασταυρούμενων πινάκων μεταξύ των κατηγορικών μεταβλητών και έχει μια αναλογία με τον πίνακα συνδιακύμανσης των συνεχών μεταβλητών. Η ανάλυση του πίνακα Burt είναι μια πιο φυσική γενίκευση της απλής CA και οι εγγραφές ή οι μέσοι όροι των ομάδων των εγγραφών μπορούν να προστεθούν ως συμπληρωματικά σημεία στη γραφική απεικόνιση.

Στην προσέγγιση του πίνακα δεικτών, οι συσχετίσεις μεταξύ των μεταβλητών υποδυκνύονται με τον υπολογισμό της απόστασης χ-τετράγωνο (chi-square) μεταξύ των διαφορετικών κατηγοριών των μεταβλητών και μεταξύ των εγγραφών. Αυτές οι συσχετίσεις στη συνέχεια αναπαρίστανται γραφικά ως "χάρτες", γεγονός που διευκολύνει την ερμηνεία των δομών στα δεδομένα. Στη συνέχεια, οι αντιθέσεις μεταξύ σειρών και στηλών μεγιστοποιούνται, προκειμένου να αποκαλυφθούν οι υποκείμενες διαστάσεις που μπορούν να περιγράψουν καλύτερα τις κεντρικές αντιθέσεις στα δεδομένα. Όπως και στην PCA, ο πρώτος άξονας είναι η πιο σημαντική διάσταση, ο δεύτερος άξονας η δεύτερη πιο σημαντική, και ούτω καθεξής, όσον αφορά το ποσό της διακύμανσης που λαμβάνεται υπόψη. Ο αριθμός των αξόνων που θα διατηρηθούν για ανάλυση προσδιορίζεται με τον υπολογισμό των τροποποιημένων ιδιοτιμών.

4.2.3 Factor Analysis Mixed Data (F.A.M.D.)

Στη στατιστική, η παραγοντική ανάλυση μικτών δεδομένων ή παραγοντική ανάλυση μικτών δεδομένων (FAMD), είναι η παραγοντική μέθοδος κατάλληλη για σύνολα δεδομένων στους οποίους μια ομάδα ατόμων περιγράφεται τόσο με ποσοτικές όσο και ποιοτικές μεταβλητές.

Ο όρος μικτή αναφέρεται στη χρήση τόσο ποσοτικών όσο και ποιοτικών μεταβλητών. Γενικά, μπορούμε να πούμε ότι η FAMD λειτουργεί ως PCA για ποσοτικές μεταβλητές και ως MCA για ποιοτικές μεταβλητές.

Η αναπαράσταση των εγγραφών γίνεται απευθείας από τους παράγοντες (factors). Η αναπαράσταση των ποσοτικών μεταβλητών κατασκευάζεται όπως στην PCA. Η αναπαράσταση των κατηγοριών των ποιοτικών μεταβλητών είναι όπως στην MCA: μια κατηγορία βρίσκεται στο κέντρο των ατόμων που την κατέχουν. Σημειώνεται ότι λαμβάνεται το ακριβές κέντρο και όχι, όπως συνηθίζεται στην MCA, το κέντρο μέχρι έναν συντελεστή που εξαρτάται από τον άξονα.

5 Βιβλιογραφική Ανασκόπηση Μοντελοποίησης Πιστωτικού Κινδύνου Μέσω Μηχανικής Μάθησης

Εισαγωγικά, όπως αναφέρθηκε και προηγουμένως, οι προβλέψεις των μοντέλων πιστοληπτικής ικανότητας έχουν πλέον γίνει σημαντικό στοιχείο του χρηματοπιστωτικού/τραπεζικού κόσμου. Τα τελευταία χρόνια έχουν χαρακτηριστεί από μια τεχνολογική εξέλιξη και ένα «ψηφιακό» κύμα, τα οποία προσφέρουν νέες ευκαιρίες για τη βελτίωση των επιχειρησιακών πρακτικών και την υιοθέτηση πιο προηγμένων μεθοδολογικών προσεγγίσεων σε διαφορετικούς τομείς έρευνας. Σε ένα πλαίσιο αυξανόμενου ανταγωνισμού και πτώσης των περιθωρίων κέρδους, η μηχανική μάθηση (Machine Learning) μπορεί να διαδραματίσει σημαντικό ρόλο τόσο στην τεχνολογία όσο και στις επιχειρήσεις, επιτρέποντας στα χρηματοπιστωτικά ιδρύματα να μεγιστοποιήσουν την αξία των δικών τους δεδομένων. Πολλές μελέτες επικεντρώνονται

στην υιοθέτηση τεχνικών μηχανικής μάθησης για τη διαμόρφωση παραμέτρων πιστωτικού κινδύνου, χρησιμοποιώντας διαφορετικές μεθοδολογίες για την εκτίμηση της πιθανότητας αθέτησης.

5.1 Βαθμολόγηση πιστοληπτικής ικανότητας

Η σημαντική πτυχή των περιουσιακών στοιχείων που χρησιμοποιούνται στον τραπεζικό τομέα, προέρχεται άμεσα από το κέρδος που αποκτάται από τη διανομή για παράδειγμα, πιστωτικών καρτών μεταξύ των πελατών και δανείων. Έχουν γίνει διάφορες προσπάθειες βαθμολόγησης της πιστοληπτικής ικανότητας ενός πελάτη ώστε να εντοπιστούν οι ‘άξιοι’ πελάτες και να μελετηθούν οι συνέπειες όσων καταλήξουν σε αθέτηση πληρωμής. Με τον τρόπο αυτό, μπορεί μια τράπεζα να αποκομίσει το μέγιστο κέρδος από την επένδυση στα περιουσιακά στοιχεία. Ο τραπεζικός τομέας μπορεί να επηρεάσει τη ζωή ενός κατόχου δανείου ή ενός κατόχου πιστωτικής κάρτας. Πιο συγκεκριμένα, αν και οι χρηματοπιστωτικές εταιρείες μπορούν να εκδώσουν πιστωτικές κάρτες μετά από μια διεξοδική διαδικασία επαλήθευσης και επικύρωσης, δεν υπάρχει καμία εγγύηση ότι οι πιστωτικές κάρτες χορηγήθηκαν σε ‘άξιους’ υποψηφίους. Η πιστοληπτική αξιολόγηση είναι ένα συμβατικό μοντέλο απόφασης, το οποίο εστιάζει στην προσέγγιση κινδύνου που σχετίζεται με πιστωτικά προϊόντα όπως πιστωτικές κάρτες, δάνεια κ.λπ. Εκτιμάται με βάση τα ιστορικά δεδομένα των αιτούντων και βοηθά τους δανειστές στη χορήγηση πιστωτικών προϊόντων. Επί του παρόντος, τα χρηματοπιστωτικά ιδρύματα υιοθετούν διάφορα εργαλεία και τεχνικές αξιολόγησης κινδύνου για συστήματα βαθμολόγησης πιστοληπτικής ικανότητας ούτως ώστε να ελαχιστοποιήσουν τον κίνδυνο (έως ένα βαθμό). Η χρήση στατιστικών εργαλείων για την ανάλυση δεδομένων πίστωσης πελατών, βοηθά στον εντοπισμό των αθετήσεων από τους πελάτες και ο εντοπισμός αυτός συνεισφέρει και στην μείωση του πιστωτικού κινδύνου και κατά συνέπεια, στην κερδοφορία του ιδρύματος.

Ορισμένες μελέτες έχουν επικεντρωθεί στα πλεονεκτήματα της χρήσης συστημάτων μηχανικής μάθησης σε προβλήματα βαθμολόγησης πιστωτικών μονάδων και τον τρόπο με τον οποίο μπορούν να επιτύχουν ανώτερη απόδοση σε σχέση με τις παραδοσιακές τεχνικές, π.χ. Λογιστική παλινδρόμηση (Logistic Regression). Η εμφάνιση αυτών των μεθόδων σε βιβλιοθήκες ανοιχτού κώδικα (όπως R ή Weka) και σε ιδιόκτητες λύσεις

λογισμικού (π.χ. SAS) τις έχει καταστήσει ευρέως διαθέσιμες στον γενικό πληθυσμό και στους ίδιους τους χρηματοδότες.

Μελέτες έχουν δείξει ότι οι προσεγγίσεις που βασίζονται σε μοντέλα στοχαστικής διαδικασίας, είναι πολύ αποτελεσματικές στην αξιολόγηση της πιστωτικής βαθμολογίας. Για την εκτίμηση του πιστωτικού κινδύνου χρησιμοποιούνται εμπειρικά μοντέλα για τη λήψη αποφάσεων τόσο για εταιρικές όσο και για λιανικές πιστωτικές επιχειρήσεις. Σύμφωνα με τους (Serrano-Cinca et al., 2015), τα εν λόγω μοντέλα έχουν ως πυρήνα την αυτόματη εκτίμηση του κακοπληρωτή χρησιμοποιώντας μεθόδους μηχανικής εκμάθησης. Η βαθμολόγηση μπορεί να χωριστεί σε δύο μέρη. Το πρώτο μέρος είναι η βαθμολόγηση συμπεριφοράς όπου χρησιμοποιούνται δυναμικές διαδικασίες διαχείρισης χαρτοφυλακίου. Σε αυτό λαμβάνονται υπόψη οι τρέχοντες καταναλωτές προκειμένου να εξεταστούν τα ατομικά πρότυπα συμπεριφοράς τους. Το δεύτερο μέρος είναι η βαθμολόγηση 'συλλογής' η οποία προσπαθεί να ταξινομήσει τους πελάτες σε διάφορες συστάδες με βάση την αλλαγή στη συμπεριφορά τους. Γενικά, τα συστήματα πιστωτικής βαθμολόγησης χρησιμοποιούνται για την εκτίμηση της πιθανότητας αθέτησης ενός δανείου. Ωστόσο, ο πρωταρχικός στόχος αυτού του μοντέλου είναι να ανακαλύψει τη βαθμολογία για όλους τους υπάρχοντες πελάτες από ένα χρηματοπιστωτικό ίδρυμα και να βοηθήσει τους δανειστές να διαθέσουν κεφάλαια για έναν αξιόπιστο υποψήφιο αξιοποιώντας τις προβλέψεις του μοντέλου. Το μοντέλο χρησιμοποιεί αλγόριθμο βασισμένο σε βαθιά μάθηση για να εντοπίσει το βάρος εισόδου. Εκεί υπάρχει το ρίσκο κάποια βάρη να οδηγούν σε μεροληπτικά (biased) αποτελέσματα και λόγω της πολυπλοκότητας του μοντέλου να μη γίνονται αντιληπτά με αποτέλεσμα να έχουμε κρυφές μεροληψίες (hidden biases). Οι μεροληψίες είναι μη-επιθυμητές διότι αν ο αλγόριθμος μας εκπαιδευτεί με μεροληψία θα έχουμε πρόβλημα στην γενίκευση του αλγορίθμου και στην πρόβλεψη του σε άγνωστες τιμές. Η μελέτη αυτή χρησιμοποιεί έναν αποτελεσματικό ταξινομητή για τη λήψη έξυπνων αποφάσεων και την εκτέλεση μιας αυστηρής εκπαίδευσης, που οδηγεί εν τέλει στην εκπαίδευση του προγνωστικού μοντέλου. Τα μοντέλα μηχανικής μάθησης εντοπίζουν μοτίβα και σχέσεις από το σύνολο δεδομένων εκπαίδευσης και παρέχουν προβλέψεις για το μέλλον.

Εκτός από το παραπάνω μοντέλο, έχουν, επίσης, χρησιμοποιηθεί και διάφορα άλλα μοντέλα βαθμολόγησης πιστωτικών μονάδων που βασίζονται σε αλγόριθμους μηχανικής

μάθησης. Ειδικότερα, οι Louzada et al. (2016) διεξήγαγαν μια έρευνα μεθόδων δυαδικής ταξινόμησης, για την αξιολόγηση της πιστοληπτικής ικανότητας. Ο κύριος στόχος τους ήταν να προτείνουν μια νέα προσέγγιση για την κατάταξη στη βαθμολογία πιστοληπτικής ικανότητας, ειδικά με υβριδικές τεχνικές, όπου υπάρχει ομοιότητα στα προγνωστικά αποτελέσματα. Οι Liberati και Camillo (2018) εξέτασαν τη χρήση των προσωπικών αξιών/χαρακτηριστικών ως νέα πληροφορία για τη βελτίωση των αξιολογήσεων πιστωτικού κινδύνου. Τα ευρήματά τους έδειξαν ότι οι μη γραμμικοί ταξινομητές μπορούν να προβλέπουν τον πιστωτικό κίνδυνο πολύ καλύτερα σε σχέση με τους KDA (kernel discriminant analysis). Ανακάλυψαν, επίσης, ότι τα ψυχολογικά χαρακτηριστικά βελτιώνουν την αποτελεσματικότητα των προσεγγίσεων προ συμπτωματικού ελέγχου. Κατέληξαν στο συμπέρασμα ότι όταν χρησιμοποιούνται τα οικονομικά δεδομένα και τα τραπεζικά αρχεία των αιτούντων, τα ευρήματα υποδεικνύουν μικρές αλλαγές στα σφάλματα της ταξινόμησης.

Οι Wang et al. (2018) έχουν προτείνει μια υβριδική λύση δύο φάσεων που βασίζεται στο ‘φιλτράρισμα’ και σε έναν γενετικό αλγόριθμο πολλαπλών πληθυσμών (hybrid multiple population genetic algorithm – HMPGA). Οι Zhao et al. (2015) πρότειναν ένα μοντέλο αξιολόγησης πιστοληπτικής ικανότητας υψηλής απόδοσης που βασίζεται σε νευρωνικά δίκτυα πολυστρωματικού αναγνωριστή (multilayer perceptron – MLP). Το μοντέλο MLP έχει 9 μυστικές ‘μονάδες’ που ρυθμίστηκαν χρησιμοποιώντας τον αλγόριθμο του back propagation – BP. Στην έρευνά τους, αυτό το μοντέλο αξιολόγησης πιστοληπτικής ικανότητας κατάφερε να επιτύχει ακρίβεια 87%, η οποία είναι κατά 5% καλύτερη σε σχέση με τα καλύτερα ευρήματα προηγούμενων εργασιών για ένα γερμανικό σύνολο δεδομένων. Οι Ala'raj και Abbod (2016) έχουν προτείνει μια συνδυαστική consensus-based μέθοδο για το συνδυασμό multiple classifier systems - MCS με διαφορετικούς αλγόριθμους συσταδοποίησης. Τα πειραματικά ευρήματα, η ανάλυση και οι συγκριτικές αναλύσεις τους δείχνουν ότι η προτεινόμενη συνδυαστική προσέγγιση υπερέχει όλων των βασικών ταξινομητών. Οι συγγραφείς έχουν επίσης δοκιμάσει/επιβεβαιώσει το μοντέλο χρησιμοποιώντας πέντε πραγματικά (real data) σύνολα δεδομένων πιστωτικών αποτελεσμάτων. Οι Kaveh Bastani et al., έχουν προτείνει μια μέθοδο ευρείας και βαθιάς μάθησης (wide and deep learning) για peer-to-peer δανεισμό. Ένα τέτοιο μοντέλο (wide and deep learning) αποτελείται από δύο κύρια στοιχεία: την συνιστώσα απομνημόνευσης (γραμμικό μοντέλο), ένα στοιχείο γενίκευσης (Νευρωνικό δίκτυο) και ένα γινόμενο/μια

μίξη των δύο προηγούμενων στοιχείων. Τα μοντέλα ευρείας και βαθιάς μάθησης χρησιμοποιούνται πολύ σε συστήματα συστάσεων. Στην εργασία αυτή, οι συγγραφείς έχουν χρησιμοποιήσει τη μέθοδο βαθμολόγησης δύο σταδίων για να 'υποστηρίξουν' τους δανειστές ώστε να αποφασίσουν ποιον είναι προτιμότερο να χρηματοδοτήσουν. Τέλος, οι Abellán και Mantas (2014) παρουσίασαν τα Credal Decision Trees, μια νέα διαδικασία για την κατασκευή δέντρων αποφάσεων που χειρίζεται την ανακρίβεια με διαφορετικό τρόπο συγκριτικά με τις παραδοσιακές διαδικασίες.

Συνοψίζοντας και σύμφωνα με τις προαναφερθέντες μελέτες, γίνεται εμφανές ότι η ταξινόμηση χαρακτηριστικών επιλογής (feature selection - FS) ενισχύει τις επιδόσεις του προγνωστικού μοντέλου και το βοηθά ούτως ώστε να παρέχει τη βέλτιστη λύση. Οι εφαρμογές που βασίζονται σε βαθιά μάθηση παρέχουν έναν σημαντικό τρόπο επίτευξης αποτελεσματικότητας ενώ το σύστημα 'υποστήριξης' αποφάσεων βοηθά στην εύρεση του συνδυασμού ταξινομητών, παρέχοντας πιο ακριβείς και 'έξυπνες' συστάσεις.

5.2 Πίνακας μετάβασης (transition matrix)

Στη συνέχεια, υπάρχει ένας σημαντικός όγκος έρευνας σχετικά με την εφαρμογή του πίνακα μετάβασης (transition matrix) για την εκτίμηση της πιθανότητας αθέτησης (probability of default). Οι πίνακες μετάβασης ουσιαστικά μετρούν τις πιθανότητες μετάβασης από μία βαθμολόγηση πιστοληπτικής ικανότητας (credit score) σε μια άλλη, σε συγκεκριμένα χρονικά διαστήματα. Είναι από τα πιο χρήσιμα εργαλεία που έχει στη διάθεση του ένας διαχειριστής πιστωτικού κινδύνου για την προβολή της εξέλιξης της πιθανότητας αθέτησης (PD). Ειδικότερα, στη εργασία των Aparicio et al. (2013) αναλύθηκε το πιστωτικό χαρτοφυλάκιο του περουβιανού χρηματοπιστωτικού συστήματος λαμβάνοντας υπόψη τη χρήση του πίνακα πιστωτικής μετάβασης που εξαρτάται από τον οικονομικό κύκλο. Όσον αφορά την περίπτωση μιας κολομβιανής μελέτης, οι Támara-Ayús et al. (2012) συνέκριναν διακριτούς και συνεχείς πίνακες μετάβασης, ενώ υπάρχουν επίσης μελέτες με το Ακαθάριστο Εθνικό Προϊόν (Gross National Product – GNP) όπως αυτή που διεξήχθη από τους McCulloch και Tsay (1994). Οι εν λόγω συγγραφείς παρατήρησαν ότι η αβεβαιότητα σχετικά με την κατάσταση μιας δεδομένης περιόδου, εξαρτάται από τις προδιαγραφές του εκάστοτε μοντέλου. Από την πλευρά του, ο Espinoza (2013) καταλήγει στο συμπέρασμα ότι οι οικονομικοί κύκλοι και οι μακροοικονομικές μεταβλητές επηρεάζουν την μείωση της πιστωτικής ποιότητας.

Τέλος και όσον αφορά μια περίπτωση της Βενεζουέλας, οι Porras et al. (2002) πραγματοποίησαν την ανάλυσή τους μέσω μεταβατικών πινάκων για τις εκ των προτέρων και εκ των υστέρων περιόδους της εισροής ξένων κεφαλαίων ούτως ώστε να προσδιορίσουν εάν αυξάνεται ο ανταγωνισμός μεταξύ των χρηματοπιστωτικών ιδρυμάτων.

Περνώντας σε άλλες μελέτες, οι Zhu et al. (2019) ασχολήθηκαν με τα XVA (X-Value-Adjustment). Εν συντομία, XVA είναι ένας γενικός όρος που αναφέρεται στους διαφορετικούς τύπους αξιολόγησης και στις προσαρμογές (adjustments) που πρέπει να κάνουν οι τράπεζες σχετικά με τα συμβόλαια παραγώγων (derivative contracts) που έχουν συνάψει και τα νέα που πρόκειται να συνάψουν. Οι συγγραφείς εφάρμοσαν τεχνικές μηχανικής μάθησης για τον προσδιορισμό και την επίλυση προβλημάτων (X-Value-Adjustment) μέσω της προσομοίωσης Monte Carlo και της μεθόδου συσταδοποίησης K-means. Στην ανάλυση πιστωτικού κινδύνου των Μικρών και Μεσαίων Επιχειρήσεων (Small and Medium Enterprises (SMEs), οι συγγραφείς Wahyudin et al. (2016) διαμόρφωσαν ομάδες κινδύνου χρησιμοποιώντας K-means και μέτρηση κινδύνου, υπολογίζοντας ποιοτικές βαθμολογίες σημασίας σε συνδυασμό με βαθμολογίες συναισθήματος. Το αποτέλεσμα της εν λόγω έρευνας δείχνει ότι το μοντέλο είναι επαρκές για συσταδοποίηση και μέτρηση του επιπέδου κινδύνου. Για μια περίπτωση του Μεξικού, οι Támara-Ayús et al. (2017) παρουσίασαν μια ανάλυση διαφορετικών μελετών σχετικά με τον πιστωτικό κίνδυνο ενός εμπορικού χαρτοφυλακίου. Ανέπτυξαν ένα μοντέλο προκειμένου να προβλέψουν την πιθανότητα αθέτησης υποχρεώσεων ενός οφειλέτη, μέσω μιας παραγοντικής και διακριτικής ανάλυσης. Από την άλλη πλευρά, οι Lagunas και Ramírez (2017) υπολόγισαν διάφορα σενάρια για τον προσδιορισμό του αριθμού των πράξεων που ενδέχεται να λάβουν παράνομες οικονομικές συναλλαγές, μέσω στοχαστικών πινάκων μετάβασης.

5.3 Μη επιβλεπόμενη μάθηση

Μέθοδοι μη επιβλεπόμενης μάθησης και ειδικότερα ομαδοποίησης, εφαρμόζονται για την εκτίμηση των πιθανοτήτων αθέτησης για δάνεια CMBS. Τα τελευταία χρόνια, το ανεξόφλητο υπόλοιπο της αγοράς τίτλων που υποστηρίζονται από εμπορικά στεγαστικά δάνεια έχει φτάσει σε τεράστια ποσά (δεκάδες δισεκατομμύρια ανά έτος) και εξακολουθεί να αυξάνεται. Το μέγεθος της αγοράς αυτών των τίτλων δείχνει την ανάγκη

για μία πιο ακριβή αποτίμηση κάθε δανείου CMBS (complete commercial mortgage backed securities). Ειδικότερα, τα εκτιμώμενα ποσοστά αθέτησης έχουν αντίκτυπο στην εύρεση των προβλεπόμενων αξιών των δανείων. Κινούμενος προς αυτή την κατεύθυνση, ο Yildirim (2008) πρότεινε ένα μοντέλο 'μείγματος' (mixture) με μακροπρόθεσμους επιζώντες (long-term survivors) για την ανάλυση συσταδοποιημένων δεδομένων χρόνου αποτυχίας, σε περιπτώσεις κατά τις οποίες ορισμένα δάνεια ενδέχεται να μην αντιμετωπίσουν ποτέ το συμβάν που βρίσκεται υπό μελέτη. Το πλεονέκτημα αυτού του μοντέλου είναι ότι επιτρέπει την ταυτόχρονη εκτίμηση των επιδράσεων των μεταβλητών τόσο στην πιθανότητα όσο και στον χρόνο ενός γεγονότος. Ένα σημαντικό ποσοστό των δανείων δεν θα αθετηθεί ποτέ και είναι πιθανό ορισμένοι παράγοντες που επηρεάζουν την πιθανότητα αθέτησης πληρωμών, να μην επηρεάζουν το χρονοδιάγραμμα αυτής της αθέτησης και το αντίστροφο. Τα τυπικά μοντέλα επιβίωσης, π.χ. λογιστική παλινδρόμηση και τα αναλογικά μοντέλα κινδύνου που συναντώνται στη μελέτη του Cox (1972), δεν κάνουν σαφή διάκριση μεταξύ αυτών των επιπτώσεων. Αυτά τα μοντέλα χρησιμοποιούνται συχνά για την εκτίμηση της επίδρασης των συμμεταβλητών στην πιθανότητα να συμβεί ένα γεγονός.

Ένα άλλο πρόβλημα αφορά τις λογοκριμένες παρατηρήσεις, δηλαδή τα δάνεια που δεν έχουν υποστεί αθέτηση πληρωμών. Το μοντέλο αναλογικών κινδύνων του Cox (1972) χειρίζεται το ζήτημα της λογοκρισίας υποθέτοντας ότι η λογοκρισία είναι ανεξάρτητη από το συμβάν. Η υπόθεση της ανεξαρτησίας είναι λογική εάν όλα τα δάνεια του δείγματος, αντιμετωπίσουν τελικά αθέτηση υποχρεώσεων. Σε περίπτωση που ένα σημαντικό μέρος του δείγματος δεν αντιμετωπίσει ποτέ αθέτηση των υποχρεώσεων, η λογοκριμένη παρατήρηση θα περιέχει τόσο εκείνους που τελικά θα βιώσουν το συμβάν όσο και εκείνους που δεν θα το βιώσουν. Λόγω του ότι τα δάνεια που δεν θα αντιμετωπίσουν ποτέ αθέτηση υποχρεώσεων θα εμπίπτουν πάντα στην ομάδα που λογοκρίνεται, το γεγονός και η λογοκρισία δε θα πρέπει να εκλαμβάνονται ως ανεξάρτητα. Επομένως, όταν η περίοδος του δείγματος δεν είναι αρκετά μεγάλη και η αθέτηση δεν αντιμετωπίζεται από όλα τα δάνεια, το μοντέλο μείγματος μπορεί να ξεπεράσει αυτό το πρόβλημα.

Το μοντέλο 'μείγματος', λοιπόν, χρησιμοποιείται για την ανάλυση δεδομένων χρόνου αθέτησης, στα δάνεια που θα μπορούσαν τελικά να αντιμετωπίσουν την αθέτηση

πληρωμών κατά τη διάρκεια της περιόδου παρατήρησης. Υπάρχουν δύο λόγοι για τους οποίους δεν αντιμετωπίζεται ένα συμβάν αθέτησης. Ο πρώτος είναι ότι ορισμένα δάνεια μπορεί να αφορούν μακροπρόθεσμους επιζώντες οι οποίοι δεν θα βιώσουν το υπό μελέτη συμβάν, ενώ ο δεύτερος ότι είναι επιρρεπή σε αθέτηση, αλλά σωστά τροποποιημένα στο δείγμα. Τα τυπικά μοντέλα επιβίωσης υποθέτουν ότι όλα τα δάνεια υπόκεινται σε αθέτηση υποχρεώσεων, αλλά τα περισσότερα χρηματοοικονομικά δεδομένα παρουσιάζουν έντονη λογοκρισία. Επίσης το μοντέλο μείγματος χρησιμοποιείται για να ξεπεραστούν οι πιθανές προκαταλήψεις στην εκτίμηση των παραμέτρων, εκτός από τις τυχαίες αδυναμίες για την καταγραφή του φαινομένου ομαδοποίησης. Το μοντέλο μείγματος έχει δύο στοιχεία: ένα που υποδεικνύει εάν θα μπορούσε τελικά να συμβεί η αθέτηση υποχρεώσεων και το δεύτερο που υποδηλώνει την χρονική στιγμή που θα συμβεί το γεγονός. Το πιο σημαντικό πλεονέκτημα που προσφέρει είναι ότι επιτρέπει την ταυτόχρονη εκτίμηση των ξεχωριστών επιδράσεων των συμμεταβλητών στην πιθανότητα και το χρόνο ενός γεγονότος.

5.4 Τεχνικές επιβλεπόμενης μάθησης

Επιπλέον, υπάρχουν ορισμένες τεχνικές επιβλεπόμενης μάθησης, οι οποίες μπορούν να βοηθήσουν στην μοντελοποίηση του πιστωτικού κινδύνου σε περιπτώσεις διαδικτυακού δανεισμού. Η Συμφωνία για το Κεφάλαιο ΙΙ της Βασιλείας (Basel II Capital Accord) απαιτεί από τα χρηματοπιστωτικά ιδρύματα να εκτιμούν την πιθανότητα αθέτησης (default probability - PD) των δανειοληπτών. Έτσι, τα μοντέλα PD είναι εξαιρετικά μελετημένα και συνεχίζουν να προσελκύουν μεγάλο ενδιαφέρον. Ένα από τα παλαιότερα μοντέλα είναι το μοντέλο γραμμικής ανάλυσης διάκρισης (linear discriminant analysis - LDA). Ο (Wiginton, 1980) εφάρμοσε για πρώτη φορά ένα μοντέλο λογιστικής παλινδρόμησης για να αξιολογήσει τον πιστωτικό κίνδυνο και διαπίστωσε ότι αυτό το μοντέλο είχε υψηλή ακρίβεια ταξινόμησης και ισχυρή πρακτικότητα. Στη συνέχεια, οι αναλύσεις λογιστικής παλινδρόμησης έγιναν μια κοινή μέθοδος που χρησιμοποιείται για την αξιολόγηση του πιστωτικού κινδύνου. Δεδομένου ότι ο σκοπός της αξιολόγησης πιστωτικού κινδύνου είναι να διευκολύνει τη λήψη αποφάσεων σχετικά με το εάν θα χορηγηθεί πίστωση σε έναν νέο αιτούντα, είναι πιο σημαντικό να προβλεφθεί ο κίνδυνος αθέτησης παρά να γίνει επεξήγησή του. Επιπροσθέτως, δεδομένου ότι ακόμη και ένα μικρό κλάσμα βελτίωσης θα μπορούσε να αφήσει σημαντικές οικονομίες και κέρδη, είναι απαραίτητο να χρησιμοποιηθούν οποιεσδήποτε μέθοδοι μπορούν να βελτιώσουν

την ικανότητα διάκρισης. Ως εκ τούτου, ένας μεγάλος όγκος μελετών αφιερώνεται στη βελτίωση της ακρίβειας της πρόβλεψης πιστώσεων, μεταξύ των οποίων οι μη παραμετρικές μέθοδοι μηχανικής μάθησης, οι οποίοι θεωρείται ότι επιτυγχάνουν υψηλότερη ακρίβεια. Η μηχανική μάθηση ξεπερνά τους περιορισμούς των τυπικών μοντέλων όπως το logit και το μοντέλο probit και μπορεί να ανιχνεύσει μη γραμμικές αλληλεπιδράσεις μεταξύ των μεταβλητών εισόδου, γεγονός που αυξάνει δραματικά τους τύπους σχέσεων που μπορούν να συλληφθούν και τον αριθμό των ανεξάρτητων μεταβλητών που μπορούν να χρησιμοποιηθούν.

Τα τελευταία χρόνια, περισσότεροι από 20 τύποι μοντέλων ταξινομητών μηχανιής μάθησης έχουν μελετηθεί για την εκτίμηση της πιθανότητας αθέτησης των δανειοληπτών. Μεταξύ αυτών, ο αλγόριθμος k-πλησιέστερου γείτονα (k-nearest neighbor – k-NN), η support vector machine (SVM) και τα δάση τυχαίας απόφασης (random forest - RF) είναι υπολογιστικά γρήγορα και εύκολα στην εφαρμογή τους και έχουν ήδη αποδείξει την πολύ καλή τους απόδοση και τη συνέπεια τους. Ειδικότερα, όσον αφορά τον k-NN, οι Chatterjee και Barcun (1970) ήταν οι πρώτοι που τον εφάρμοσαν στην αξιολόγηση πιστωτικού κινδύνου. Ως τυπικός ταξινομητής μάθησης με απλό μηχανισμό λειτουργίας, ο k-NN έχει χρησιμοποιηθεί ευρέως στην αξιολόγηση πιστωτικού κινδύνου. Όσον αφορά την SVM, οι Van Gestel et al. (2003) ήταν οι πρώτοι που την εφάρμοσαν στην αξιολόγηση πιστοληπτικής ικανότητας και διαπίστωσαν ότι σε σύγκριση με τα μοντέλα νευρωνικών δικτύων, η SVM μπορεί να αποκτήσει μεγαλύτερη ακρίβεια ταξινόμησης. Άλλες μελέτες όπως αυτή των Bellotti και Crook (2009), δείχνουν ότι η SVM αποδίδει αρκετά καλά στις αξιολογήσεις πιστωτικού κινδύνου των αιτούντων πιστωτικών καρτών. Σύμφωνα με τους Yu et al. (2018), η βελτιωμένη μέθοδος SVM είναι ανώτερη από την παραδοσιακή μέθοδο SVM σε ότι έχει να κάνει με την πρόβλεψη της πιθανότητας αθέτησης των δανειοληπτών. Αυτό συμβαίνει διότι έχει καλύτερη ικανότητα γενίκευσης και πρακτική αξία. Όσον αφορά το μοντέλο RF, ο Breiman (2001) πρότεινε τον εν λόγω αλγόριθμο με βάση την ιδέα της 'ολοκλήρωσης' και σημείωσε ότι ένα ισχυρό μοντέλο μάθησης που παράγεται από το συνδυασμό αδύναμων μοντέλων μάθησης θα είναι ανώτερο από ένα μοντέλο δέντρου παλινδρόμησης ή μεμονωμένης ταξινόμησης/συσταδοποίησης. Πιο συγκεκριμένα, η έρευνα των Kruppa et al. (2013) έδειξε ότι το μοντέλο RF ήταν ανώτερο συγκριτικά με αυτό του k-NN ως προς την ανάλυση του καταναλωτικού πιστωτικού κινδύνου.

Επιπλέον, οι Lessmann et al. (2015) συνέκριναν 24 διαφορετικούς αλγόριθμους ταξινόμησης μεταξύ των οποίων, ο αλγόριθμος RF είχε την καλύτερη απόδοση. Από την άλλη πλευρά, η έρευνα των Arora και Kaur (2020) έδειξε ότι το βελτιωμένο RF είναι ανώτερο από την SVM και τον k-NN, όσον αφορά την πρόβλεψη της πιθανότητας αθέτησης του δανειολήπτη.

5.5 Αλγόριθμοι βαθιάς μάθησης

Διάφοροι αλγόριθμοι βαθιάς μάθησης έχουν εφαρμοστεί κατά καιρούς για την πρόβλεψη της πιθανότητας αθέτησης. Πέρα από τις μεθόδους που αναφέρθηκαν σε προηγούμενο κεφάλαιο για την βαθμολόγηση της πιστοληπτικής ικανότητας, οι Ha και Nguyen (2016) και Luo et al. (2017) προσδιόρισαν δύο αρχιτεκτονικές βαθιάς μάθησης με ένα πολυστρωματικό αναγνωριστικό δίκτυο (Multilayer Perceptron Network - MLP) και ένα δίκτυο βαθιάς πεποίθησης (deep belief network - DBN) αντίστοιχα και τις εφάρμοσαν για να προβλέψουν την πιθανότητα αθέτησης. Ενώ οι Gunnarsson et al. (2021) πραγματοποίησαν τις προγνωστικές τους επιδόσεις σε μια έρευνα ανασκόπησης και υποστήριξαν ότι οι υπάρχοντες αλγόριθμοι βαθιάς μάθησης δεν ήταν σημαντικά καλύτεροι από εκείνους του μοντέλου ‘ολοκληρωμένης’ μάθησης.

5.6 Τρόποι αξιολόγησης μοντέλων

Κλείνοντας την βιβλιογραφική ανασκόπηση, οι περισσότερες έρευνες που αφορούν την μηχανική μάθηση στην οικονομετρία, κρίνουν τα πλεονεκτήματα και τα μειονεκτήματα των μοντέλων συγκρίνοντας την προβλεπόμενη απόδοση τους. Κατά την αξιολόγηση των διαφορών μεταξύ των προβλέψεων εκτός δείγματος του μοντέλου και των πραγματικών αποτελεσμάτων, υπάρχουν τρεις βασικοί τύποι κριτηρίων: Η AUC (Area under the ROC Curve), η βαθμολογία Brier και η ακρίβεια (accuracy). Αυτοί χρησιμοποιούνται γενικά για την αξιολόγηση της διακριτικής ικανότητας (discriminatory power), της ορθότητας των κατηγορικών προβλέψεων και της ακρίβειας των προβλέψεων πιθανοτήτων. Οι συγγραφείς Pencina et al. (2008) πρότειναν την μέθοδο IDI (Integrated Discrimination Improvement) για να ελέγξουν εάν υπάρχει σημαντική διαφορά μεταξύ των αποτελεσμάτων πρόβλεψης των δύο μοντέλων όσον αφορά την αξιολόγηση της βελτίωσης του μοντέλου σε σύγκριση με το μοντέλο αναφοράς. Ωστόσο, οι παραπάνω δείκτες δεν λαμβάνουν υπόψη το κόστος των σφαλμάτων τύπου ΙΙ στα αποτελέσματα πρόβλεψης. Σφάλμα τύπου ΙΙ , είναι η αδυναμία να απορριφθεί μια ψευδής μηδενική υπόθεση ("ψευδές αρνητικό", δηλαδή, η αποδοχή λανθασμένης

υπόθεσης, θεωρώντας την σωστή). Στην πραγματικότητα, οι απώλειες των σφαλμάτων τύπου ΙΙ είναι σημαντικά υψηλες. Οι Lessmann et al. (2015) επεσήμαναν ότι εκτός από την απαίτηση πιο περίπλοκων μεθόδων, όπως η ανάλυση επιβίωσης και οι διαδικασίες Markov, απαιτούνται επίσης πρόσθετα δεδομένα για την εκτίμηση του χρόνου εμφάνισης της παραβίασης, της ζημιάς που έχει οριστεί ως προεπιλογή (Loss Given Default - LGD) και της έκθεσης σε προεπιλογή (Exposure at Default - EAD).

6 Εμπειρική μελέτη

6.1 Δεδομένα

Σε αυτή την διπλωματική εργασία για να εξεταστεί η έννοια του πιστωτικού κινδύνου και συγκεκριμένα η πιθανότητα αθέτησης πληρωμής θα χρησιμοποιηθεί το σύνολο δεδομένων «default of credit card clients Data Set» το οποίο είναι δημόσια διαθέσιμο στο αποθετήριο UCI.

Αυτό το σύνολο δεδομένων περιέχει 30000 δείγματα και 24 μεταβλητές σχετικά με την πληρωμή ή αθέτηση πληρωμών πιστωτικών καρτών στην Ταϊβάν. Η μεταβλητή που πρέπει να προβλεφθεί είναι η πτώχευση και έχει δύο τιμές (0/1). Οι υπόλοιπες μεταβλητές είναι:

LIMIT_BAL: Ποσό πίστωσης (δολάρια): περιλαμβάνει τόσο την ατομική όσο και την οικογενειακή (συμπληρωματική) πίστωση.

SEX: Φύλο (1 = αρσενικό, 2 = θηλυκό).

EDUCATION: Εκπαίδευση (1 = μεταπτυχιακό, 2 = πανεπιστήμιο, 3 = λύκειο, 4 = άλλα).

MARRIAGE: Οικογενειακή κατάσταση (1 = έγγαμος, 2 = άγαμος, 3 = άλλοι).

AGE: Ηλικία (έτος).

PAY_0 - PAY_6: Ιστορικό προηγούμενων πληρωμών. Αρχείο προηγούμενων μηνιαίων πληρωμών (από τον Απρίλιο έως τον Σεπτέμβριο του 2005) ως εξής:
PAY_0 = η κατάσταση αποπληρωμής τον Σεπτέμβριο του 2005. PAY_1 = η κατάσταση αποπληρωμής τον Αύγουστο του 2005. . . .; PAY_6 = η κατάσταση αποπληρωμής τον Απρίλιο του 2005.
Η κλίμακα μέτρησης για την κατάσταση αποπληρωμής είναι: -1 = κανονική πληρωμή. 1 = καθυστέρηση πληρωμής για ένα μήνα. 2 = καθυστέρηση πληρωμής για δύο μήνες. . . .; 8 = καθυστέρηση πληρωμής για οκτώ μήνες. 9 = καθυστέρηση πληρωμής για εννέα μήνες και άνω.

BILL_AMT1 - BILL_AMT6: Ποσό bill statement (δολάρια). X12 = ποσό bill statement τον Σεπτέμβριο, 2005. BILL_AMT1 = ποσό bill statement τον Αύγουστο, 2005. . . .; BILL_AMT6 = ποσό bill statement τον Απρίλιο του 2005.

PAY_AMT1 - PAY_AMT6: Ποσό προηγούμενης πληρωμής (NT δολάριο). PAY_AMT1 = ποσό που καταβλήθηκε τον Σεπτέμβριο του 2005. PAY_AMT2 = ποσό που καταβλήθηκε τον Αύγουστο του 2005. . . .; PAY_AMT6 = ποσό που καταβλήθηκε τον Απρίλιο του 2005.

Σε αυτά τα δεδομένα θα εφαρμοστούν οι αλγόριθμοι μη επιβλεπόμενης μηχανικής μάθησης που παρουσιάστηκαν στα προηγούμενα κεφάλαια. Συγκεκριμένα, θα εφαρμοστούν PCA, MCA και FAMD για να μειωθεί ο αριθμός των μεταβλητών. Στη συνέχεια θα πραγματοποιηθεί συσταδοποίηση με τους αλγορίθμους kmeans και Hierarchical Clustering, των οποίων τα αποτελέσματα θα συγκριθούν. Η υλοποίηση θα γίνει σε γλώσσα προγραμματισμού Python, η οποία έχει πολλές βιβλιοθήκες που είναι κατάλληλες για διαχείριση δεδομένων, μηχανική μάθηση και απεικόνιση δεδομένων.

6.2 Υλοποίηση

Πρώτο βήμα της υλοποίησης είναι να γίνει load το dataset και να γίνει μια αρχική επισκόπηση των μεταβλητών και των τιμών τους. Συγκεκριμένα, φαίνεται να υπάρχουν 24 μεταβλητές και 30.000 παρατηρήσεις.

Καθ' όλη την υλοποίηση, για την αποδοτικότερη διαχείριση και ανάλυση των δεδομένων χρησιμοποιούνται οι παρακάτω βιβλιοθήκες της Python

Πίνακας 1: Python και βιβλιοθήκες

Βιβλιοθήκη	**Χρησιμότητα**
Numpy	Πίνακες, γραμμική άλγεβρα
Pandas	Δομές δεδομένων, ανάλυση δεδομένων
Sklearn	Μηχανική μάθηση
Matplotlib	Οπτικοποίηση (visualization)
Seaborn	Στατιστικές γραφικές

Prince	Στατιστική ανάλυση συνιστωσών (Statistical factor analysis)

Αρχικά γίνεται μια αρχική εξερεύνηση του dataset (exploratory analysis).

Η μεταβλητή που πρέπει να προβλεφθεί, η πιθανότητα να συμβεί αθέτηση πληρωμής (Y) από τον εκάστοτε πελάτη είναι δυαδική (παίρνει τιμές 0 και 1). Παρακάτω φαίνεται ο αριθμός των εγγραφών για κάθε μία από τις δύο τιμές προκειμένου να υπάρχει μια συνολική εικόνα της πιθανότητας αθέτησης.

Πίνακας 2: Διερεύνηση μεταβλητής πρόβλεψης

Τιμή μεταβλητής Y	**Αριθμός εγγραφών**
0	23.364
1	6.636

Ξεκινώντας από τις κατηγορηματικές μεταβλητές, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3 ,PAY_4, PAY_5 και PAY_6 και με την εντολή describe() παρατηρείται ότι οι μοναδικές (unique) τιμές που έχουν δεν αντιστοιχούν στην επίσημη περιγραφή του dataset. Πιο συγκεκριμένα, για την μεταβλητή EDUCATION μη έγκυρες τιμές είναι οι 0,5 και 6, για την μεταβλητή MARRIAGE μη έγκυρη τιμή είναι το 0. Σύμφωνα με την περιγραφή των μεταβλητών στις έγκυρες τιμές και των δύο αυτών μεταβλητών περιλαμβάνεται η κατηγορία ‘άλλο’ με την τιμή 4 για τη μεταβλητή EDUCATION και την τιμή 3 για τη μεταβλητή MARRIAGE. Στους Πίνακας 3 και Πίνακας 4 παρουσιάζονται οι κατανομές των παρατηρήσεων για τις δύο αυτές μεταβλητές. Με κόκκινο σημειώνονται οι μη έγκυρες τιμές.

Πίνακας 3: Κατανομή των παρατηρήσεων για τη μεταβλητή EDUCATION

EDUCATION		
Τιμή	**Κατηγορία**	**Αριθμός παρατηρήσεων**
0	-	14
1	‘Μεταπτυχιακό’	10.585
2	‘Πανεπιστήμιο’	14.030

3	‘Λύκειο’	4.917
4	‘Άλλο’	123
5	-	280
6	-	51

Πίνακας 4: Κατανομή των παρατηρήσεων για τη μεταβλητή MARRIAGE

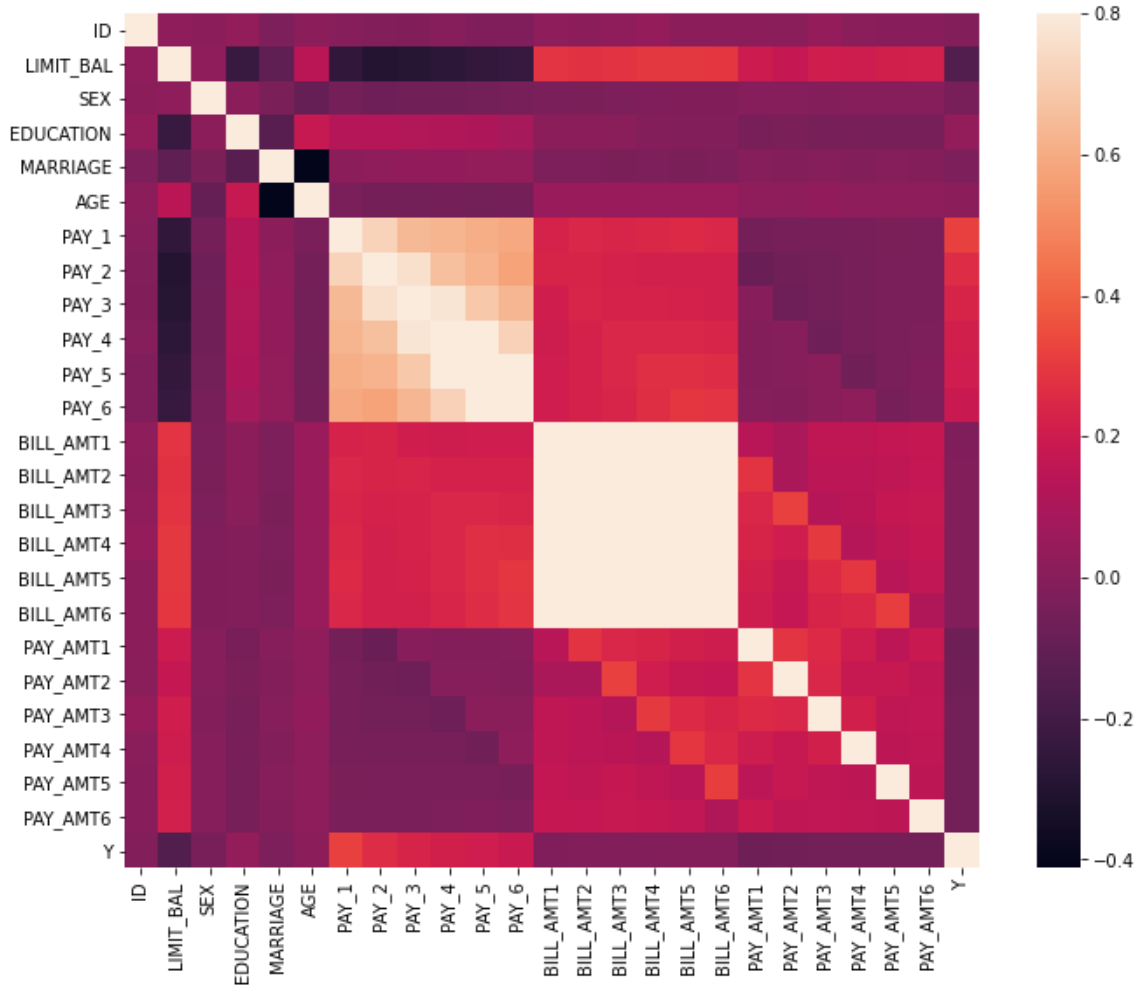
MARRIAGE		
Τιμή	**Κατηγορία**	**Αριθμός παρατηρήσεων**
0	-	54
1	‘Έγγαμος’	13.659
2	‘Άγαμος	15.964
3	‘Άλλο	323

Δεδομένου ότι υπάρχει ένας αξιόλογος αριθμός παρατηρήσεων με μη έγκυρες τιμές στις δύο αυτές μεταβλητές, γίνεται η παραδοχή/υπόθεση (assumption) ότι ανήκουν στην κατηγορία ‘άλλο’, ώστε να μη θεωρηθούν άκυρες και διαγραφούν από το dataset.

Περνώντας στις μεταβλητές PAY_0, PAY_2, PAY_3 ,PAY_4, PAY_5 και PAY_6, παρατηρείται αρχικά ότι η αρίθμηση των μεταβλητών είναι λάθος. Προκειμένου να είναι πιο λογική και να αποφευχθεί ενδεχόμενο λάθος, η μεταβλητή PAY_0 μετονομάζεται σε PAY_1. Στη συνέχεια, σύμφωνα με την περιγραφή των μεταβλητών του dataset, οι μεταβλητές αυτές παίρνουν τιμές από -1 έως 9. Παρατηρείται ότι σε όλες τις μεταβλητές η μικρότερη τιμή είναι το -2 και η μεγαλύτερη το 8. Προκειμένου και πάλι να μην υπάρχουν άκυρες τιμές, γίνεται ανακατανομή (rescaling) των τιμών των μεταβλητών στο εύρος -1 έως 9, ώστε να είναι σύμφωνα με την επίσημη περιγραφή του dataset.

Επόμενο βήμα είναι η διερεύνηση της συσχέτισης των μεταβλητών. Για τον λόγο αυτό, υπολογίζεται ο παρακάτω πίνακας συσχέτισης, Πίνακας 5.

Πίνακας 5: Πίνακας συσχέτισης (correlation matrix)



Παρατηρείται ότι μεταξύ των μεταβλητών BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6 υπάρχει πολύ ισχυρή συσχέτιση, πάνω από 0.8%. Η συσχέτιση μεταξύ των παραπάνω είναι τόσο έντονη που μπορεί να υποδηλώνει μία κατάσταση πολυσυγγραμικότητας (multicollinearity) - χαρακτηριστικά που πρακτικά "δίνουν" την ίδια πληροφορία. Για τον λόγο αυτό, και για να μειωθεί η διάσταση του dataset, αφαιρούνται οι μεταβλητές BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6 εκτός από την BILL_AMT1 (ώστε να μη χαθεί η πληροφορία).

Περνώντας στις συνεχείς αριθμητικές μεταβλητές, γίνεται μια προσπάθεια να κανονικοποιηθούν στο εύρος (0,1). Πριν από αυτό το βήμα, κρατάται ένα αντίγραφο του dataset όπως αυτό είναι σε αυτό το στάδιο ώστε σε μεταγενέστερο βήμα να γίνει εφαρμογή των αλγορίθμων και στα μη κανονικοποιημένα δεδομένα και να συγκριθούν

τα αποτελέσματα. Παρατηρείται ότι η μεταβλητή LIMIT_BAL παίρνει πολύ μεγάλες τιμές και για το λόγο αυτό, κανονικοποιείται μέσω της συνάρτησης του λογαρίθμου. Η κανονικοποίηση μέσω του λογαρίθμου βοηθά ώστε οι μεταβλητές με πολύ 'λοξή – skew' κατανομή να γίνουν λιγότερο 'λοξές – skewed'. Οι υπόλοιπες μεταβλητές, BILL_AMT1, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6 μετατρέπονται μέσω του StandardScaler της βιβλιοθήκης Sklean. Ο StandardScaler εφαρμόζει την κανονικοποίηση $Z = (X-\mu)/\sigma$, οπου μ η μέση τιμή και σ η τυπική απόκλιση των δειγμάτων.

Επόμενο βήμα της προεπεξεργασίας των δεδομένων είναι η αφαίρεση πιθανών ακραίων τιμών (outliers) των συνεχών αριθμητικών μεταβλητών. Αυτό γίνεται μέσω του στατιστικού z-score της βιβλιοθήκης scipy. Το z-score, που αναφέρεται επίσης ως standard score, τιμή z ή και normal score, μεταξύ άλλων, είναι μια ποσότητα χωρίς διαστάσεις (dimensionless quantity), μια τιμή δηλαδή που χρησιμοποιείται για να υποδείξει κατά 'πόσες τυπικές αποκλίσεις' μία παρατήρηση απέχει από τη μέση τιμή της μεταβλητής στην οποία ανήκει. Οι τιμές πάνω από τον μέσο όρο έχουν θετικά z-score, ενώ οι τιμές κάτω από τον μέσο όρο έχουν αρνητικά z-score. Σύμφωνα με τη βιβλιογραφία, ο εμπειρικός κανόνας υποδεικνύει ότι αν το z-score μιας παρατήρησης είναι πάνω από 3, τότε θεωρείται outlier. Η εφαρμογή αυτής της στατιστικής μεθόδου δείχνει ως αποτέλεσμα 2.479 παρατηρήσεις με ακραίες τιμές. Για την καλύτερη απόδοση των αλγορίθμων λοιπόν αφαιρούνται αυτές οι παρατηρήσεις από τα datasets (υπάρχουν πλέον δύο datasets, ένα με τις αρχικές τιμές των μεταβλητών και ένα με τις κανονικοποιημένες). Επίσης, δεδομένου ότι η μεταβλητή ID δεν προσδίδει κάποια πληροφορία σχετικά με την πιθανότητα αθέτησης, αφαιρείται από τα datasets.

Στη συνέχεια, γίνεται διαχωρισμός των μεταβλητών των datasets σε κατηγορηματικές και συνεχείς αριθμητικές μεταβλητές. Πλέον υπάρχουν 4 σύνολα. Επίσης, διαχωρίζεται η μεταβλητή Y η οποία δηλώνει την αθέτηση και είναι αυτή που πρέπει να προβλεφθεί.

Σε αυτό το σημείο έχει τελειώσει η προεπεξεργασία των δεδομένων.

Συνολικά γίνονται οι παρακάτω δοκιμές για τη μορφή των δεδομένων που μπαίνουν σαν είσοδος στους αλγορίθμους:

-το dataset μετά την προεπεξεργασία που περιγράφηκε παραπάνω

-το dataset μετά την κανονικοποίηση ορισμένων μεταβλητών που περιγράφηκε παραπάνω

-PCA στις συνεχείς αριθμητικές μεταβλητές

-one hot encode στις κατηγορηματικές μεταβλητές ώστε να γίνει PCA σε όλες (συνεχείς και τροποποιημένες κατηγορηματικές)

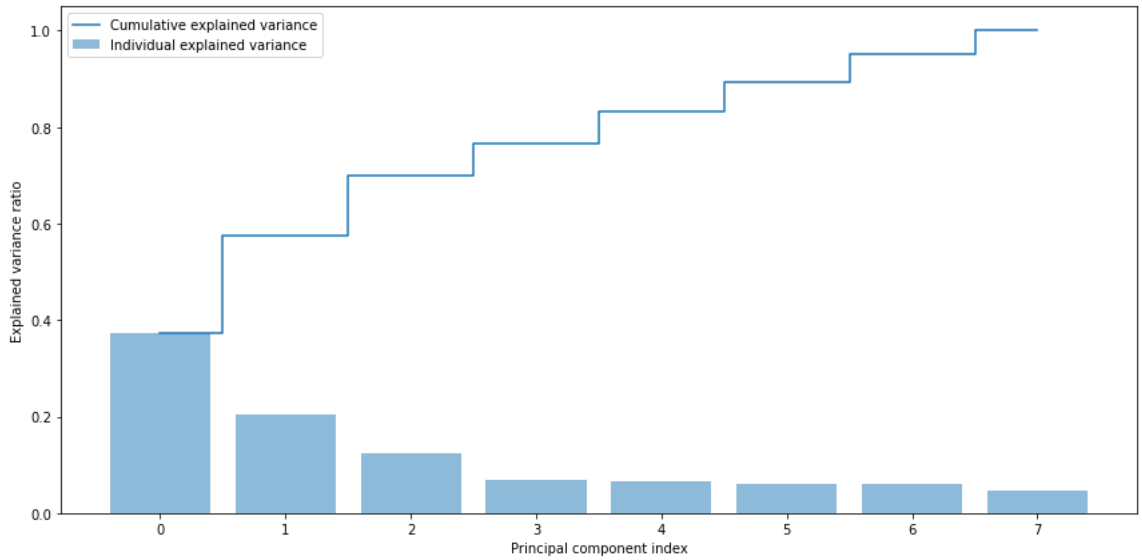
-MCA στις κατηγορηματικές

-FAMD σε όλες τις μεταβλητές

Για τις δύο πρώτες περιπτώσεις τα δεδομένα έχουν ήδη δημιουργηθεί.

Για την τρίτη περίπτωση, γίνεται εφαρμογή της PCA στις 9 συνεχείς αριθμητικές κανονικοποιημένες μεταβλητές του dataset. Για να βρεθεί ο κατάλληλος αριθμός συνιστωσών (components) χρησιμοποιείται το Explained Variance Ratio. Το Explained Variance Ratio είναι το ποσοστό διακύμανσης που αποδίδεται στον καθένα από τους επιλεγμένους components. Στην ιδανική περίπτωση, θα επιλέγεται ο αριθμός των componets που θα συμπεριληφθούν στο μοντέλο προσθέτοντας τη διασπορά που αντιστοιχεί σε κάθε component μέχρι να φτάσει το άθροισμα περίπου 0,8 ή 80% για να αποφευχθεί overfitting.

Εικόνα 4: Explained Variance Ratio - Cumulative Variance



Αφού έχει βρεθεί ο κατάλληλος αριθμός components εφαρμόζεται η PCA από την οποία προκύπτουν οι principal components που θα χρησιμοποιηθούν στους αλγορίθμους.

Για την τέταρτη περίπτωση, εφαρμόζεται η τεχνική one hot encoding στις κατηγορηματικές μεταβλητές. Πιο συγκεκριμένα, εφαρμόζεται στις μεταβλητές EDUCATION, MARRIAGE και SEX. Η διαδικασία του one hot encoder μετατρέπει τις κατηγορηματικές αυτές μεταβλητές σε δυαδικές (μορφή: αριθμητικές 1.0 και 0.0). Παρακάτω παρουσιάζεται ως παράδειγμα η μετατροπή της μεταβλητής EDUCATION, Πίνακας 6 και Πίνακας 7.

Πίνακας 6: Αρχική μορφή της μεταβλητής EDUCATION

Μεταβλητή	EDUCATION
Τιμές	4
	2
	2
	1
	3

Πίνακας 7: One hot encoding στη μεταβλητή EDUCATION

Μεταβλητή	EDUCATION_1.0	EDUCATION_2.0	EDUCATION_3.0	EDUCATION_4.0
Τιμές	0	0	0	1
	0	1	0	0
	0	1	0	0
	1	0	0	0
	0	0	1	0

Αντίστοιχα μετατρέπονται και οι άλλες δύο μεταβλητές.

Δεδομένου ότι η διαδικασία one hot encoding έγινε στο αρχικό dataset, ξαναγίνεται κανονικοποίηση των τιμών των μεταβλητών ώστε να είναι δυνατή η υλοποίηση της PCA.

Για την τέταρτη και πέμπτη περίπτωση εφαρμόζονται αντίστοιχα οι MCA και FAMD στα αντίστοιχα σύνολα δεδομένων (κατηγορηματικές για την MCA και όλες οι μεταβλητές για την FAMD).

Σε αυτό το στάδιο τα δεδομένα σε όλες τις πιθανές μορφές με τις οποίες θα χρησιμοποιηθούν σαν είσοδος στους αλγορίθμους είναι έτοιμα. Οι δοκιμές που γίνονται είναι οι εξής:

k-Means

- Μόνο numerical από το αρχικό dataset (χωρις κανονικοποίηση)
- Μόνο numerical κανονικοποιημένα
- PCA στις numerical (κανονικοποιημένες όπως απαιτεί ο αλγόριθμος)
- One hot encoder στις κατηγορηματικές και PCA σε όλες (κανονικοποιημένες όπως απαιτεί ο αλγόριθμος)
- MCA στις categorical
- FAMD

Hierarchical (Single linkage – Complete linkage – Ward linkage)

- Μόνο numerical από το αρχικό dataset (χωρις κανονικοποίηση)
- Μόνο numerical κανονικοποιημένα
- PCA στις numerical (κανονικοποιημένες όπως απαιτεί ο αλγόριθμος)
- One hot encoder στις κατηγορηματικές και PCA σε όλες (κανονικοποιημένες όπως απαιτεί ο αλγόριθμος)
- MCA στις categorical
- FAMD

Για λόγους σύγκρισης υλοποιούνται και δύο αλγόριθμοι εποπτευόμενης μάθησης, ο αλγόριθμος Random Forest και η Logistic Regression η οποία όπως έχει αναφερθεί σε προηγούμενα κεφάλαια είναι η κύρια μέθοδος που χρησιμοποιείται από τα χρηματοπιστωτικά ιδρύματα για την μοντελοποίηση του πιστωτικού κινδύνου. Δεδομένου ότι δεν αποτελούν μέρος της παρούσας έρευνας οι δύο αυτοί αλγόριθμοι δεν

αναλύονται σε βάθος και γίνεται απλώς μια σύγκριση της απόδοσης των αλγορίθμων ως προς την πρόβλεψη της πιθανότητας αθέτησης.

Επιπλέον, όλες οι δοκιμές έγιναν και στο dataset χωρίς την αφαίρεση των outliers για να αξιολογηθεί αν χάνεται χρήσιμη πληροφορία και κακώς αφαιρούνται.

7 Αποτελέσματα – Συμπεράσματα

Όπως αναφέρθηκε, η μεταβλητή που πρέπει να προβλεφθεί είναι η πιθανότητα να συμβεί αθέτηση πληρωμής (Y) από τον εκάστοτε πελάτη και είναι δυαδική (παίρνει τιμές 0 και 1). Στον Πίνακας 2 φαίνεται ότι ο αριθμός των εγγραφών για κάθε μία από τις δύο τιμές δεν είναι ισόποσα κατανεμημένος. Πιο συγκεκριμένα, το 77.88% του συνόλου δεδομένων (23.364 εγγραφές) έχει τιμή Y μηδέν και μόνο το 22.12% έχει τιμή Y το ένα (6.636 εγγραφές).

Το γεγονός αυτό έχει επίπτωση στα αποτελέσματα των αλγορίθμων καθώς δεν επιτυγχάνουν ξεκάθαρα clusters για τις εγγραφές που θα προβούν σε αθέτηση πληρωμών, αφού είναι πολύ λιγότερες και σε πλήθος αλλά και ποσοστιαία από αυτές που δε θα προβούν σε αθέτηση πληρωμών. Έγιναν πολλές δοκιμές επιλέγοντας διαφορετικό αριθμό clusters κάθε φορά. Ο μικρότερος δυνατός αριθμός clusters ώστε κάποια δοκιμή να καταφέρει να δημιουργήσει ένα τουλάχιστον cluster με πλειοψηφία των αντίστοιχων Y να είναι 1, είναι 10 clusters.

Παρακάτω λοιπόν παρουσιάζονται τα αποτελέσματα για όλες τις προαναφερθέντες δοκιμές με αριθμό clusters 10 και <u>έχοντας αφαιρέσει τις ακραίες τιμές.</u> Επιπλέον, παρουσιάζονται και τα γραφήματα κάθε διαφορετικής δοκιμής που αναπαριστούν τα ποσοστά συγκέντρωσης ως προς τον συνολικό αριθμό των εγγραφών, ανάλογα με την τιμή Y που τους αντιστοιχεί, σε κάθε cluster (“ποσοστό συγκέντρωσης 0 και 1 σε κάθε cluster”, δηλαδή, αριθμός των 0 στο εκάστοτε cluster / συνολικό αριθμό 0 και αντίστοιχα αριθμός των 1 στο εκάστοτε cluster/ συνολικό αριθμό 1). Σημειώνεται ότι στα γραφήματα που ακολουθούν με γαλάζιο αναπαρίστανται τα ποσοστά που αφορούν την τιμή Y = 0, δηλαδή εγγραφές που δε θα προβούν σε αθέτηση ενώ με πράσινο

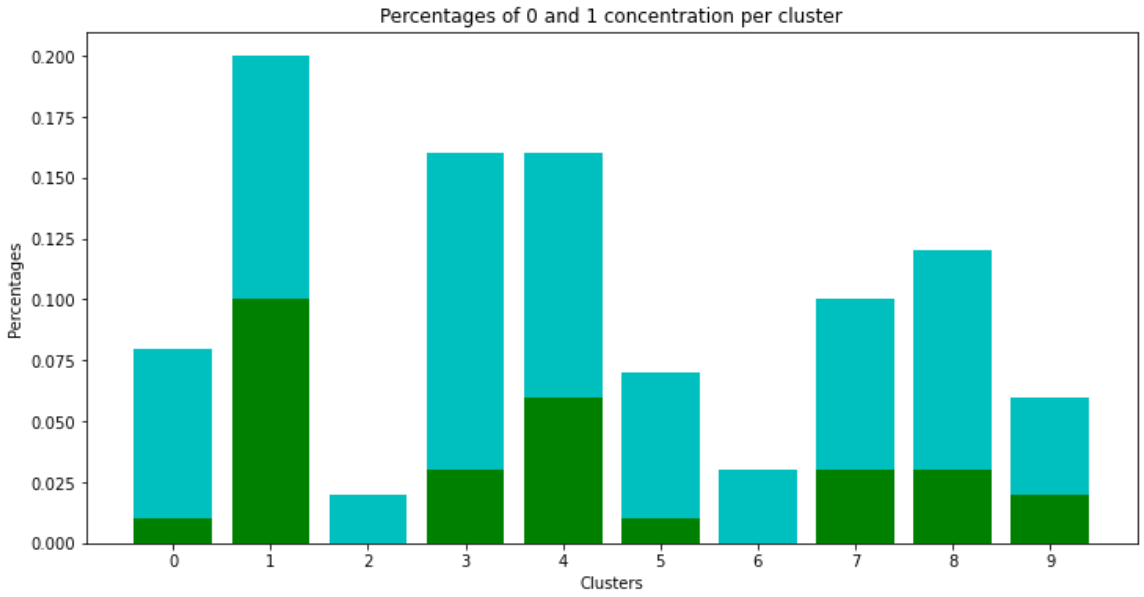
αναπαρίστανται τα ποσοστά που αφορούν την τιμή Y = 1, δηλαδή εγγραφές που θα προβούν σε αθέτηση.

Kmeans

Πίνακας 8: Kmeans simple

Kmeans simple
Το cluster 0 αντιστοιχεί σε 0 με: 1761 -> 0 και 282 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.04
Το cluster 1 αντιστοιχεί σε 0 με: 4215 -> 0 και 2087 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.2 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.33
Το cluster 2 αντιστοιχεί σε 0 με: 407 -> 0 και 53 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 3 αντιστοιχεί σε 0 με: 3434 -> 0 και 719 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.16 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 4 αντιστοιχεί σε 0 με: 3319 -> 0 και 1211 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.16 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.19
Το cluster 5 αντιστοιχεί σε 0 με: 1416 -> 0 και 241 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.04
Το cluster 6 αντιστοιχεί σε 0 με: 728 -> 0 και 104 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.03 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02
Το cluster 7 αντιστοιχεί σε 0 με: 2030 -> 0 και 562 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 8 αντιστοιχεί σε 0 με: 2590 -> 0 και 706 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 9 αντιστοιχεί σε 0 με: 1327 -> 0 και 329 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.05

Εικόνα 5: Kmeans simple

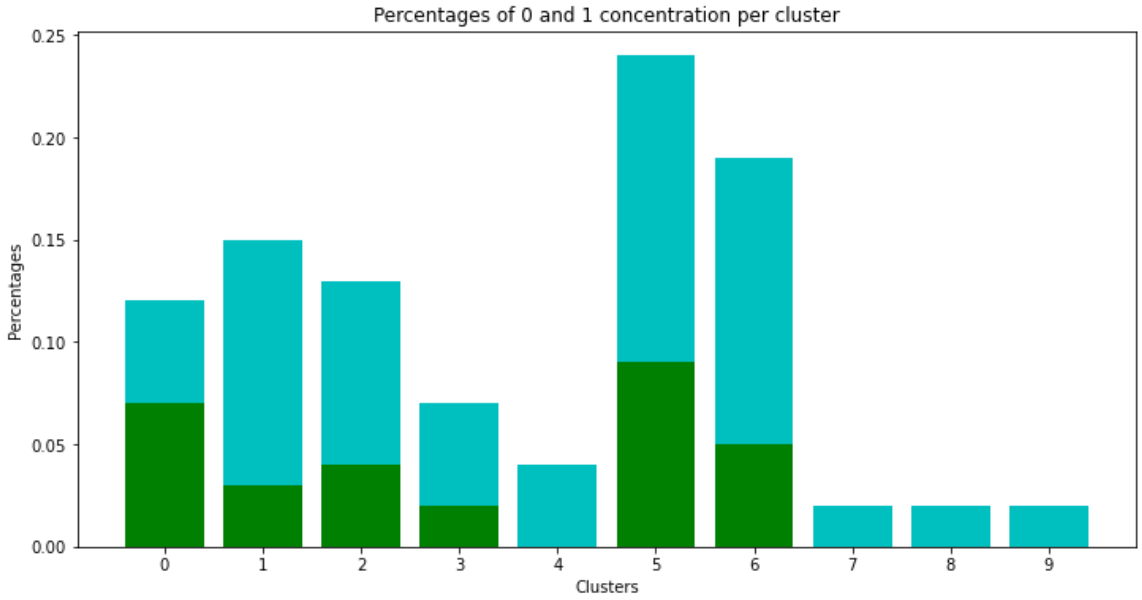


Πίνακας 9: Kmeans normalization

Kmeans normalization
Το cluster 0 αντιστοιχεί σε 0 με: 2609 -> 0 και 1464 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.23
Το cluster 1 αντιστοιχεί σε 0 με: 3172. -> 0 και 589 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 2 αντιστοιχεί σε 0 με: 2829 -> 0 και 754 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.13 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.12
Το cluster 3 αντιστοιχεί σε 0 με: 1542 -> 0 και 352 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06
Το cluster 4 αντιστοιχεί σε 0 με: 780 -> 0 και 69 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 5 αντιστοιχεί σε 0 με: 5080 -> 0 και 1867 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.24 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.3
Το cluster 6 αντιστοιχεί σε 0 με: 4079 -> 0 και 1061 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.19 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.17
Το cluster 7 αντιστοιχεί σε 0 με: 376 -> 0 και 40 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που

συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 8 αντιστοιχεί σε 0 με: 410 -> 0 και 44 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 9 αντιστοιχεί σε 0 με: 350 -> 0 και 54 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01

Εικόνα 6: Kmeans normalization

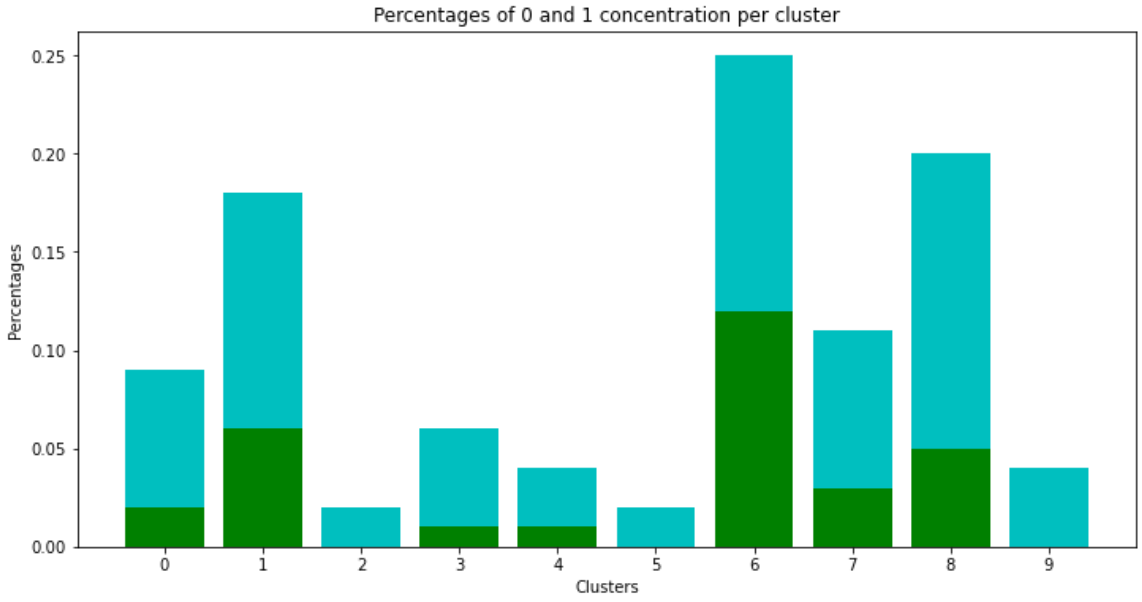


Πίνακας 10: Kmeans PCA

Kmeans PCA
Το cluster 0 αντιστοιχεί σε 0 με: 1987 -> 0 και 347 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06
Το cluster 1 αντιστοιχεί σε 0 με: 3741-> 0 και 1273 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.18 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.2
Το cluster 2 αντιστοιχεί σε 0 με: 402 -> 0 και 14 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 3 αντιστοιχεί σε 0 με: 1268 -> 0 και 295 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.05
Το cluster 4 αντιστοιχεί σε 0 με: 840 -> 0 και 127 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02

Το cluster 5 αντιστοιχεί σε 0 με: 404 -> 0 και 70 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 6 αντιστοιχεί σε 0 με: 5284 -> 0 και 2511 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.25 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.4
Το cluster 7 αντιστοιχεί σε 0 με: 2274 -> 0 και 563 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.11 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 8 αντιστοιχεί σε 0 με: 4148 -> 0 και 993 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.2 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.16
Το cluster 9 αντιστοιχεί σε 0 με: 879 -> 0 και 101 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02

Εικόνα 7: Kmeans PCA

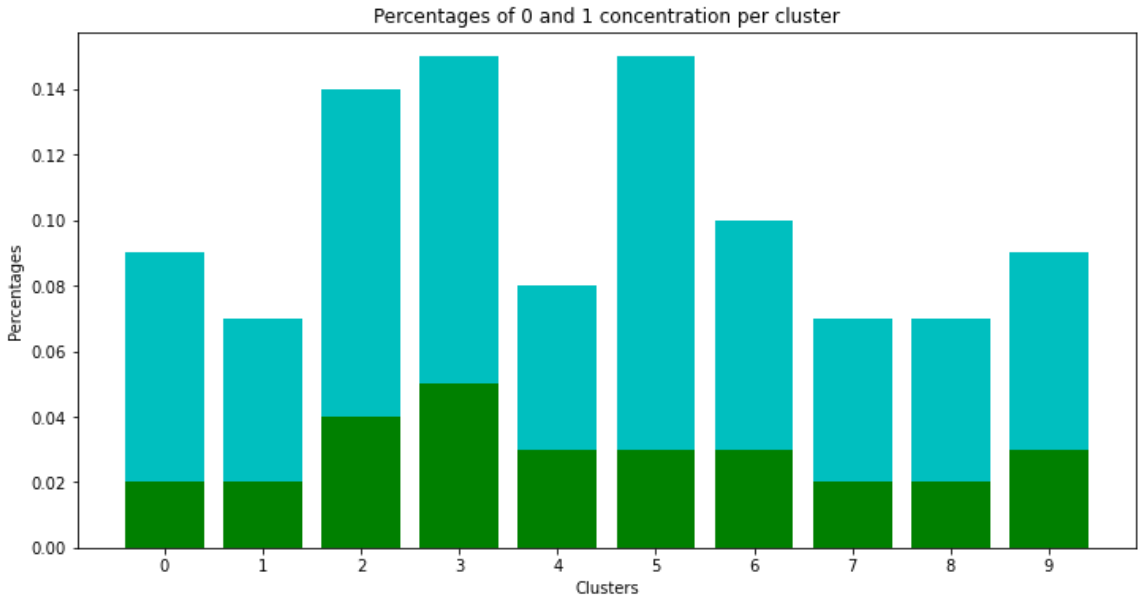


Πίνακας 11: Kmeans one-hot-encode and PCA

Kmeans one-hot-encode and PCA
Το cluster 0 αντιστοιχεί σε 0 με: 1934 -> 0 και 492 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.08
Το cluster 1 αντιστοιχεί σε 0 με: 1417 -> 0 και 458 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.07

Το cluster 2 αντιστοιχεί σε 0 με: 2978 -> 0 και 839 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.14 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.13
Το cluster 3 αντιστοιχεί σε 0 με: 3234 -> 0 και 984 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.16
Το cluster 4 αντιστοιχεί σε 0 με: 1595-> 0 και 627 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1
Το cluster 5 αντιστοιχεί σε 0 με: 3226-> 0 και 718 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 6 αντιστοιχεί σε 0 με: 2068 -> 0 και 720 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 7 αντιστοιχεί σε 0 με: 1381 -> 0 και 448 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.07
Το cluster 8 αντιστοιχεί σε 0 με: 1550 -> 0 και 358 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06
Το cluster 9 αντιστοιχεί σε 0 με: 1844 -> 0 και 650 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1

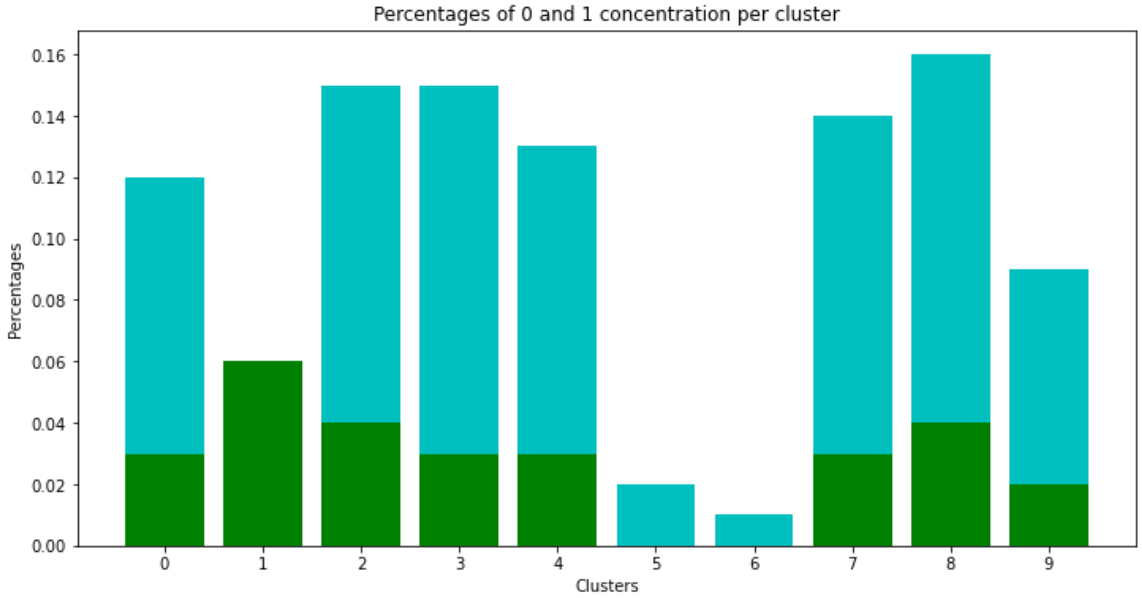
Εικόνα 8: Kmeans one-hot-encode and PCA



Πίνακας 12: Kmeans FAMD

Kmeans FAMD
Το cluster 0 αντιστοιχεί σε 0 με: 2504 -> 0 και 574-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 1 αντιστοιχεί σε 1 με: 1302 -> 1 και 629 -> 0.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.03 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.21
Το cluster 2 αντιστοιχεί σε 0 με: 3229 -> 0 και 940 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.15
Το cluster 3 αντιστοιχεί σε 0 με: 3088 -> 0 και 603 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1
Το cluster 4 αντιστοιχεί σε 0 με: 2841-> 0 και 659 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.13 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1
Το cluster 5 αντιστοιχεί σε 0 με: 368 -> 0 και 28 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 6 αντιστοιχεί σε 0 με: 272 -> 0 και 87 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 7 αντιστοιχεί σε 0 με: 3070 -> 0 και 740 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.14 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.12
Το cluster 8 αντιστοιχεί σε 0 με: 3390 -> 0 και 954-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.16 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.15
Το cluster 9 αντιστοιχεί σε 0 με: 1836-> 0 και 407 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06

Εικόνα 9: Kmeans FAMD

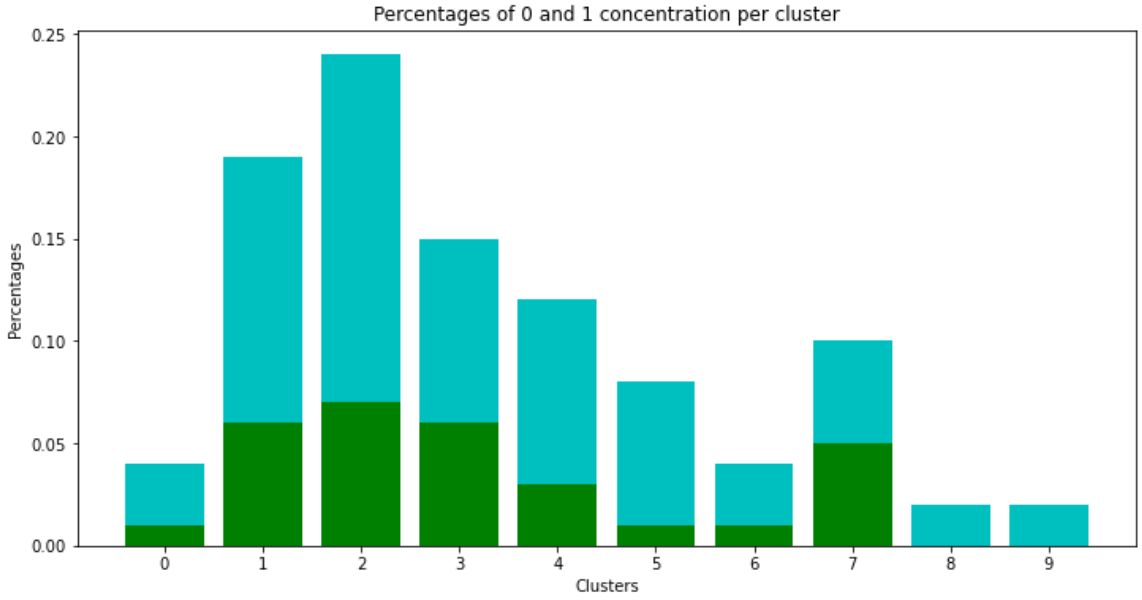


Πίνακας 13: Kmeans MCA

Kmeans MCA
Το cluster 0 αντιστοιχεί σε 0 με: 860 -> 0 και 210 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.03
Το cluster 1 αντιστοιχεί σε 0 με: 3968 -> 0 και 1347 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.19 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.21
Το cluster 2 αντιστοιχεί σε 0 με: 5087 -> 0 και 1420-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.24 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.23
Το cluster 3 αντιστοιχεί σε 0 με: 3217 -> 0 και 1194 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.19
Το cluster 4 αντιστοιχεί σε 0 με: 2530 -> 0 και 620 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1
Το cluster 5 αντιστοιχεί σε 0 με: 1675 -> 0 και 142 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02
Το cluster 6 αντιστοιχεί σε 0 με: 819 -> 0 και 241 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.04
Το cluster 7 αντιστοιχεί σε 0 με: 2086 -> 0 και 958 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που

συγκεντρώνονται σε αυτό το cluster: 0.15
Το cluster 8 αντιστοιχεί σε 0 με: 490 -> 0 και 81 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 9 αντιστοιχεί σε 0 με: 495 -> 0 και 81 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01

Εικόνα 10: Kmeans MCA



Όσων αφορά τον αλγόριθμο kmeans, φαίνεται ότι οι περισσότερες δοκιμές δεν καταφέρνουν να δημιουργήσουν cluster με πλειοψηφία εγγραφές που θα προβούν σε αθέτηση πληρωμών. Στα περισσότερα clusters είναι φανερά η πλειοψηφία οι εγγραφές που δε θα προβούν σε αθέτηση, και φαίνεται ότι οι εγγραφές που θα προβούν σε αθέτηση είναι μοιρασμένες μεταξύ των clusters με αποτέλεσμα να είναι σχεδόν σε όλες τις περιπτώσεις η μειοψηφία. Η μοναδική δοκιμή που δημιούργησε cluster με πλειοψηφία εγγραφές που έχουν Y ίσο με 1 είναι η τεχνική FAMD που στο δεύτερο cluster συγκέντρωσε 1302 τέτοιες εγγραφές. Πρακτικά όμως και πάλι το ποσοστό αυτό είναι χαμηλό, δεδομένου ότι αποτελεί μόνο το 21% του συνολικού αριθμού αθετήσεων που θα έπρεπε να προβλεφθούν.

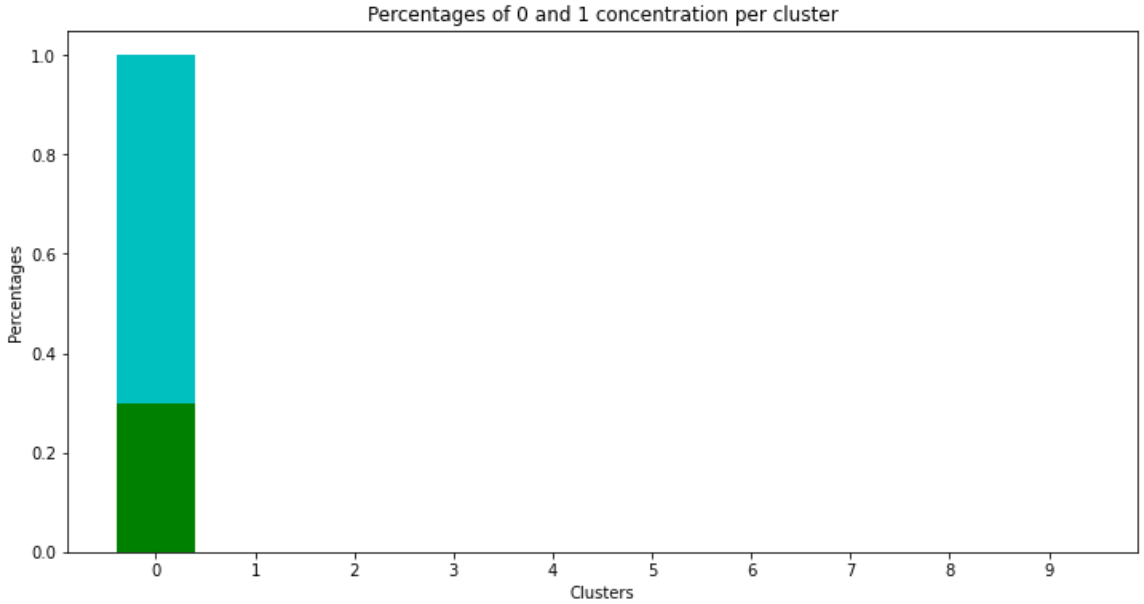
Hierarchical Clustering

-SINGLE LINKAGE

Πίνακας 14: Hierarchical Single linkage simple

Hierarchical Single linkage simple
Το cluster 0 αντιστοιχεί σε 0 με: 21218 -> 0 και 6294 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 1.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 1.0
Το cluster 1 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 2 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 3 αντιστοιχεί σε 0 με: 1-> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 4 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 5 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 6 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 7 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 9 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 11: Hierarchical Single linkage simple

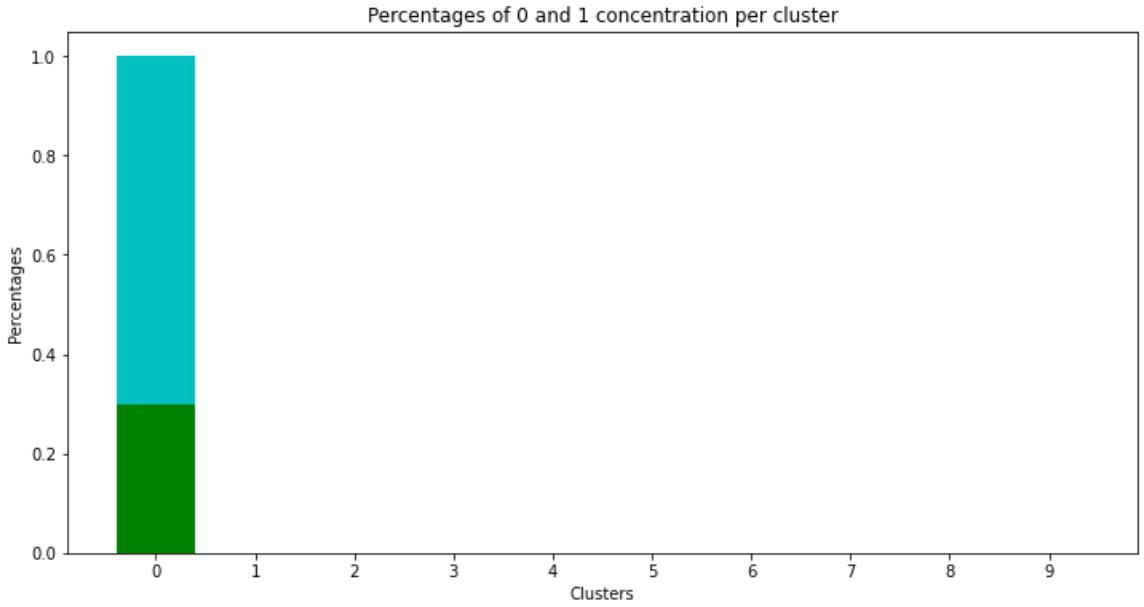


Πίνακας 15: Hierarchical Single linkage normalization

Hierarchical Single linkage normalization
Το cluster 0 αντιστοιχεί σε 0 με: 21217 -> 0 και 6294 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 1.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 1.0
Το cluster 1 αντιστοιχεί σε 0 με: 2 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 2 αντιστοιχεί σε 0 με: 1-> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 3 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 4 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 5 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 6 αντιστοιχεί σε 0 με: 1-> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 7 αντιστοιχεί σε 0 με: 1-> 0 και 0 -> 1.

Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 9 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 12: Hierarchical Single linkage normalization

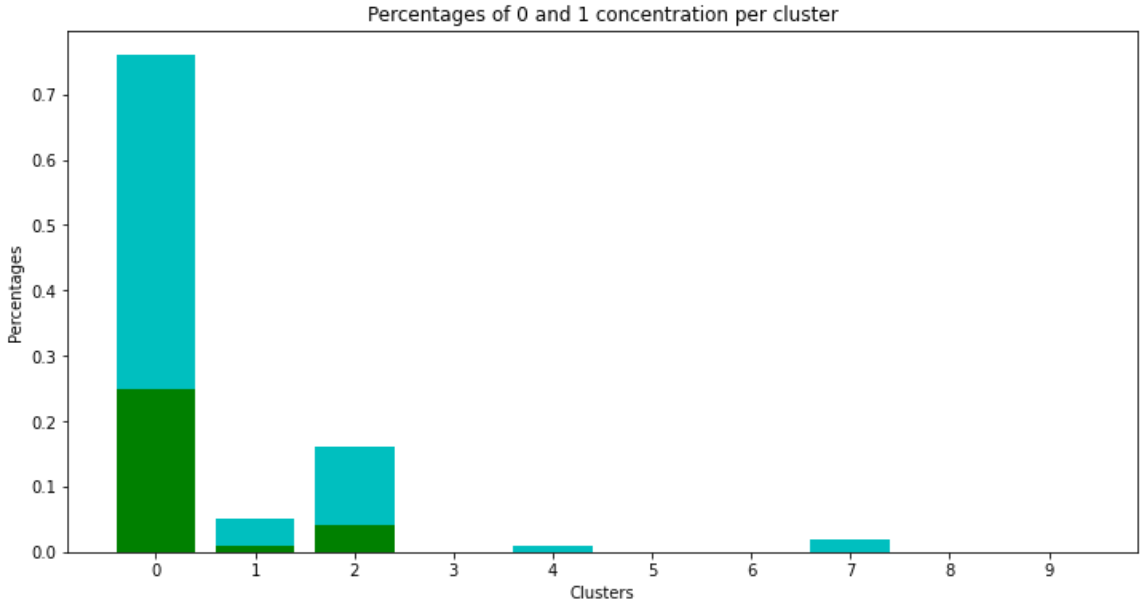


Πίνακας 16: Hierarchical Single linkage PCA

Hierarchical Single linkage PCA
Το cluster 0 αντιστοιχεί σε 0 με: 2 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 1 αντιστοιχεί σε 0 με: 21215 -> 0 και 6294 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 1.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 1.0
Το cluster 2 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 3 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Το cluster 4 αντιστοιχεί σε 0 με: 2 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 5 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 6 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 7 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 9 αντιστοιχεί σε 0 με: 2 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 13: Hierarchical Single linkage PCA

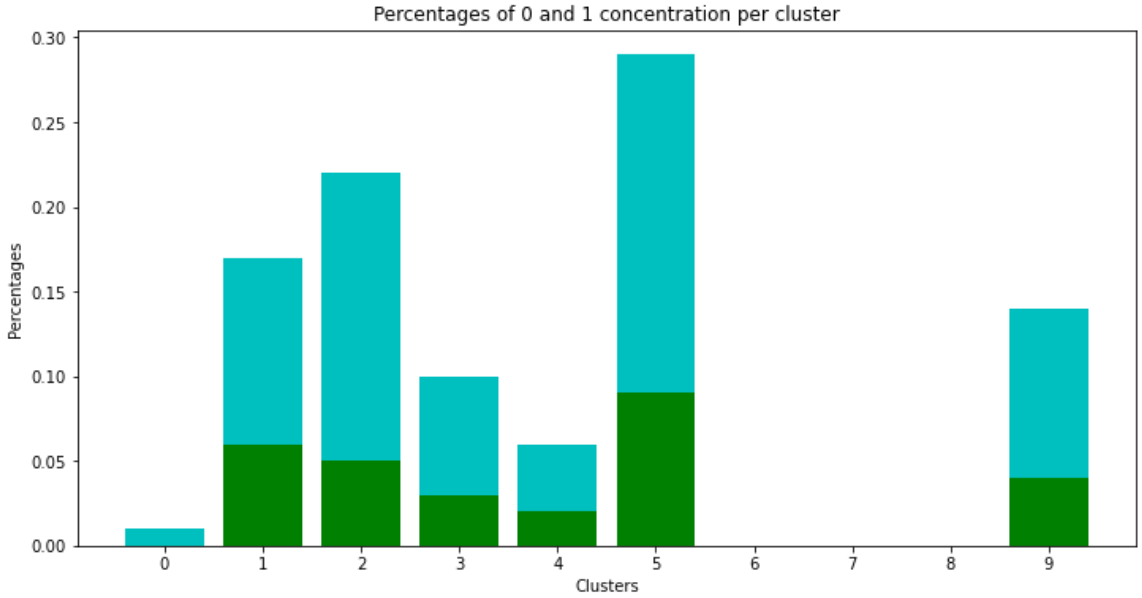


Πίνακας 17: Hierarchical Single linkage one-hot-encode and PCA

Hierarchical Single linkage one-hot-encode and PCA
Το cluster 0 αντιστοιχεί σε 0 με: 241 -> 0 και 16 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Το cluster 1 αντιστοιχεί σε 0 με: 3635 -> 0 και 1346 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.17 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.21
Το cluster 2 αντιστοιχεί σε 0 με: 4602-> 0 και 1068-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.22 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.17
Το cluster 3 αντιστοιχεί σε 0 με: 2105 -> 0 και 670 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 4 αντιστοιχεί σε 0 με: 1333 -> 0 και 521 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.08
Το cluster 5 αντιστοιχεί σε 0 με: 6189 -> 0 και 1819 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.29 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.29
Το cluster 6 αντιστοιχεί σε 0 με: 72 -> 0 και 4-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 7 αντιστοιχεί σε 0 με: 2 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 60 -> 0 και 8 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 9 αντιστοιχεί σε 0 με: 2988 -> 0 και 842 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.14 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.13

Εικόνα 14: Hierarchical Single linkage one-hot-encode and PCA

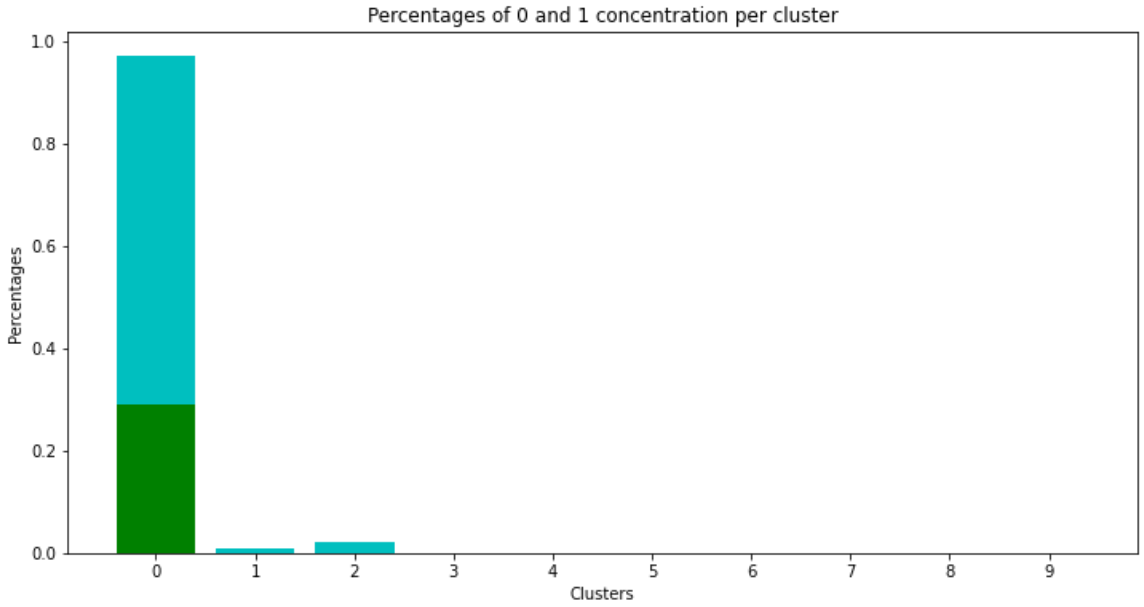


Πίνακας 18: Hierarchical Single linkage FAMD

Hierarchical Single linkage FAMD
Το cluster 0 αντιστοιχεί σε 0 με: 20582-> 0 και 6179 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.97 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.98
Το cluster 1 αντιστοιχεί σε 0 με: 264 -> 0 και 87 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 2 αντιστοιχεί σε 0 με: 368 -> 0 και 28 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 3 αντιστοιχεί σε 0 με: 7 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 4 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 5 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 6 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 7 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που

συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 1 -> 0 και 0-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 9 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 15: Hierarchical Single linkage FAMD

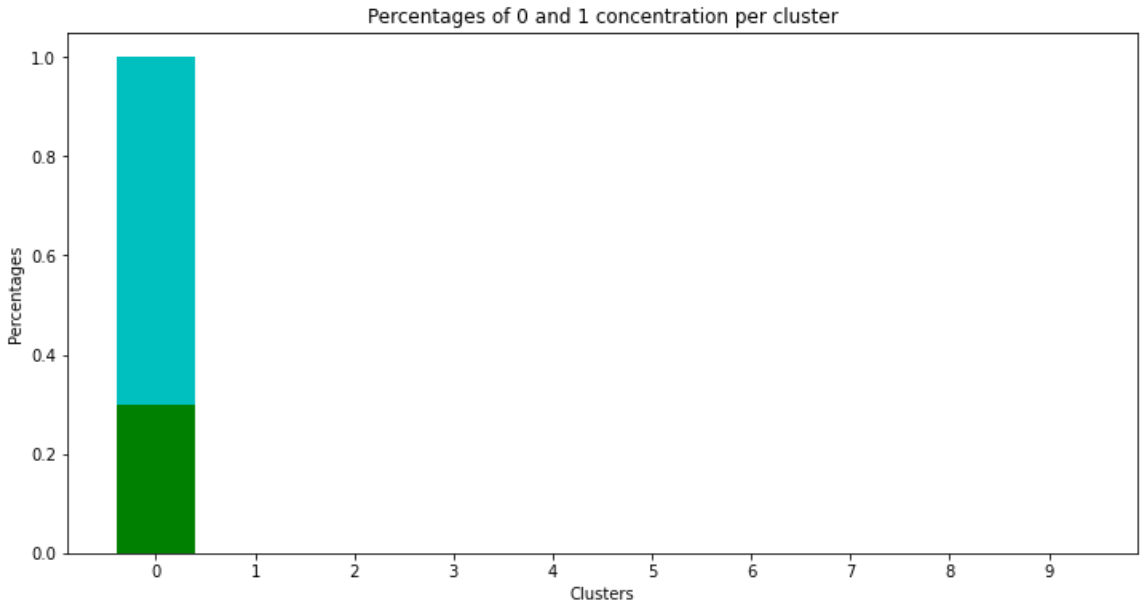


Πίνακας 19: Hierarchical Single linkage MCA

Hierarchical Single linkage MCA
Το cluster 0 αντιστοιχεί σε 0 με: 21221 -> 0 και 6291 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 1.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 1.0
Το cluster 1 αντιστοιχεί σε 1 με: 1 -> 1 και 0 -> 0.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 2 αντιστοιχεί σε 0 με: 1 -> 0 και 0-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 3 αντιστοιχεί σε 1 με: 1 -> 1 και 0 -> 0.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 4 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που

συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 5 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 6 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 7 αντιστοιχεί σε 0 με: 1 -> 0 και 0 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 1 με: 1 -> 1 και 0 -> 0.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 9 αντιστοιχεί σε 0 με: 1-> 0 και 0-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.0 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 16: Hierarchical Single linkage MCA



Παρατηρείται ότι ο αλγόριθμος Hierarchical με single linkage δημιουργεί πολύ άνισα clusters σε όλες σχεδόν τις δοκιμές. Όλες οι εγγραφές μαζεύονται σε ένα cluster, με αποτέλεσμα να μην υπάρχει διαχωρισμός των αθετήσεων και στα υπόλοιπα clusters παρατηρείται μονοψήφιος αριθμός εγγραφών. Εξαίρεση αποτελεί η δοκιμή με την τεχνική one-hot-encode, η οποία καταφέρνει να δημιουργήσει clusters στα οποία να

μοιράζονται οι εγγραφές αλλά και πάλι σε όλα τα clusters πλειοψηφία είναι οι εγγραφές που δε θα προβούν σε αθέτηση άρα δεν επιτυγχάνεται ο σκοπός της συσταδοποίησης.

Hierarchical Clustering

-COMPLETE LINKAGE

Όσων αφορά τα αποτελέσματα του Hierarchical με complete linkage, παρατηρούνται παρόμοια αποτελέσματα με τον Hierarchical με single linkage με τον αλγόριθμο να δημιουργεί πολύ άνισα clusters στις περισσότερες δοκιμές. Για λόγους συντομίας παραλείπονται οι αναλυτικοί πίνακες και εικόνες. Οι περισσότερες εγγραφές μαζεύονται σε λίγα clusters (1, 2 ή 3 σε αριθμό) με αποτέλεσμα να μην υπάρχει διαχωρισμός των αθετήσεων και στα υπόλοιπα clusters παρατηρείται πολύ μικρός αριθμός εγγραφών. Εξαίρεση και πάλι αποτελεί η δοκιμή με την τεχνική one-hot-encode, η οποία καταφέρνει να δημιουργήσει clusters στα οποία να μοιράζονται καλύτερα οι εγγραφές Επίσης σημειώνεται ότι αυτή η τεχνική δημιουργεί τέσσερα clusters στα οποία πλειοψηφία είναι οι εγγραφές που θα προβούν σε αθέτηση. Παρόλα αυτά συνολικά από αυτά τα 4 clusters προβλέπονται μόνο 941 αθετήσεις από τις 6.294, δηλαδή σε ποσοστό περίπου το 15% του συνολικού αριθμού αθετήσεων.

Hierarchical Clustering

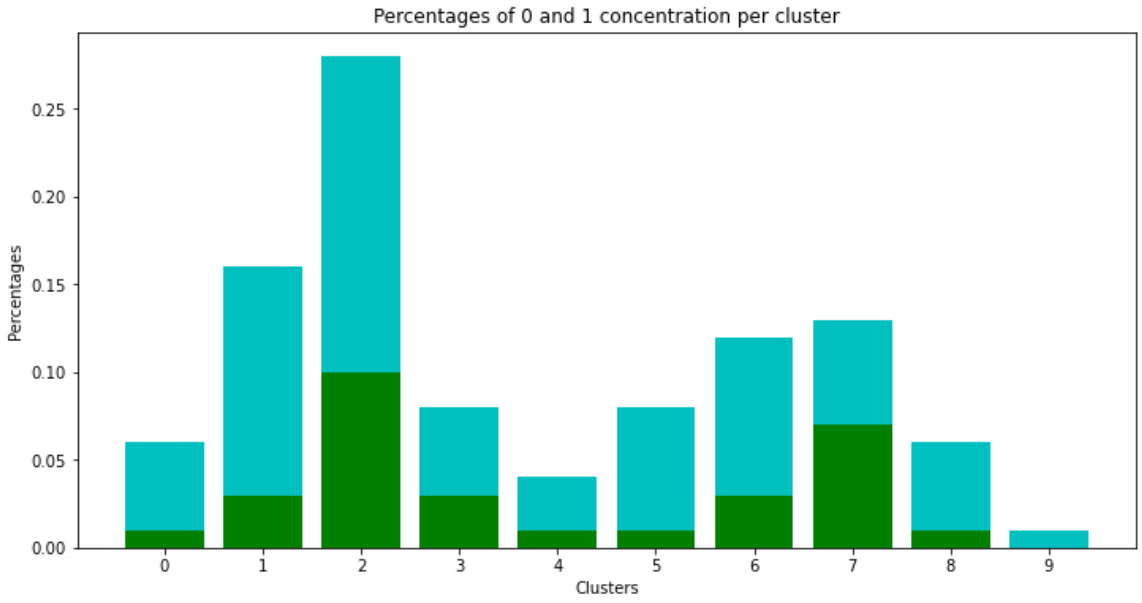
-WARD LINKAGE

Πίνακας 20: Hierarchical Ward linkage simple

Hierarchical Ward linkage simple
Το cluster 0 αντιστοιχεί σε 0 με: 1280 -> 0 και 169 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.03
Το cluster 1 αντιστοιχεί σε 0 με: 3335 -> 0 και 646 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.16 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1
Το cluster 2 αντιστοιχεί σε 0 με: 5894 -> 0 και 2090 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.28 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.33
Το cluster 3 αντιστοιχεί σε 0 με: 1729 -> 0 και 554 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 4 αντιστοιχεί σε 0 με: 775 -> 0 και 155 -> 1.

Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02
Το cluster 5 αντιστοιχεί σε 0 με: 1720 -> 0 και 303 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.05
Το cluster 6 αντιστοιχεί σε 0 με: 2443 -> 0 και 581 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 7 αντιστοιχεί σε 0 με: 2680 -> 0 και 1491 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.13 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.24
Το cluster 8 αντιστοιχεί σε 0 με: 1193 -> 0 και 285-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.05
Το cluster 9 αντιστοιχεί σε 0 με: 178 -> 0 και 20 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 17: Hierarchical Ward linkage simple

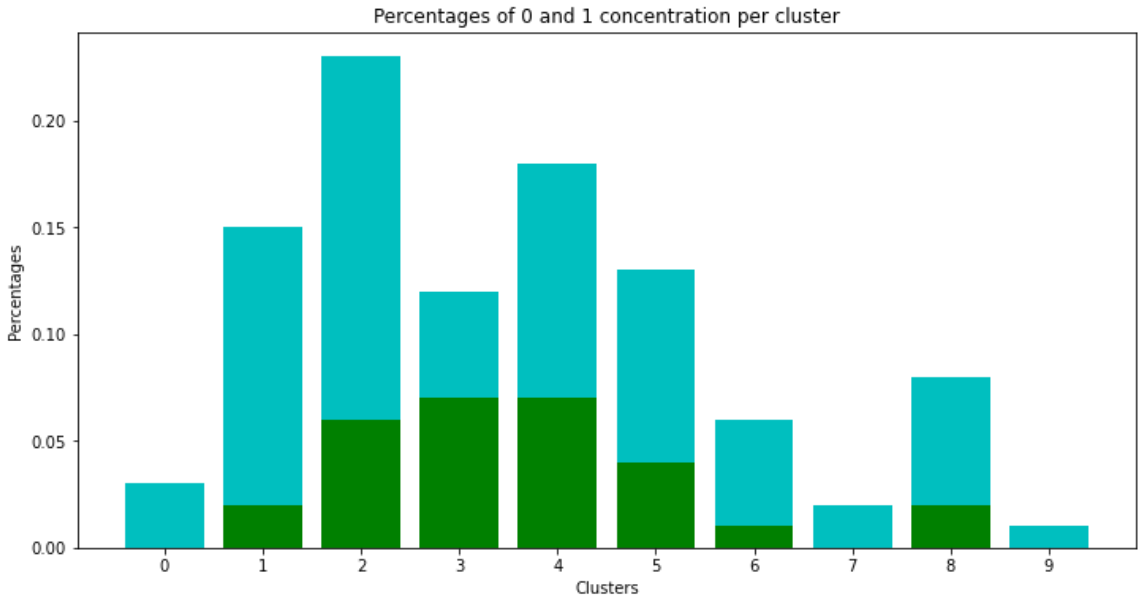


Πίνακας 21: Hierarchical Ward linkage normalization

Hierarchical Ward linkage normalization
Το cluster 0 αντιστοιχεί σε 0 με: 690 -> 0 και 94 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.03 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 1 αντιστοιχεί σε 0 με: 3089 -> 0 και 458 -> 1.

Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.07
Το cluster 2 αντιστοιχεί σε 0 με: 4825 -> 0 και 1221 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.23 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.19
Το cluster 3 αντιστοιχεί σε 0 με: 2442 -> 0 και 1407 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.22
Το cluster 4 αντιστοιχεί σε 0 με: 3866 -> 0 και 1429 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.18 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.23
Το cluster 5 αντιστοιχεί σε 0 με: 2764 -> 0 και 920 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.13 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.15
Το cluster 6 αντιστοιχεί σε 0 με: 1245 -> 0 και 290 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.05
Το cluster 7 αντιστοιχεί σε 0 με: 337 -> 0 και 46 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01
Το cluster 8 αντιστοιχεί σε 0 με: 1688 -> 0 και 385-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06
Το cluster 9 αντιστοιχεί σε 0 με: 281 -> 0 και 44 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01

Εικόνα 18: Hierarchical Ward linkage normalization

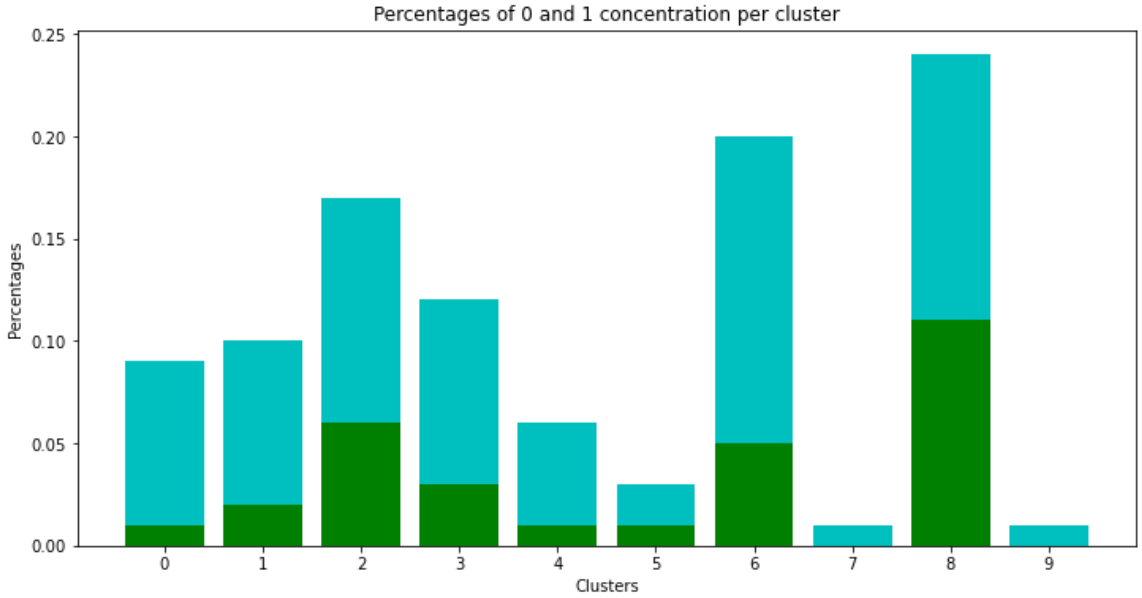


Πίνακας 22: Hierarchical Ward linkage PCA

Hierarchical Ward linkage PCA
Το cluster 0 αντιστοιχεί σε 0 με: 1854 -> 0 και 181 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.03
Το cluster 1 αντιστοιχεί σε 0 με: 2100 -> 0 και 433. -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.07
Το cluster 2 αντιστοιχεί σε 0 με: 3541 -> 0 και 1319 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.17 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.21
Το cluster 3 αντιστοιχεί σε 0 με: 2445 -> 0 και 583-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.09
Το cluster 4 αντιστοιχεί σε 0 με: 1173 -> 0 και 271 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.04
Το cluster 5 αντιστοιχεί σε 0 με: 622 -> 0 και 116 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.03 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02
Το cluster 6 αντιστοιχεί σε 0 με: 4146 -> 0 και 1019 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.2 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.16
Το cluster 7 αντιστοιχεί σε 0 με: 187 -> 0 και 30 -> 1.

Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 5006 -> 0 και 2336 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.24 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.37
Το cluster 9 αντιστοιχεί σε 0 με: 153 -> 0 και 6 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0

Εικόνα 19: Hierarchical Ward linkage PCA

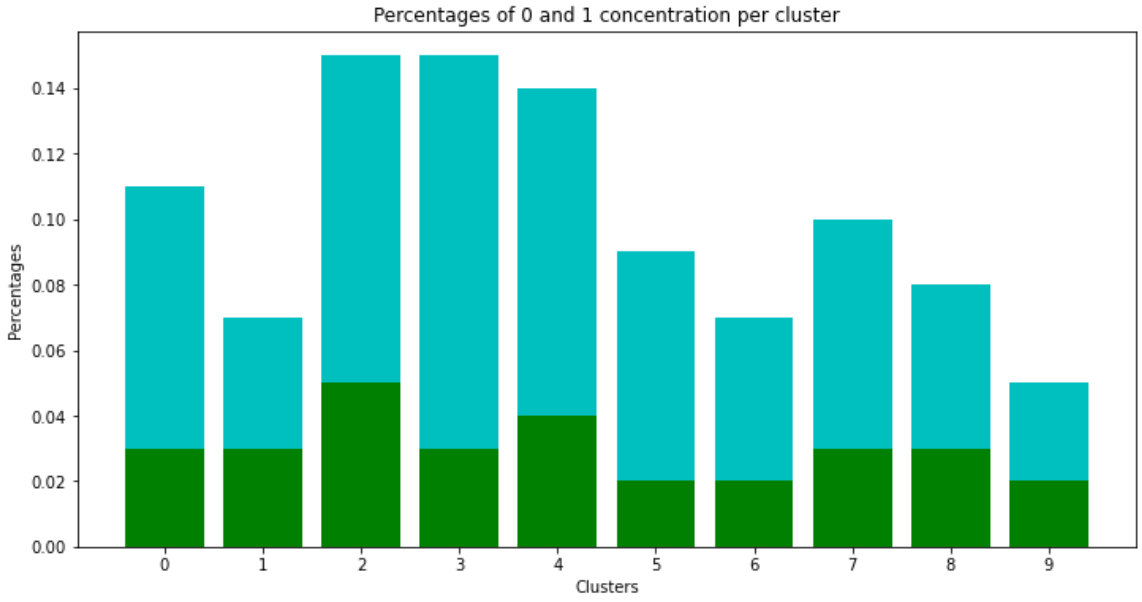


Πίνακας 23: Hierarchical Ward linkage one-hot-encode and PCA

Hierarchical Ward linkage one-hot-encode and PCA
Το cluster 0 αντιστοιχεί σε 0 με: 2346 -> 0 και 686 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.11 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 1 αντιστοιχεί σε 0 με: 1467 -> 0 και 533 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.08
Το cluster 2 αντιστοιχεί σε 0 με: 3257-> 0 και 996 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.16
Το cluster 3 αντιστοιχεί σε 0 με: 3132 -> 0 και 710 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 4 αντιστοιχεί σε 0 με: 2932 -> 0 και 823 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.14 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.13

Το cluster 5 αντιστοιχεί σε 0 με: 1870 -> 0 και 480 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.08
Το cluster 6 αντιστοιχεί σε 0 με: 1470 -> 0 και 358 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.07 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06
Το cluster 7 αντιστοιχεί σε 0 με: 2026 -> 0 και 713 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 8 αντιστοιχεί σε 0 με: 1609 -> 0 και 633 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.1
Το cluster 9 αντιστοιχεί σε 0 με: 1118 -> 0 και 362 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.05 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06

Εικόνα 20: Hierarchical Ward linkage one-hot-encode and PCA

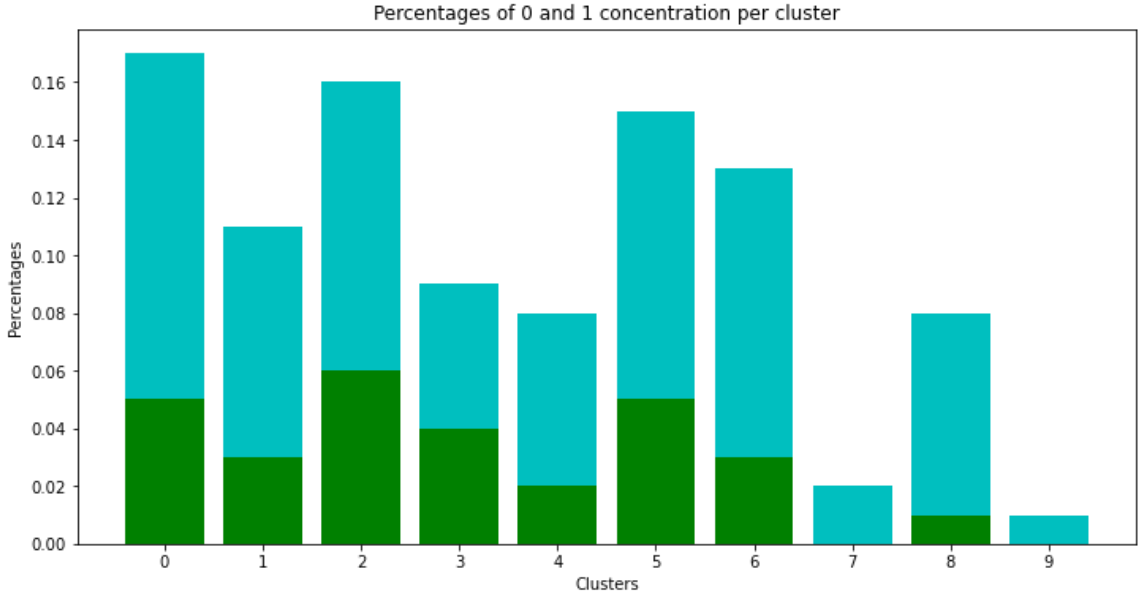


Πίνακας 24: Hierarchical Ward linkage FAMD

Hierarchical Ward linkage FAMD
Το cluster 0 αντιστοιχεί σε 0 με: 3616-> 0 και 1011 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.17 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.16
Το cluster 1 αντιστοιχεί σε 0 με: 2329 -> 0 και 672 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.11 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11

Το cluster 2 αντιστοιχεί σε 0 με: 3291 -> 0 και 1251 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.16 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.2
Το cluster 3 αντιστοιχεί σε 0 με: 1957 -> 0 και 766. -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.12
Το cluster 4 αντιστοιχεί σε 0 με: 1641 -> 0 και 453 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.07
Το cluster 5 αντιστοιχεί σε 0 με: 3138 -> 0 και 1096 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.15 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.17
Το cluster 6 αντιστοιχεί σε 0 με: 2844 -> 0 και 673 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.13 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 7 αντιστοιχεί σε 0 με: 368-> 0 και 28 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.02 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.0
Το cluster 8 αντιστοιχεί σε 0 με: 1771 -> 0 και 257 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.08 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.04
Το cluster 9 αντιστοιχεί σε 0 με: 272 -> 0 και 87 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.01 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.01

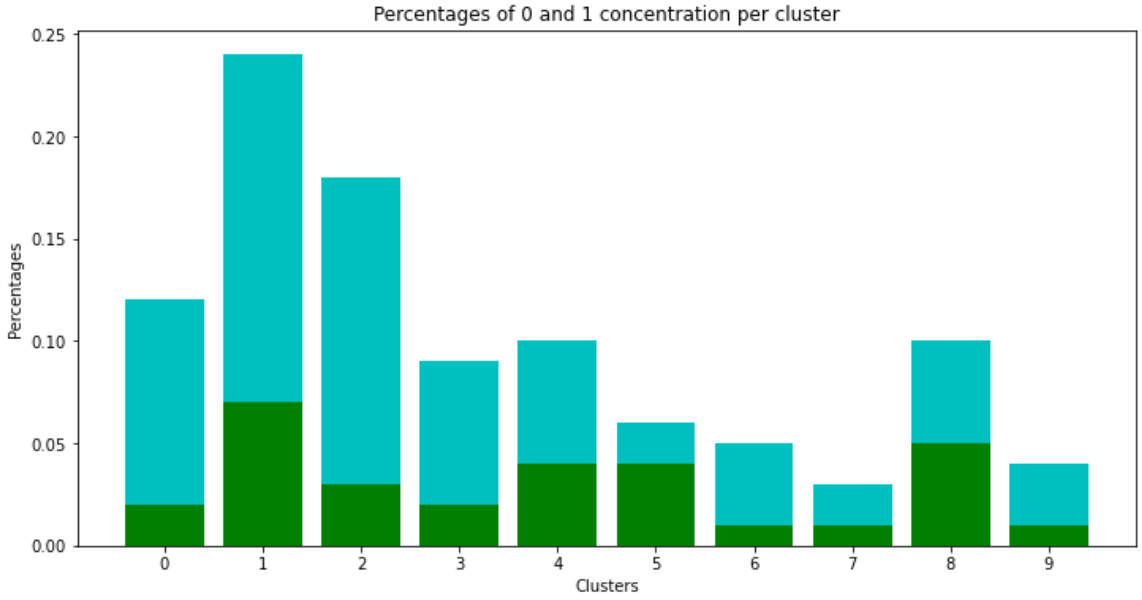
Εικόνα 21: Hierarchical Ward linkage FAMD



Πίνακας 25: Hierarchical Ward linkage MCA

Hierarchical Ward linkage MCA
Το cluster 0 αντιστοιχεί σε 0 με: 2532 -> 0 και 348 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.12 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.06
Το cluster 1 αντιστοιχεί σε 0 με: 5017 -> 0 και 1429 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.24 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.23
Το cluster 2 αντιστοιχεί σε 0 με: 3861 -> 0 και 718 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.18 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.11
Το cluster 3 αντιστοιχεί σε 0 με: 1825 -> 0 και 463-> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.09 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.07
Το cluster 4 αντιστοιχεί σε 0 με: 2126 -> 0 και 817 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.13
Το cluster 5 αντιστοιχεί σε 0 με: 1301 -> 0 και 928 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.06 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.15
Το cluster 6 αντιστοιχεί σε 0 με: 1034 -> 0 και 286 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.05 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.05
Το cluster 7 αντιστοιχεί σε 0 με: 679 -> 0 και 120 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.03 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.02
Το cluster 8 αντιστοιχεί σε 0 με: 2019 -> 0 και 959 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.1 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.15
Το cluster 9 αντιστοιχεί σε 0 με: 833 -> 0 και 226 -> 1.
Ποσοστό των 0 που συγκεντρώνονται σε αυτό το cluster: 0.04 και ποσοστό των 1 που συγκεντρώνονται σε αυτό το cluster: 0.04

Εικόνα 22: Hierarchical Ward linkage MCA



Όσων αφορά τα αποτελέσματα της εφαρμογής του hierarchical αλγορίθμου με ward linkage, φαίνεται ότι σε σύγκριση με το single και το complete linkage μοιράζει καλύτερα τις εγγραφές στα clusters. Παρόλα αυτά καμία τεχνική δε κατάφερε να δημιουργήσει clusters στα οποία πλειοψηφία να είναι οι εγγραφές που θα προβούν σε αθέτηση.

Όλες οι παραπάνω δοκιμές εφαρμόστηκαν και στο dataset χωρίς να αφαιρεθούν οι ακραίες τιμές και τα αποτελέσματα ήταν αντίστοιχα με αυτά χωρίς τις ακραίες τιμές που παρουσιάστηκαν παραπάνω. Αυτό υποδηλώνει ότι δεν χάνεται χρήσιμη πληροφορία αφαιρώντας αυτές τιμές και καλώς αφαιρούνται. Παρατηρήθηκε ότι χωρίς τις ακραίες τιμές οι αλγόριθμοι ήταν ελαφρώς πιο γρήγοροι σε χρόνο υλοποίησης.

Γενικά από τα αποτελέσματα των τεχνικών μη επιβλεπόμενης μάθησης, προκύπτει ότι ο αλγόριθμος kmeans καταφέρνει καλύτερα να δημιουργήσει ισόποσα clusters από ότι ο hierarchical. Επίσης, η καλύτερη απόδοση του kmeans είναι με την τεχνική FAMD με την οποία κατάφερε να προβλέψει το 20% των αθετήσεων έναντι της καλύτερης απόδοσης του hierarchical με complete linkage και με την τεχνική One-hot-encode που κατάφερε να προβλέψει μόλις το 15% των αθετήσεων. Σημειώνεται επίσης ότι ο αλγόριθμος kmeans ήταν σημαντικά γρηγορότερος στην υλοποίηση του από ότι ο hierarchical με όλα τα διαφορετικά linkage.

Παρακάτω παρουσιάζονται τα αποτελέσματα της εφαρμογής των τεχνικών επιβλεπόμενης μάθησης, Random Forest και Logistic Regression.

Πίνακας 26: Επιβλεπόμενη μάθηση

Supervised learning technique	Accuracy
Random Forest – χωρίς κανονικοποίηση	0.778
Random Forest – με κανονικοποίηση	0.779
Logistic Regression	0.788

Παρατηρείται αρκετά καλή απόδοση των τεχνικών επιβλεπόμενης μάθησης με την Logistic Regression να επιτυγχάνει την μέγιστη απόδοση, ~79%.

Συμπερασματικά, είναι φανερό ότι η απόδοση των τεχνικών μη επιβλεπόμενης μάθησης δεν ήταν καλή στο συγκεκριμένο dataset ενώ αντιθέτως οι τεχνικές επιβλεπόμενης μάθησης είχαν πολύ καλή απόδοση. Το αποτέλεσμα αυτό αποδίδεται στο γεγονός ότι δεν ήταν καθόλου ισόποσα κατανεμημένες οι δύο διαφορετικές τιμές που έγινε προσπάθεια να προβλεφθούν και κατά συνέπεια οι αλγόριθμοι δεν δημιούργησαν clusters για τις εγγραφές που θα προβούν σε αθέτηση πληρωμών.

Σίγουρα η καλύτερη απόδοση της logistic regression δικαιολογεί την ευρεία χρήση της από τις τράπεζες και τα χρηματοπιστωτικά ιδρύματα δεδομένου ότι στόχος τους είναι η μέγιστη καλύτερη πρόβλεψη αθετήσεων πληρωμών από τους πελάτες τους.

8 . Επεκτάσεις και μελλοντική διερεύνηση

Ως μελλοντικές προτάσεις για επέκταση αυτής της έρευνας προτείνεται αρχικά η εφαρμογή των αλγορίθμων σε ένα διαφορετικό σύνολο δεδομένων το οποίο να έχει μεγαλύτερο ποσοστό αθετήσεων και να είναι καλύτερα διαμοιρασμένες οι εγγραφές σε 0 και 1 ως προς την μεταβλητή πρόβλεψης.

Σε επόμενο βήμα θα μπορούσε να γίνει η σύγκριση των αλγορίθμων με βάση τον χρόνο υλοποίησης, ώστε να λαμβάνεται και αυτό το μέτρο ως προς τον υπολογισμό της απόδοσης των αλγορίθμων. Επίσης θα μπορούσαν να χρησιμοποιηθούν διαφορετικές

μετρικές για την απόδοση των αλγορίθμων όπως το silhouette score και το rand index καθώς και να μελετηθεί περισσότερο ο τρόπος με τον οποίο μπορούν να βρεθούν ακραίες τιμές του συνόλου δεδομένων.

Από την βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε προκύπτει ότι θα μπορούσε να γίνει περισσότερη έρευνα σχετικά με τη χρήση αλγορίθμων βαθιάς μάθησης στη μοντελοποίηση και διαχείριση του πιστωτικού κινδύνου.

Επιπλέον, θα ήταν χρήσιμο να γίνει μελέτη της απόδοσης των αλγορίθμων μη επιβλεπόμενης μάθησης στα μοντέλα LGD (Loss Given Default) και EAD (Exposure At Default) τα οποία έχουν επίσης αρκετά σημαντικό ρόλο στη μοντελοποίηση και διαχείριση του πιστωτικού κινδύνου.

Σίγουρα, η σημαντικότερη ίσως μελλοντική πρόταση είναι η δοκιμή μεθόδων μη εποπτευόμενης μάθησης σε κάποιο μεγαλύτερο σύνολο δεδομένων και ακόμη καλύτερα σε ένα σύνολο πραγματικών δεδομένων. Με τον τρόπο αυτό θα τεκμηριωθεί καλύτερα η απόδοση των τεχνικών μη εποπτευόμενης μάθησης στην μοντελοποίηση και κατ' επέκταση, διαχείριση του πιστωτικού κινδύνου ώστε ίσως να συμπεριληφθούν περισσότερο στην πολιτική των τραπεζών και χρηματοπιστωτικών ιδρυμάτων.

Βιβλιογραφία

Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *41*(8), 3825-3830.

Achtert, E., Kriegel, H. P., Reichert, L., Schubert, E., Wojdanowski, R., & Zimek, A. (2010, April). Visual evaluation of outlier detection models. In *International Conference on Database Systems for Advanced Applications* (pp. 396-399). Springer, Berlin, Heidelberg.

Agrafiotis. Stochastic proximity embedding. Journal of Computational Chemistry, 24(10):1215–1221, 2003.

Ala'raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89-105.

Aparicio, C., Gutiérrez, J., Jaramillo, M., & Moreno, H. (2013). Indicadores alternativos de riesgo de crédito en el Perú: matrices de transición crediticia condicionadas al ciclo económico. *Documentos de Trabajo*, 01-2013.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, *89*(428), 1329-1339.

Atkinson, A. C., & Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics*, *15*(2), 460-476.

Ayadi, A., Ghorbel, O., Obeid, A. M., & Abid, M. (2017). Outlier detection approaches for wireless sensor networks: A survey. *Computer Networks*, *129*, 319-333.

Baesens, B., Roesch, D., & Scheule, H. (2016). *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons.

Barnett, V., & Lewis, T. (1994). Evolution by gene duplication. *Outliers in Statistical Data, 3rd ed.; Wiley: Hoboken, NJ, USA*.

Barrett, B. E., & Gray, J. B. (1997). Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression. *Computational statistics & data analysis*, *26*(1), 39-52.

Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. *Expert Systems with Applications*, *134*, 209-224.

Bolstad, W. M., & Manda, S. O. (2001). Investigating child mortality in Malawi using family and community random effects: a Bayesian analysis. *Journal of the American Statistical Association*, *96*(453), 12-19.

C.K.I. Williams. On a connection between Kernel PCA and metric multidimensional scaling. Machine Learning, 46(1-3):11–19, 2002.

Cai, Z., He, Z., Guan, X., & Li, Y. (2016). Collective data-sanitization for preventing sensitive information inference attacks in social networks. *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 577-590.

Castelli, M., Manzoni, L., & Popovič, A. (2016). An artificial intelligence system to predict quality of service in banking organizations. *Computational intelligence and neuroscience*, *2016*.

Chen, G. G., & Astebro, T. (2001). The economic value of reject inference in credit scoring. *Department of Management Science, University of Waterloo*.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, *16*(1), 41-46.

Cook, J., Sutskever, I., Mnih, A., & Hinton, G. (2007, March). Visualizing similarity data with a mixture of maps. In *Artificial intelligence and statistics* (pp. 67-74). PMLR.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

Cox, M. A., & Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization* (pp. 315-347). Springer, Berlin, Heidelberg.

D. Xu, X. Zhang, J. Hu, J. Chen, "A Novel Ensemble Credit Scoring Model Based on Extreme Learning Machine and Generalized Fuzzy Soft Sets", Mathematical Problems in Engineering vol. 2020, pp. 1-12 , (2020).

Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, *115*(9).

Demartines, P., & Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on neural networks*, *8*(1), 148-154.

Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A survey on urban traffic anomalies detection algorithms. *IEEE Access*, *7*, 12192-12205.

Donepudi, P. K. (2016). Influence of cloud computing in business: are they robust. *Asian journal of applied science and engineering*, *5*(3), 193-196.

Donepudi, P. K. (2017). Machine learning and artificial intelligence in banking. *Engineering International*, *5*(2), 83-86.

Dubuisson, B., & Masson, M. (1993). A statistical decision rule with incomplete knowledge about classes. *Pattern recognition*, *26*(1), 155-165.

Ekins, S., Balakin, K. V., Savchuk, N., & Ivanenkov, Y. (2006). Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *Journal of medicinal chemistry*, *49*(17), 5059-5071.

Emmott, A., Das, S., Dietterich, T., Fern, A., & Wong, W. K. (2016). Anomaly detection meta-analysis benchmarks.

Espinoza, L. A. P. (2013). Matrices de Transición del Crédito en Nicaragua. *Managua, Nicaragua*.

Gavira-Durón, N., Gutierrez-Vargas, O., & Cruz-Aké, S. (2021). Markov Chain K-Means Cluster Models and Their Use for Companies' Credit Quality and Default Probability Estimation. *Mathematics*, *9*(8), 879.

Gebremeskel, G. B., Yi, C., He, Z., & Haile, D. (2016). Combined data mining techniques based patient data outlier detection for healthcare safety. *International Journal of Intelligent Computing and Cybernetics*.

Gogoi, P., Bhattacharyya, D. K., Borah, B., & Kalita, J. K. (2011). A survey of outlier detection methods in network anomaly identification. *The Computer Journal*, *54*(4), 570-588.

Guo, G., & Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in Guatemala. *Journal of the American Statistical Association*, *87*(420), 969-976.

Hadi. A. S., Imon R., and Werner. M. (2009). Detection of Outliers, vol. 1, no. 1. Hoboken, NJ, USA: Wiley, pp. 57–70.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, 357-384.

Hand, D. J., & Henley, W. E. (1993). Can reject inference ever work?. *IMA Journal of Management Mathematics*, *5*(1), 45-55.

Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, 271-320.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, *22*(2), 85-126.

Jian, M., & Xu, M. (2012). Determinants of the guarantee circles: The case of Chinese listed firms. *Pacific-Basin Finance Journal*, *20*(1), 78-100.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009, April). Outlier detection in axis-parallel subspaces of high dimensional data. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 831-838). Springer, Berlin, Heidelberg.

Kumar, S. A., & Chandrakala, D. (2016). A survey on customer churn prediction using machine learning techniques. *International Journal of Computer Applications*, *975*, 8887.

Lagunas Puls, S., & Ramírez Pacheco, J. C. (2017). Expectativas para operaciones financieras en los sectores vulnerables mediante Matrices de Transición. *Revista mexicana de economía y finanzas*, *12*(2), 71-101.

Liberati, C., & Camillo, F. (2018). Personal values and credit scoring: new insights in the financial prediction. *Journal of the Operational Research Society*, *69*(12), 1994-2005.

Liu, H., Li, X., Li, J., & Zhang, S. (2017). Efficient outlier detection for high-dimensional data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *48*(12), 2451-2461.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117-134.

McCulloch, R. E., & Tsay, R. S. (1994). Statistical analysis of economic time series via Markov switching models. *Journal of time series analysis*, *15*(5), 523-539.

Minsky, H. P. (1992). *The capital development of the economy and the structure of financial institutions*. Jerome Levy Economics Institute, Bard College.

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision support systems*, *118*, 33-45.

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagao, T. (2016, December). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In *2016 15th ieee international conference on machine learning and applications (icmla)* (pp. 954-960). IEEE.

Platt, J. (2005, January). Fastmap, metricmap, and landmark mds are all nyström algorithms. In *International Workshop on Artificial Intelligence and Statistics* (pp. 261-268). PMLR.

Porras, E. R. G., Anchundia, S., & Vieira, M. (2002). Competencia, Rivalidad y Entrada del Capital Extranjero en la Banca Venezolana. *Observatorio de la Economía Latinoamericana*.

Purdy M. & Daugherty, P. (2016). Why artificial intelligence is the Future of Growth. Accenture, Available online at https://www.accenture.com/t20170524t055435__w__/ca-en/_acnmedia/pdf52/accenture-why-ai-is-the-future-of-growth.pdf

Riani, M., & Atkinson, A. C. (2007). Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in data analysis and classification*, *1*(2), 123-141.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American statistical association, 79(388), 871-880.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323-2326.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, *100*(5), 401-409.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. *Journal of the American Statistical Association*, *92*(438), 426-435.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS one*, *10*(10), e0139427.

Smith, S. W. (2010). An experiment in bibliographic mark-up: Parsing metadata for XML export. In *Proceedings of the 3rd. annual workshop on Librarians and*

Computers (LAC'10, Vol. 3), Reginald N. Smythe and Alexander Noble (Eds.). Paparazzi Press, Milan Italy (pp. 422-431).

Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in Bangladesh. *Lifetime Data Analysis*, *9*(2), 155-174.

Takatsuka, M. (2001, September). An application of the Self-Organizing Map and interactive 3-D visualization to geospatial data. In *Proceedings of the 6th International Conference on GeoComputation* (pp. 24-26).

Támara Ayús, A. L., Villarraga Peña, A. M., & Vera Álvarez, Y. C. (2017). El análisis factorial y el análisis discriminante en la estimación de la pérdida esperada para una institución financiera.

Támara-Ayús, A., Aristizábal, R., & Velásquez, E. (2012). Matrices de transición en el análisis del riesgo crediticio como elemento fundamental en el cálculo de la pérdida esperada en una institución financiera colombiana. *Revista Ingenierías Universidad de Medellín*, *11*(20), 105-114.

Tamboli, J., & Shukla, M. (2016, March). A survey of outlier detection algorithms for data streams. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 3535-3540). IEEE.

Tenenbaum, J. B., Silva, V. D., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, *290*(5500), 2319-2323.

Tobias, A., & Brunnermeier, M. K. (2016). CoVaR. *The American Economic Review*, *106*(7), 1705.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401-419.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*(66-71), 13.

Wahyudin, I., Djatna, T., & Kusuma, W. A. (2016). Cluster analysis for SME risk analysis documents based on Pillar K-Means. *Telkomnika*, *14*(2), 674.

Wang, D., Zhang, Z., Bai, R., & Mao, Y. (2018). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. *Journal of Computational and Applied Mathematics*, *329*, 307-321.

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *Ieee Access*, *7*, 107964-108000.

Yildirim, Y. (2008). Estimating default probabilities of CMBS loans with clustering and heavy censoring. *The Journal of Real Estate Finance and Economics*, *37*(2), 93-111.

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). Data mining and analysis: fundamental concepts and algorithms. *Cambridge University Press*.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, *42*(7), 3508-3516.

Zhu, S., Chan, J., & Bright, D. (2019). Applying Machine Learning for Troubleshooting Credit Exposure and xVA Profiles. *Available at SSRN 3404863*.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(5), 363-387.