ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία

Βελτιστοποίηση Ανάλυσης Δεδομένων με Χρήση Πολυκριτήριων Μεθόδων Ανάλυσης Αποφάσεων και Μηχανικής Μάθησης

Αλέξανδρος Μιχαήλ

Χανιά 2020

Διπλωματική Εργασία

Βελτιστοποίηση Ανάλυσης Δεδομένων με Χρήση Πολυκριτήριων Μεθόδων Ανάλυσης Αποφάσεων και Μηχανικής Μάθησης

Αλέξανδρος Μιχαήλ

Λαγουδάκης Μιχαήλ	Ζερβάκης Μιχαήλ	Ματσατσίνης Νικόλαος
Αναπληρωτής Καθηγητής	Καθηγητής	Καθηγητής

Εξεταστική Επιτροπή

Περίληψη

Στον κόσμο των Μεγάλων Δεδομένων (Big Data), όπου ο όγκος αυτών αυξάνεται με ραγδαίους ρυθμούς, το ερευνητικό ενδιαφέρον επικεντρώνεται στην ανακάλυψη και εφαρμογή νέων και πιο αποτελεσματικών μεθόδων ανάλυσης δεδομένων. Στα πλαίσια αυτά διερευνώνται μέθοδοι από πολλούς ερευνητικούς χώρους, όπως η επιχειρησιακή έρευνα (βελτιστοποίηση, πολυκριτήρια ανάλυση, κλπ.), καθώς και νέοι τρόποι εφαρμογής, αλλά και συνεργασίας μεθόδων με την ανάπτυξη νέων ταξινομητών (classifiers) ή συνόλων ταξινομητών (ensemble classifiers). Σκοπός αυτής της διπλωματικής εργασίας είναι να αναπτυχθεί ένα σύστημα που θα υποστηρίζει την εφαρμογή μεθόδων βελτιστοποίησης στην ανάλυση Μεγάλων Δεδομένων (Big Data Analysis), συνδυάζοντας τις πολυκριτήριες μεθόδους ανάλυσης αποφάσεων TOPSIS, UTASTAR και UTADIS με μεθόδους μηχανικής μάθησης, όπως η k-means. Αρχικά, συγκεντρώθηκαν σύνολα δεδομένων από το διαδίκτυο και εν συνεχεία ακολούθησαν οι φάσεις προ-επεξεργασίας και διαμόρφωσης των τελικών αρχείων δεδομένων. Στα τελικά αρχεία δεδομένων, εφαρμόστηκαν οι αλγόριθμοι που έχουν υλοποιηθεί στο σύστημα, με στόχο την υποστήριξη εξαγωγής γνώσης από δεδομένα σε προβλήματα κατάταξης (ranking) και ταξινόμησης / συσταδοποίησης (classification / clustering) δεδομένων. Τέλος, έγινε αξιολόγηση των αποτελεσμάτων εφαρμογής των ανωτέρω ταξινομητών με στόχο την εξαγωγή χρήσιμων συμπερασμάτων και προτάσεων για μελλοντική έρευνα.

Keywords: Data Mining, Big Data Analysis, Multi Criteria Decision Analysis Methods, TOPSIS, UTASTAR, UTADIS, Machine Learning, Ensemble Methods.

Technical University of Crete

School of Electrical and Computer Engineering

Diploma Thesis

Title

Data Analysis Optimization Using Multi-Criteria Decision Analysis and Machine Learning Methods

Author

Alexandros Michail

Abstract

In the Big Data world, where information continuously and rapidly keeps increasing, research interest focuses on the discovery of new and more effective methods for data analysis. On this premise, studies are being combined and conducted from multiple research fields, such as Business Administration (Optimization, Multi-Criteria Aid, etc.), as well as new ways of implementation and collaboration with the development of new classifiers or ensemble classifiers. The purpose of this thesis is to develop a system that supports an implementation that optimizes Big Data Analysis, combining the Multi-Criteria Decision Aid methods TOPSIS, UTASTAR and UTADIS with machine learning methods, such as k-means. First off, various datasets were gathered from open data libraries on the web, followed by the Extract-Transform-Load (ETL) procedure. The methods implemented on the system were applied on these datasets with the purpose to support the greater knowledge extraction from data of ranking and classification/clustering problems. Lastly, the results of the aforementioned classifiers were evaluated for greater information gain and future research proposals.

Keywords: Data Mining, Big Data Analysis, Multi Criteria Decision Analysis Methods, TOPSIS, UTASTAR, UTADIS, Machine Learning, Ensemble Methods.

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή 7

1.1 Στόχος της διπλωματικής εργασίας 7

1.2 Σχετικές Εργασίες 8

Κεφάλαιο 2 Θεωρητικό Υπόβαθρο Πολυκριτήριων Μεθόδων 11

2.1 Εισαγωγή (Θεωρία Αποφάσεων) 11

2.2 Η Αναλυτική-Συνθετική Προσέγγιση 13

2.3 Η μέθοδος UTA 14

2.4 Η μέθοδος UTASTAR 19

2.5 Η μέθοδος UTADIS 22

2.6 Η μέθοδος TOPSIS 25

Κεφάλαιο 3 Μέθοδοι Μηχανικής Μάθησης 28

3.1 Εισαγωγή 28

3.2 Clustering - Συσταδοποίηση 28

3.3 Αλγόριθμος k-means 29

Κεφάλαιο 4 Προτεινόμενη Μεθοδολογία 31

4.1 Εισαγωγή 31

4.2 Υλοποίηση πολυκριτήριων μεθόδων και μεθόδων μηχανικής μάθησης 31

4.3 Η Μεθοδολογία κατασκευής του Μοντέλου 32

Κεφάλαιο 5 Σύγκριση και Παρουσίαση Αποτελεσμάτων 36

5.1 Πολυκριτήριοι Πίνακες Δεδομένων 36

5.2 Αρχική Ανάλυση Διαστάσεων 42

5.3 Πολυκριτήριες Μέθοδοι και Αποτελέσματα στα Αρχικά Δεδομένα 46

5.3.1 UTASTAR Weights and Classification Accuracy 46

5.3.2 UTADIS Weights and Classification Accuracy 48

5.3.3 TOPSIS with UTASTAR weights and classification 49

5.3.4 TOPSIS with UTADIS weights and classification 51

5.4 Μέθοδοι Μηχανικής Μάθησης και Αποτελέσματα 53

5.4.1 Ταξινόμηση k-means 53
5.5 Συνδυασμός Μεθόδων και Αποτελέσματα 57
5.5.1 UTASTAR & k-means 57
5.5.2 UTADIS & k-means 60
5.5.3 TOPSIS with UTASTAR Weights & k-means 63
5.5.4 TOPSIS with UTADIS Weights & k-means 66
5.6 Μείωση διαστάσεων Πολυκριτήριων Μεθόδων 70
5.6.2 UTADIS Feature Reduction 72
5.7 Συνδυασμός Μεθόδων και Τελικά Αποτελέσματα με Μειωμένες Διαστάσεις 74
5.7.1 k-means with original data and feature reduction 74
5.7.2 UTASTAR & k-means 77
5.7.3 UTADIS & k-means 80
5.7.4 TOPSIS with UTASTAR weights & k-means 83
5.7.5 TOPSIS with UTADIS weights & k-means 86
5.8 Συγκρίσεις και Ολικά αποτελέσματα 89
Κεφάλαιο 6 Συμπεράσματα και Μελλοντικές Προεκτάσεις 91
6.1 Συμπεράσματα αποτελεσμάτων 91
6.2 Μελλοντικές προεκτάσεις 91
Βιβλιογραφία 92

Κεφάλαιο 1 Εισαγωγή

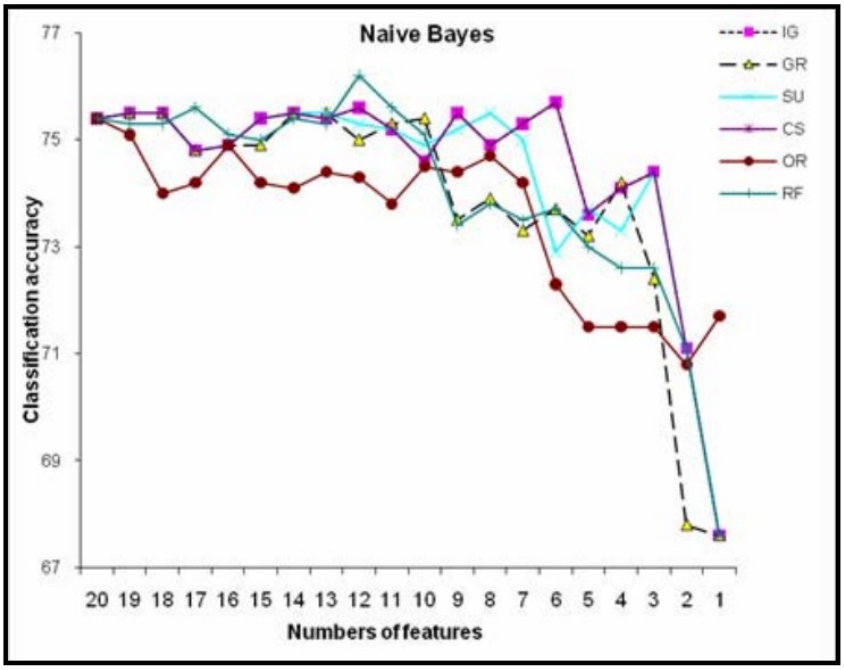
1.1 Στόχος της διπλωματικής εργασίας

Βαδίζοντας στο κόσμο των Μεγάλων Δεδομένων (Big Data), ο οποίος συνεχώς αυξάνεται με ραγδαίους ρυθμούς και το ερευνητικό ενδιαφέρον στον τομέα βρίσκεται σε έξαρση, προσπαθήσαμε σε αυτή την έρευνα να ανακαλύψουμε και να εφαρμόσουμε συνδυαστικά νέες και πιο αποτελεσματικές μεθόδους ανάλυσης δεδομένων. Πλέον, έχοντας στην διάθεσή μας τα εργαλεία (Python) να χειριστούμε αυτή την μεγάλη ποσότητα δεδομένων, είχαμε την δυνατότητα να χρησιμοποιήσουμε μεθόδους από πολλούς ερευνητικούς χώρους, όπως η επιχειρησιακή έρευνα (βελτιστοποίηση, πολυκριτήρια ανάλυση, κλπ.), καθώς και νέους τρόπους εφαρμογής, αλλά και συνεργασίας, μεθόδων. Σκοπός αυτής της διπλωματικής εργασίας είναι να γίνουν τα αρχικά βήματα της διαδικασίας, έτσι ώστε να αναπτυχθεί ένα σύστημα που θα υποστηρίζει την εφαρμογή μεθόδων βελτιστοποίησης στην Ανάλυση Μεγάλων Δεδομένων (Big Data Analysis), συνδυάζοντας τις πολυκριτήριες μεθόδους ανάλυσης αποφάσεων TOPSIS, UTASTAR και UTADIS με μεθόδους μηχανικής μάθησης, όπως η k-means, κ.α.. Αρχικά, έγινε η κατάλληλη σχετική έρευνα με την υπάρχουσα βιβλιογραφία και το θεωρητικό υπόβαθρο που χρειάζεται συγκεκριμένα για τον κλάδο της αξιολόγησης χρηματο-οικονομικής πίστωσης (credit scoring), και έπειτα συγκεντρώθηκαν σύνολα δεδομένων (datasets) από το διαδίκτυο (UCI German Credit Score, UCI Taiwan Credit Score, UCI Australian Credit Score) και εν συνεχεία ακολούθησαν οι φάσεις προ-επεξεργασίας και διαμόρφωσης των τελικών αρχείων δεδομένων. Στα δεδομένα αυτά εφαρμόστηκαν οι αλγόριθμοι που έχουμε υλοποιήσει (UTASTAR, UTADIS, TOPSIS, k-means), με στόχο την υποστήριξη και την βελτιστοποίηση ακρίβειας εξαγωγής γνώσης από δεδομένα σε προβλήματα κατάταξης (ranking) και ομαδοποίησης / συσταδοποίησης (classification / clustering) δεδομένων. Τέλος, έγινε αξιολόγηση των αποτελεσμάτων εφαρμογής των ανωτέρω μεθόδων με επιτυχία, κατά την οποία παρατηρήθηκε βελτίωση στην ακρίβεια ομαδοποίησης και ταξινόμησης με πιο ξεκάθαρο στόχο για περαιτέρω εξαγωγή χρήσιμων συμπερασμάτων και προτάσεων για μελλοντικές έρευνες.

1.2 Σχετικές Εργασίες

Σχετικές έρευνες χρησιμοποιώντας συνδυαστικά μεθόδους μηχανικής μάθησης και πολυκριτήριες μεθόδους υπάρχουν σε διάφορους τομείς και προβλήματα. Θα δούμε αναλυτικά τις περιπτώσεις μείωσης διαστάσεων, βελτιστοποίησης κατάταξης, ταξινόμησης, συσταδοποίησης και της συνδυαστικά μεταξύ τους σύγκρισης αποτελεσμάτων.

Συγκεκριμένα, στο Toward Optimal Feature Selection [1], χρησιμοποίησαν δεδομένα χρηματο-οικονομικών (credit card approval) και προσπάθησαν να λύσουν το πρόβλημα της μείωσης διαστάσεων για καλύτερη ακρίβεια σε μεθόδους ταξινόμησης (IB1, Naive Bayes, C4.5 decision tree, RBF network), συγκρίνοντας διάφορες μεθόδους κατάταξης, είτε στατιστικές, είτε βασιζόμενες στην εντροπία (Information Gain, Gain Ratio, Symmetrical Uncertainty, Chi-Squared, One-R, Relief-F). Χρησιμοποίησαν δηλαδή δύο σύνολα δεδομένων με παρόμοιου τύπου διαστάσεις (ποσοτικές και ποιοτικές) και την σειρά κατάταξης για κάθε μία από τις μεθόδους, συγκρίνοντας την ακρίβεια των μεθόδων ταξινόμησης, αλλά σε κάθε επανάληψη χρησιμοποιώντας μια διάσταση λιγότερη. Για παράδειγμα στο Σχήμα 1.2.1 για Naive Bayes βλέπουμε την ακρίβεια των μεθόδων κατάταξης ως προς τις διαστάσεις που χρησιμοποιήθηκαν.



Σχήμα 1.2.1: Σειρά κατάταξης μεθόδων και ακρίβεια ταξινόμησης ως προς τις διαστάσεις για το German credit dataset, με Naive Bayes ταξινομητή.

Εν τέλει, είχαν ως αποτέλεσμα τη μη σαφή λύση για την καλύτερη μέθοδο κατάταξης για μείωση διαστάσεων, αλλά την επιβεβαίωση της ανάγκης για σύγκριση και δοκιμή μεταξύ πολλών μεθόδων και συνόλων διαστάσεων για καλύτερη ακρίβεια.

Επίσης, στο [2], έφτιαξαν ένα μοντέλο για να καθορίσουν το CLV (Customer Lifetime Value). Αφού κανονικοποίησαν τα δεδομένα με min-max (0-1), χρησιμοποίησαν την πολυκριτήρια μέθοδο FAHP (Fuzzy Analytical Hierarchy Process) για να αναθέσουν βάρη στα κριτήρια Recency, Frequency and Monetary (RFM) και μετέπειτα τη μέθοδο μηχανικής μάθησης k-means για συσταδοποίηση πελατών, όπου οι πέντε κλάσεις κατατάχθηκαν με πολυκριτήρια μέθοδο TOPSIS βάσει των RFM κριτηρίων. Έτσι, βάσει της κατάταξης καθορίζουν τον προορισμό των πόρων της εταιρείας. Για να παρουσιάσουν την έρευνά τους, συγκέντρωσαν δεδομένα από εταιρείες διαφήμισης καλλυντικών.

Στην έρευνα [3] προσπάθησαν να δημιουργήσουν ένα κατάλληλο credit scoring μοντέλο, βρίσκοντας τις παραμέτρους που το επηρεάζουν περισσότερο. Μείωση διαστάσεων για τα δεδομένα έγινε με τη συμβολή ειδικών του τομέα και με τη χρήση στοιχείων από αναλύσεις ιστογραμμάτων και συναρτήσεων διασποράς. Έγινε βελτιστοποίηση της συσταδοποίησης k-means με πέντε συστάδες και τέλος χρήση πολυκριτήριας μεθόδου UTADIS για ταξινόμηση των κριτηρίων. Το μοντέλο αυτό συγκρίθηκε με μεθόδους ταξινόμησης μηχανικής μάθησης (CART, C5, ANN, ANN1, ANN2, ANN3, RL, KNN, and BN) και αποδείχθηκε το καλύτερο σε ακρίβεια πρόβλεψης.

Η έρευνα [4] είχε σκοπό την εύρεση του βέλτιστου αριθμού συστάδων. Για ένα σύνολο δεδομένων χρησιμοποίησαν τις πολυκριτήριες μεθόδους PROMETHEE II, WSM, και TOPSIS για να κατατάξουν τα αποτελέσματα συσταδοποίησης k-means από 15 διαφορετικά σύνολα δεδομένων, με εναλλακτικές στον αριθμό των συστάδων και ως κριτήρια την απόδοση του k-means για κάθε σύνολο δεδομένων. Τα αποτελέσματα έδειξαν ότι οι πολυκριτήριες μέθοδοι είναι αρκετά τελεσφόρες στην εκτίμηση του αριθμού των συστάδων και αποδίδουν καλύτερα από τα δέκα κριτήρια που δημιουργήθηκαν συγκεκριμένα για σύγκριση.

Σε μια άλλη έρευνα [5], θέλοντας να λύσουν το πρόβλημα του credit score classification, μετατρέπουν την γνωστή πολυκριτήρια μέθοδο TOPSIS σε μέθοδο ταξινόμησης. Αρχικά, χρησιμοποιώντας logistic regression για να μειώσουν διαστάσεις και μετέπειτα αναπτύσσοντας έναν τρόπο κοντά στη λογική της TOPSIS για να καθορίζουν τα βάρη για τα κριτήρια. Τέλος, παραδεχόμενοι ότι όσο πιο ψηλό credit score έχει κάποιος τόσο το καλύτερο, θέτουν ένα όριο στην ταξινόμηση, χρησιμοποιώντας τα δυο ευρέως γνωστά σύνολα δεδομένων για credit score classification από το UCI. Μετέπειτα, η μέθοδος συγκρίνεται με τις μεθόδους linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), decision tree (DT), logistic regression (LogR), k-nearest neighbor classifier (k-NN), support vector machine (SVM) και least-squares support vector machine (LSSVM) και αποδίδει καλά αποτελέσματα βάσει κριτηρίων ακρίβειας, πολυπλοκότητας και διερμηνείας.

Η έρευνα [6] χρησιμοποιεί ordinary least-squares (OLS) regression για τη δημιουργία βαρών για τα κριτήρια και έπειτα καθορίζει το φράγμα ταξινόμησης από το closeness coefficient, αφού δημιουργήθηκε το credit score. Καθορίστηκε μη αναγκαία η μείωση διαστάσεων, καθώς

επηρέαζε την ακρίβεια πρόβλεψης. Το μοντέλο δοκιμάστηκε σε σύνολα δεδομένων Καναδικής τράπεζας και τα αποτελέσματα συγκρίθηκαν με decision tree και με Monte Carlo simulation, όπου και η TOPSIS έδωσε καλύτερα αποτελέσματα.

Οι ερευνητές στο [7] χρησιμοποίησαν συνάρτηση πολυωνύμων, αντί συνάρτησης μερικών χρησιμοτήτων, έτσι ώστε να μπορεί να χρησιμοποιεί μεταβλητές ονομαστικές και διάταξης. Υπολογίζοντας τους πολυωνυμικούς συντελεστές, φράγματα και βάρη κριτηρίων, μειώθηκαν ως αποτέλεσμα τα λάθη της ταξινόμησης. Άγνωστες παράμετροι για την ταξινόμηση εκτιμώνται με την χρήση ενός συνδυαστικού αλγορίθμου Particle Swarm Optimization (PSO) algorithm και Genetic Algorithm (GA). Η μέθοδος δοκιμάστηκε σε διάφορα σύνολα δεδομένων και συγκριτικά με άλλες μεθόδους αποδείχθηκε υψηλή η αποδοτικότητά της.

Κεφάλαιο 2 Θεωρητικό Υπόβαθρο Πολυκριτήριων Μεθόδων

2.1 Εισαγωγή (Θεωρία Αποφάσεων)

Για την καλύτερη κατανόηση της έννοιας των Πολυκριτήριων Μεθόδων δόκιμη είναι η εκκίνηση από την Θεωρία Αποφάσεων. Η λήψη αποφάσεων αποτελεί το πιο συνηθισμένο φαινόμενο της δραστηριότητας του ανθρώπου, από το πιο απλό μέχρι το πιο πολύπλοκο τμήμα της καθημερινότητάς του, μια ξεκάθαρη έννοια, όπως αναφέρεται και στο βιβλίο του κ. Ματσατσίνη: «Σαν απόφαση θεωρούνται όλες εκείνες οι ενέργειες που γίνονται από έναν ή περισσότερους ανθρώπους με στόχο την επιλογή ενός τρόπου ενέργειας-δράσης μέσα από ένα σύνολο εναλλακτικών επιλογών. Απόφαση έχουμε όταν ο αποφασίζων έχει τη δυνατότητα επιλογής μεταξύ τουλάχιστον δύο διαφορετικών εναλλακτικών ενεργειών» [22].

Η λήψη αποφάσεων γίνεται σε πολύ μεγάλο βαθμό με ορθολογικό τρόπο. Αυτό σημαίνει ότι ακολουθώντας σαφείς και ξεκάθαρους κανόνες, και λαμβάνοντας υπ' όψιν τις διαθέσιμες πληροφορίες, τις γνώσεις και την εμπειρία του αποφασίζοντα, αναλύεται το εξεταζόμενο πρόβλημα και οδηγούμαστε στην επιλογή μιας λύσης. Ένα μοντέλο ορθολογικής λήψης αποφάσεων ακολουθεί τα εξής βασικά βήματα:

Βήμα 1: Καθορισμός του προβλήματος

Βήμα 2: Καθορισμός κριτηρίων απόφασης

Βήμα 3: Απόδοση βαρών στα κριτήρια

Βήμα 4: Καθορισμός εναλλακτικών επιλογών

Βήμα 5: Εκτίμηση κάθε εναλλακτικής σε κάθε κριτήριο

Βήμα 6: Υπολογισμός της βέλτιστης απόφασης

Μεγάλο μέρος των προς επίλυση προβλημάτων έχει πολυδιάστατη φύση και χαρακτηρίζεται από αδυναμία αντιμετώπισης με χρήση ενός μόνο κριτηρίου, κάτι που οδήγησε στην ανάπτυξη της Πολυκριτήριας Λήψης Αποφάσεων. Τα προβλήματα αυτά αναλύονται σε πολλαπλά κριτήρια, τα οποία συνήθως χαρακτηρίζονται από μεγάλη πολυπλοκότητα στις μεταξύ τους σχέσεις, κάτι που κάνει εξαιρετικά δύσκολη για τον αποφασίζοντα την τελική επιλογή. Συνεπώς, δημιουργείται η ανάγκη για υποστήριξη του αποφασίζοντα μέσω της ανάπτυξης των κατάλληλων πολυκριτήριων μοντέλων.

Η πολυκριτήρια λήψη αποφάσεων χρησιμοποιεί ένα σύστημα αξιών, προκειμένου να ερμηνεύσει τις προτιμήσεις που έχουν οι αποφασίζοντες σε ένα σύνολο εναλλακτικών επιλογών. Το σύστημα αξιών συμβάλλει στη διαμόρφωση μιας συνάρτησης χρησιμότητας και των σχετικών βαρών των προτεραιοτήτων των κριτηρίων. Σύμφωνα με αυτή, οι προτιμήσεις που εκφράζει ο αποφασίζων για ένα σύνολο εναλλακτικών επιλογών, λαμβάνονται υπόψη, ώστε να δημιουργηθεί ένα σύστημα αξιών, που ικανοποιεί ένα σύνολο από συνθήκες και που θα υποβοηθήσει τον αποφασίζοντα να οδηγηθεί στη σωστότερη λύση.

Πολύ σημαντικό να αναφέρουμε ότι η πολυκριτήρια λήψη αποφάσεων δεν χρειάζεται πληθώρα δεδομένων για να δώσει αντιπροσωπευτικές λύσεις, κάτι που θα αξιοποιήσουμε ιδιαίτερα στην παρούσα εργασία.

Κύριοι στόχοι της πολυκριτήριας ανάλυσης είναι:

- Ο καθορισμός των συνθηκών, με την ικανοποίηση των οποίων, το σύστημα αξιών δύναται να υφίσταται.
- Η υποστήριξη του αποφασίζοντα ώστε να ανακαλύψει, μέσω μιας διαδικασίας, ένα σύστημα αξιών και να πάρει στη συνέχεια μια σωστή απόφαση.

Κριτήρια

Στην πολυκριτήρια ανάλυση χρησιμοποιούνται διαφορετικοί τύποι κριτηρίων και είδη δεδομένων. Αυτοί που αναφέρονται και χρησιμοποιούνται στην παρούσα διπλωματική είναι οι εξής:

- **Ονομαστικά ή κατηγορικά (nominal)** που είναι τα δεδομένα που δεν έχουν καμία ιδιότητα, π.χ. χρώμα ματιών, φύλο, τόπος γέννησης.
- **Διάταξης (ordinal)** είναι τα δεδομένα που είναι εφικτό να καθορίσουμε μια σειρά, π.χ. σειρά κατάταξης σε ένα αγώνισμα, επίπεδο εκπαίδευσης, κλίμακα σεισμών RICHTER, κτλ.
- **Διαστήματος (interval)** είναι τα δεδομένα των οποίων οι διαφορές μεταξύ των τιμών τους έχουν νόημα και μπορεί να οριστεί μια υποδιάταξη για αυτά π.χ. θερμοκρασία, ηλικία, κλίμακα βαθμών Κελσίου, κτλ.
- **Λόγου ή Αναλογίας (ratio)** είναι τα δεδομένα που το μέγεθος της τιμής τους αντιστοιχεί με το χαρακτηριστικό που υποδεικνύουν, συμπεριλαμβανομένου και του μηδέν, π.χ. ταχύτητα, ύψος.

2.2 Η Αναλυτική-Συνθετική Προσέγγιση

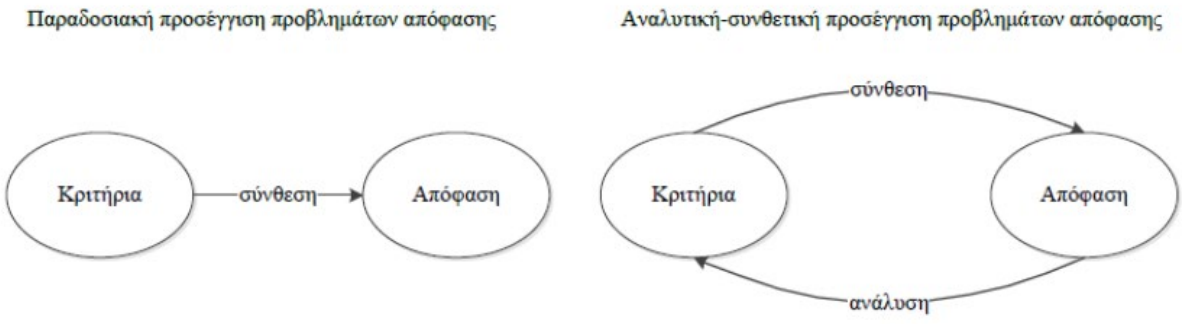
Η μέθοδος της Αναλυτικής-Συνθετικής Προσέγγισης (ΑΣΠ) είναι ένα πολύ καθοριστικό κομμάτι των πολυκριτήριων συστημάτων αποφάσεων. Βασίζεται σε ένα σύστημα αξιών και προτιμήσεων με το οποίο ο αποφασίζων λαμβάνει αποφάσεις (συνειδητά ή ασυνείδητα). Διερευνά τη σχέση μεταξύ των αποτελεσμάτων των εναλλακτικών και προτιμήσεων των κριτηρίων και καταλήγει στην μέθοδο με την οποία λαμβάνονται αυτές οι αποφάσεις, με αποτέλεσμα το μοντέλο υπόδειγμα σύνθεσης των κριτηρίων.

Συγκεκριμένα, η Αναλυτική-Συνθετική Προσέγγιση αναλύει τα δεδομένα για να εντοπίσει το υπόδειγμα που αναπαριστά, όσο πιο πιστά γίνεται, το σύστημα αξιών και προτιμήσεων του αποφασίζοντα, σε αντίθεση με την λογική που πρώτα γίνεται η σύνθεση των δεδομένων για ένα πρόβλημα, έτσι ώστε να καταλήξουν στο τελικό αποτέλεσμα που χρησιμοποιούν άλλες πολυκριτήριες προσεγγίσεις.

Για την βέλτιστη προδιαγραφή αυτού του υποδείγματος είναι πολύ σημαντική η συγκέντρωση σχετικής πληροφορίας βάσει του συστήματος αξιών και προτιμήσεων, όπως επίσης και η συγκέντρωση και η ανάλυση παραδειγμάτων από τις αποφάσεις που παίρνει ο αποφασίζων.

Οι ίδιες οι αποφάσεις είναι συνήθως οι πληροφορίες, και εκφράζονται σε διάφορες μορφές. Τα παραδείγματα μπορεί να είναι ένα σύνολο πραγματικών, κατασκευασμένων συγκεκριμένα για το σκοπό αυτό, εναλλακτικών επιλογών, οι οποίες έχουν βαθμολογηθεί ως προς τις επιδόσεις σε ένα σύνολο κριτηρίων. Αφού ο αποφασίζων εκφράσει τις σχέσεις προτίμησης μεταξύ των εναλλακτικών αυτών, έχουμε σαν αποτέλεσμα το σύστημα αξιών που χρησιμοποιήθηκε για την λήψη της απόφασης.

Η διαφορά μεταξύ της παραδοσιακής και της αναλυτικής-συνθετικής προσέγγισης φαίνεται στο Σχήμα 2.2.1.



Σχήμα 2.2.1 Παραδοσιακή και Αναλυτική-Συνθετική Προσέγγιση

2.3 Η μέθοδος UTA

Βασισμένη στην αρχή της Αναλυτικής-Συνθετικής Προσέγγισης που αναλύσαμε πιο πάνω, η μέθοδος UTA (UTility Additives) [8] αποτελεί μέρος των πιο γνωστών πολυκριτήριων μέθοδων. Λαμβάνοντας υπόψη ένα σύνολο εναλλακτικών επιλογών, και μια συγκεκριμένη προ-διάταξη από τον αποφασίζοντα, η μέθοδος αυτή, έχει σκοπό την ανάπτυξη προσθετικών συναρτήσεων αξιών (additive value function) των κριτηρίων. Χρησιμοποιώντας αλγορίθμους γραμμικού προγραμματισμού για τον υπολογισμό των συναρτήσεων αυτών, έχει στόχο η κατάταξη που δημιουργείται σαν αποτέλεσμα να αντιστοιχεί όσο πιο πολύ γίνεται με την καθορισμένη προ-διάταξη που έχει δώσει ο αποφασίζων.

Θεωρώντας, την ύπαρξη του σύνολου εναλλακτικών $A = \{a_1, a_2, \dots, a_m\}$, που βαθμολογούνται σε ένα σύνολο κριτηρίων $\mathbf{g} = (g_1, g_2, \dots, g_n)$ δημιουργείται μία προσθετική συνάρτηση αξίας με μορφή:

$$U(\mathbf{g}) = \sum_{i=1}^{n} p_i u_i(g_i) \qquad \textbf{(2.3.1)}$$

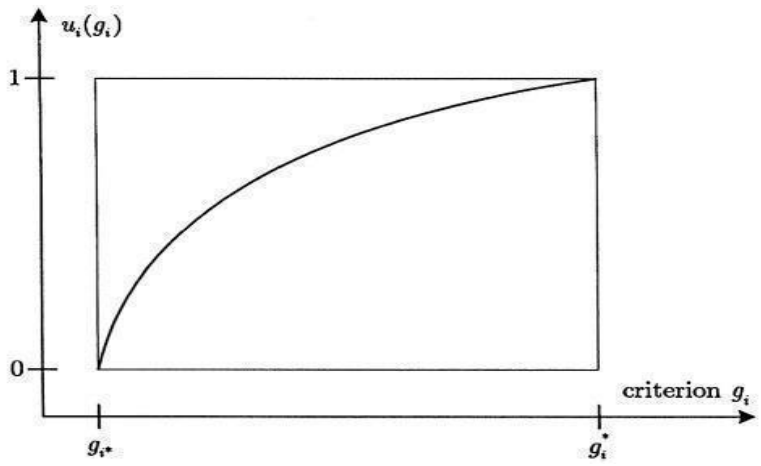
υπό τους περιορισμούς κανονικοποίησης:

- $\sum_{i=1}^{n} p_i = 1$
- $u_i(g_{i_*}) = 0, u_i(g_i^*) = 1, \forall i = 1,2, \dots, n;$

(2.3.2)

όπου

- u_i, i = 1,2, …, n οι κανονικοποιημένες στο διάστημα [0 ,1], μερικές συναρτήσεις αξίας ή χρησιμότητας.
- p_i τα βάρη των u_i και g_{i*} το κατώτατο όριο και g_i^* το ανώτατο, όπως φαίνεται και στο Σχήμα 2.3.1.



Σχήμα 2.3.1 Η καμπύλη της περιθώριας συνάρτησης αξίας [8]

Να αναφέρουμε ότι τόσο στις μερικές, αλλά και στις ολικές, συναρτήσεις χρησιμότητας ισχύουν οι παρακάτω ιδιότητες μονοτονίας :

$$U(\mathbf{g}(a)) > U(\mathbf{g}(b)) \Leftrightarrow a \succ b \text{ (Προτίμηση)}$$
$$U(\mathbf{g}(a)) = U(\mathbf{g}(b)) \Leftrightarrow a \sim b \text{ (Αδιαφορία)} \quad \textbf{(2.3.3)}$$

όπου ο συμβολισμός $a \succ b$ θα σημαίνει την Προτίμηση του a έναντι του b, και αντίστοιχα το $a \sim b$ την Αδιαφορία στην επιλογή μεταξύ τους.

Μια μορφή της προσθετικής συνάρτησης αξίας χωρίς βάρη, ισοδύναμη με τις μορφές των σχέσεων (2.3.1) και (2.3.2), είναι:

$$U(\mathbf{g}) = \sum_{i=1}^{n} u_i(g_i) \quad \textbf{(2.3.4)}$$

υπό τους περιορισμούς κανονικοποίησης:

$$\sum_{i=1}^{n} u_i(g_i^*) = 1 \text{ και } u_i(g_{i_*}) = 0, \forall i = 1,2,\dots,n; \quad \textbf{(2.3.5)}$$

Για να ισχύει η προσθετική μορφή στην ύπαρξη ενός τέτοιου μοντέλου, μια βασική προϋπόθεση μεταξύ άλλων συνθήκων, είναι η προτίμηση ανεξαρτησίας στα κριτήρια του αποφασίζοντα.

Η αξία κάθε εναλλακτικής $a \in A$ μπορεί να εμφανιστεί και ως:

$$U'\big(\mathbf{g}(a)\big) = \sum_{i=1}^{n} u_i\big(g_i(a)\big) + \sigma(a)\ \forall a \in A \quad \textbf{(2.3.6)}$$

Βάσει των συνθηκών (2.3.3) και από τις (2.3.4), (2.3.5).

όπου $\sigma(a)$ είναι ένα πιθανό σφάλμα της $U(\mathbf{g}(a))$ σε σχέση με την $U'(\mathbf{g}(a))$.

Χρησιμοποιείται γραμμική παρεμβολή για τον προσδιορισμό των συναρτήσεων μερικών (περιθωρίων) χρησιμοτήτων, που είναι γραμμικές συναρτήσεις. Με αποτέλεσμα για κάθε κριτήριο, το διάστημα $[g_{i*}, g_i^*]$ να διανέμεται σε $(a_i - 1)$ ίσα υπό διαστήματα, με τελικό σημείο g_i^j κάθε υποδιαστήματος να δίνεται από την σχέση:

$$g_i^j = g_{i_*} + \frac{j-1}{a_i - 1}\left(g_i^* - g_{i_*}\right) \ \forall j = 1,2,\dots,a_i \qquad \textbf{(2.3.7)}$$

Με χρήση γραμμικής παρεμβολής, υπολογίζεται η μερική αξία μιας εναλλακτικής a:

$$g_i(a) \in [g_i^j, g_i^{j+1}]:$$

$$u_i(g_i(a)) = u_i(g_i^j) + \frac{g_i(a) - g_i^j}{g_i^{j+1} - g_i^j}\left(u_i(g_i^{j+1}) - u_i(g_i^j)\right) \qquad \textbf{(2.3.8)}$$

Οι εναλλακτικές $A = \{a_1, a_2, \dots, a_m\}$ αναδιατάσσονται έτσι ώστε η προτιμότερη εναλλακτική a_1 να είναι η κορυφή στην κατάταξη και αντίστοιχα η εναλλακτική a_m η ρίζα. Εφόσον η κατάταξη έχει το σχήμα μιας προδιάταξης R, τότε για κάθε διαδοχικές εναλλακτικές (a_k, a_{k+1}) ισχύει:

$a_k > a_{k+1}$ (Προτίμηση),

ή

$a_k \sim a_{k+1}$ (Αδιαφορία).

Τότε για τη διαφορά αξίας των εναλλακτικών

$$\Delta(a_k, a_{k+1}) = U'(\mathbf{g}(a_k)) - U'(\mathbf{g}(a_{k+1})) \qquad \textbf{(2.3.9)}$$

ισχύει μια από τις περιπτώσεις:

$$\Delta(a_k, a_{k+1}) \geq \delta \text{ εάν } a_k \succ a_{k+1}$$

ή

$$\Delta(a_k, a_{k+1}) = 0 \text{ εάν } a_k \sim a_{k+1} \qquad \textbf{(2.3.10)}$$

όπου δ αντιπροσωπεύει ένας μικρός μη αρνητικό αριθμός με στόχο τον διαχωρισμό δύο διαδοχικών κλάσεων ισοδυναμίας της κατάταξης R.

Χρησιμοποιώντας την υπόθεση με την μονοτονία των προτιμήσεων, οι περιθώριες αξίες $u_i(g_i)$ πρέπει να ικανοποιούν το σύνολο των ακόλουθων περιορισμών:

$$u_i(g_i^{j+1}) - u_i(g_i^j) \geq s_i, \forall j = 1,2,\dots,a_{i-1}, \forall i = 1,2,\dots,n \quad \textbf{(2.3.11)}$$

όπου ως $s_i \geq 0$ ορίζονται τα κατώφλια αδιαφορίας για κάθε κριτήριο g_i και η χρήση τους συμβάλει σημαντικά στην αποφυγή περιπτώσεων, όπου για τις μερικές χρησιμότητες ισχύει

$$u_i\left(g_i^j\right) = u_i\left(g_i^{j+1}\right), \text{με } g_i^{j+1} > g_i^j.$$

Με στόχο την ελαχιστοποίηση του συνολικού προκαλούμενου σφάλματος, που είναι η αντικειμενική συνάρτηση, γίνεται ο υπολογισμός των περιθωρίων συναρτήσεων χρησιμότητας μέσω του παρακάτω γραμμικού προγραμματισμού (LP), με περιορισμούς από τις σχέσεις (2.3.4), (2.3.5), (2.3.10) και (2.3.11). Συνεπώς:

$$[min]F = \sum_{a \in A} \sigma(a)$$

Υπό τους περιορισμούς :

$$\Delta(a_k, a_{k+1}) \geq \delta \text{ εάν } a_k \succ a_{k+1} \; \forall \, k$$

$$\Delta(a_k, a_{k+1}) = 0 \text{ εάν } a_k \sim a_{k+1} \forall \, k$$

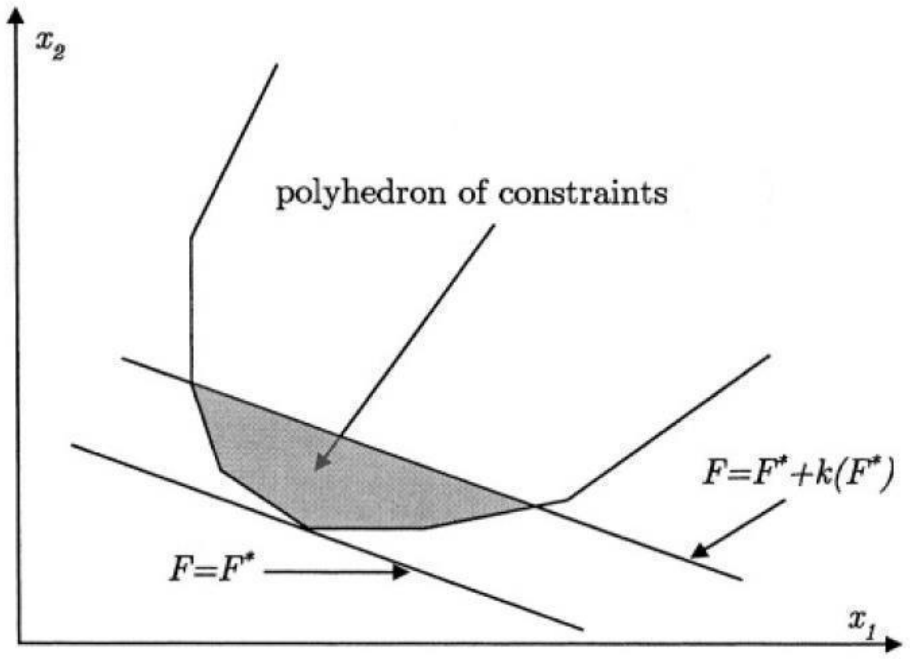
$$u_i(g_i^{j+1}) - u_i(g_i^j) \geq 0 \; \forall \, i, j$$

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

$$u_i(g_{i*}) = 0, u_i(g_i^j) \geq 0, \sigma(a) \geq 0 \; \forall a \in A, \forall \, i, j$$

(2.3.12)

Ως ένα πρόβλημα ανάλυσης μετα-βελτιστοποίησης αντιμετωπίζεται η ανάλυση ευστάθειας των αποτελεσμάτων του LP (2.3.12). Αν η βέλτιστη λύση είναι F^* = 0, τότε είναι διαθέσιμες πολλές συναρτήσεις αξίας που αντιστοιχούν ακριβώς με την προδιάταξη R, όπως φαίνεται από το υπερ-πολύεδρο των λύσεων για τα $u_i(g_i)$. Ακόμη και στην περίπτωση που η τιμή της F^* είναι μη μηδενική, υπάρχουν άλλες λύσεις, λιγότερο καλές για την F, που είναι σε θέση να βελτιώσουν άλλα εναλλακτικά κριτήρια βελτιστοποίησης (π.χ. συντελεστή συσχέτισης τ του Kendall [26]).



Σχήμα 2.3.2 Ανάλυση ευστάθειας (UTA)[8]

Στο Σχήμα 2.3.2 παρατηρείται ότι, ο χώρος των μετα-βέλτιστων λύσεων από το υπερ-πολύεδρο ορίζεται από:

$$F \leq F^* + k(F^*) \qquad \textbf{(2.3.13)}$$

και τους περιορισμούς του LP (2.3.12) με $k(F^*)$ να αντιπροσωπεύει ένα μη αρνητικό όριο, ως ένα μικρό ποσοστό του σφάλματος F^*.

Η αξιολόγηση–εξέταση των λύσεων–κορυφών του υπερ-πολυέδρου, μπορεί να πραγματοποιηθεί με διάφορες μεθόδους. Μερικές από αυτές, είναι:

- Κλάδου και φράγματος (branch and bound).
- Αντίστροφης simplex.
- Μέθοδος Manas/Nedoma

Χρησιμοποιώντας μια ευρετική μέθοδο αναζήτησης (ημι)βέλτιστων λύσεων, οι Jacquet-Siskos [9] αναλύουν το πολυέδρο (2.3.13), με τα παρακάτω LP:

$[min]\ u_i(g_i^*)$

και

$[max]\ u_i(g_i^*)$ στο πολύεδρο (2.3.13)

(2.3.14)

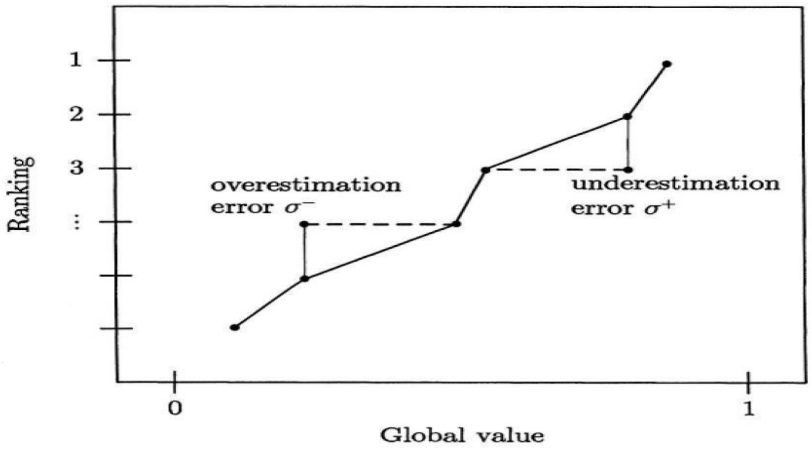
Η μέση τιμή των λύσεων των προηγούμενων LP, υπολογίζεται ως η τελική λύση του προβλήματος, που είναι και αυτή (ημι)βέλτιστη, λόγω της κυρτότητας του υπερ-πολυέδρου.

Αντίθετα σε περίπτωση αστάθειας, οι λύσεις των LP (2.3.14) έχουν μεγάλη διαφορά μεταξύ τους και έτσι γίνεται λιγότερο αντιπροσωπευτική η εκτιμώμενη μέση λύση.

Εν τέλει, έχουμε σαν αποτέλεσμα μια ένδειξη της σημαντικότητας αυτών των κριτηρίων στο σύστημα προτιμήσεων του αποφασίζοντος καθώς υποδεικνύεται από τις λύσεις αυτές η διακύμανση των βαρών των κριτηρίων g_i .

2.4 Η μέθοδος UTASTAR

Συνεχίζοντας, έχουμε από τους Siskos-Yannacopoulos [10] την μέθοδο UTASTAR που αποτελεί μια βελτίωση της μεθόδου UTA. Στην μέθοδο UTA όπως είδαμε, για κάθε εναλλακτική $a \in A$ ορίζεται ένα μοναδικό σφάλμα $\sigma(a)$ που δεν είναι αρκετό για την μείωση της ολικής εξάπλωσης των σημείων στη μονότονη καμπύλη. Αυτό γίνεται λόγω των σημείων που είναι δεξιά της καμπύλης, όπου θα ήταν καλύτερο να μειωθούν οι αξίες τους χωρίς να αυξηθούν οι αξίες των άλλων.



Σχήμα 2.4.1 - Καμπύλη μονότονης παλινδρόμησης [10].

Στη μέθοδο UTASTAR, εισάγεται μια διπλή θετική συνάρτηση σφάλματος μετατρέποντας τη σχέση (2.1) σε:

$$U'(\mathbf{g}(a)) = \sum_{i=1}^{n} u_i(g_i(a)) - \sigma^+(a) + \sigma^-(a) \; \forall a \in A \qquad \textbf{(2.15)}$$

με $\sigma^+(a)$ και $\sigma^-(a)$ τα σφάλματα υποεκτίμησης και υπερεκτίμησης.

Επίσης γίνεται μοντελοποίηση των περιορισμών μονοτονίας των κριτηρίων, χρησιμοποιώντας τους μετασχηματισμούς:

$$w_{ij} = u_i(g_i^{j+1}) - u_i(g_i^j) \geq 0 \; \forall i = 1,2, \ldots, n \; \kappa\alpha\iota \; j = 1,2, \ldots, a_i - 1 \qquad \textbf{(2.16)}$$

Έτσι οι συνθήκες μονοτονίας (2.11) αντικαθίστανται από περιορισμούς θετικότητας για τις μεταβλητές w_{ij}.

Οπότε, ο αλγόριθμος UTASTAR χρησιμοποιεί τα βήματα :

1ο Βήμα: Μέσω των παρακάτω σχέσεων οι ολικές χρησιμότητες των εναλλακτικών $U(g(a_k))$, $k = 1,2, \ldots, m$, εκφράζονται ως συναρτήσεις των περιθωρίων χρησιμοτήτων $u_i(g_i)$ και των μεταβλητών w_{ij}, όπως φαίνεται στο (2.16):

$$u_i(g_i^1) = 0 \; \forall \; i = 1,2, \ldots, n$$

$$u_i(g_i^j) = \sum_{t=1}^{j-1} w_{it} \; \forall i = 1,2, \ldots, n \; \kappa\alpha\iota \; j = 2,3, \ldots, a_i - 1 \qquad \textbf{(2.17)}$$

2ο Βήμα: Εισαγωγή δύο συναρτήσεων σφάλματος $\sigma^+(a)$ και $\sigma^-(a)$ στο A, έτσι ώστε για κάθε ζεύγος διαδοχικών εναλλακτικών στην προδιάταξη να έχουμε τις σχέσεις:

$$\Delta(a_k, a_{k+1}) = U(g(a_k)) - \sigma^+(a_k) + \sigma^-(a_k) - U(g(a_{k+1})) + \sigma^+(a_{k+1}) - \sigma^-(a_{k+1}) \qquad \textbf{(2.18)}$$

3ο Βήμα: Δημιουργείται το παρακάτω LP.

$$[min] z = \sum_{k=1}^{m} \sigma^+(a_k) + \sigma^-(a_k)$$

υπό τους περιορισμούς

$$\Delta(a_k, a_{k+1}) \geq \delta \text{ εάν } a_k \succ a_{k+1} \; \forall \, k$$

$$\Delta(a_k, a_{k+1}) = 0 \text{ εάν } a_k \sim a_{k+1} \forall \, k$$

$$\sum_{i=1}^{n} \sum_{j=1}^{a_i - 1} w_{ij} = 1 \qquad \textbf{(2.19)}$$

$$w_{ij} \geq 0, \qquad \sigma^{+}(a_k) \geq 0, \qquad \sigma^{-}(a_k) \geq 0 \; \forall i, j \; \kappa\alpha\iota \; k$$

όπου δ μικρός μη αρνητικός αριθμός.

4ο Βήμα: Γίνεται έλεγχος για τον εντοπισμό πολλαπλών βέλτιστων ή (ημι)βέλτιστων λύσεων στο LP (2.19), υπολογίζοντας τη μέση προσθετική συνάρτηση αξίας που μεγιστοποιεί τις ακόλουθες αντικειμενικές συναρτήσεις:

$$u_i(g_i^*) = \sum_{j=1}^{a_i - 1} w_{ij} \;\; \forall \, i = 1,2, \dots, n$$

(2.20)

που περιορίζεται από:

$$\sum_{k=1}^{m} \left(\sigma^{+}(a_k) + \sigma^{-}(a_k) \right) \leq z^* + \varepsilon$$

(2.21)

όπου z^* είναι η βέλτιστη τιμή (σφάλμα) της αντικειμενικής συνάρτησης του LP και “ε” είναι ένας πολύ μικρός θετικός αριθμός ή μηδέν.

2.5 Η μέθοδος UTADIS

Ακόμη μια διαφορετική ανάπτυξη της μεθόδου UTA είναι η πολυκριτήρια μέθοδος UTADIS (UTilités Additives DIScriminantes) [12] και σκοπός δεν είναι η κατάταξη των εναλλακτικών δραστηριοτήτων, αλλά η ταξινόμησή τους σε προκαθορισμένες ομοιογενείς κατηγορίες διατεταγμένες από τις καλύτερες προς τις χειρότερες ως εξής:

$$C_1 \succ C_2 \succ \ldots \succ C_q$$

Με C_1 να συμβολίζεται η προτιμότερη κατηγορία έναντι των υπόλοιπων κατηγοριών και αντίστοιχα η τελευταία κατηγορία C_q που θεωρείται από τις χειρότερες επιλογές.

Στόχος η σύνθεση των κριτηρίων, με αποτέλεσμα τις μεγάλες ολικές χρησιμότητες στις εναλλακτικές της κατηγορίας C_1 και χαμηλότερες, στις εναλλακτικές που είναι στις χαμηλότερες κατηγορίες.

Η προσθετική συνάρτηση χρησιμότητας που χρησιμοποιεί η UTADIS έχει την μορφή:

$$U(g) = \sum m_i u_i(g_i)$$

(2.24)

με κριτήρια αξιολόγησης $g = (g_1, g_2, \ldots g_n)$

$$g_i \left(\sum_{i=1}^{n} m_i = 1 \right)$$

m_i η χρησιμότητα(βάρος) του κριτηρίου

και $u_i(g_i)$ η συνάρτηση μερικής χρησιμότητας του κριτηρίου g_i.

Οι μονότονες συναρτήσεις μερικών χρησιμοτήτων ικανοποιούν τις δύο βασικές συνθήκες:

$$\left.\begin{array}{l} u_i(g_{i*}) = 0 \\ u_i(g_i^*) = 1 \end{array}\right\}$$

όπου ως g_{i*} και g_i^* ορίζονται, αντίστοιχα, η λιγότερο και η περισσότερο προτιμητέα τιμή του κριτηρίου g_i.

Έχοντας ως A το σύνολο των k εναλλακτικών και ως $g_i(x_j)$ την επίδοση της εναλλακτικής x_j στο κριτήριο g_i, οι οποίες τιμές των g_{i*} και g_i* ορίζονται ως εξής:

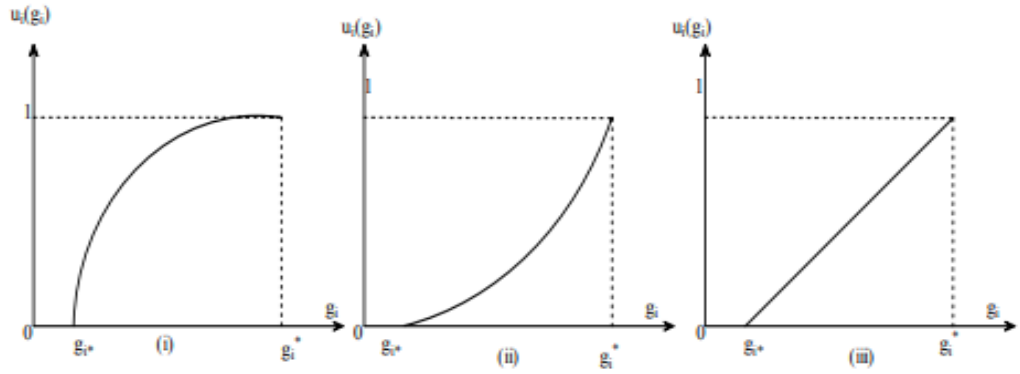
- Κριτήρια που οι **υψηλότερες** τιμές υποδεικνύουν καλύτερες εναλλακτικές:

$$g_{i*} = \min_{\forall x_j \in A}\{g_i(x_j)\} \text{ και } g_i^* = \max_{\forall x_j \in A}\{g_i(x_j)\}$$

- Κριτήρια που οι **χαμηλότερες** τιμές υποδεικνύουν καλύτερες εναλλακτικές:

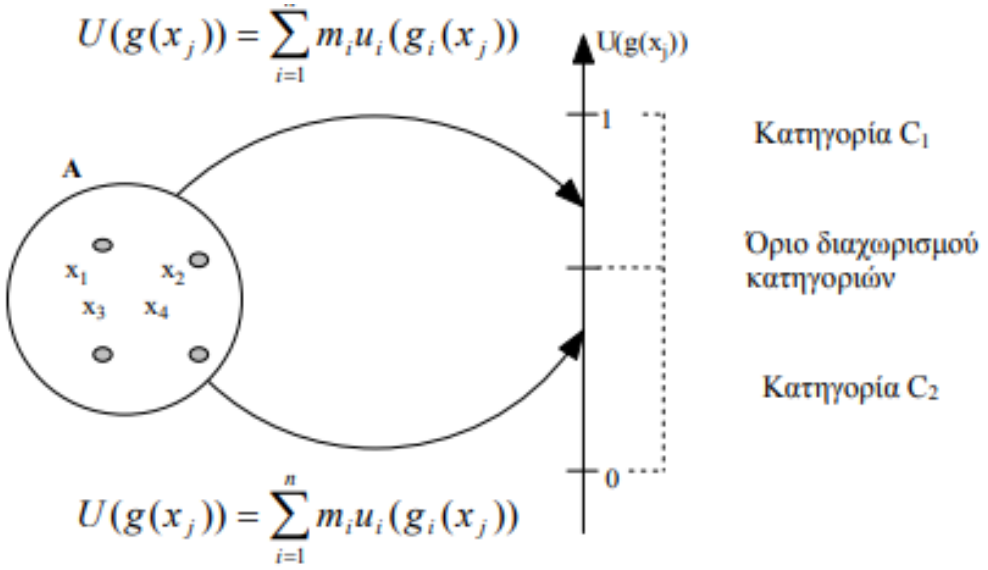
$$g_{i*} = \max_{\forall x_j \in A}\{g_i(x_j)\} \text{ και } g_i^* = \min_{\forall x_j \in A}\{g_i(x_j)\}$$

Οι συναρτήσεις μερικών χρησιμοτήτων πραγματοποιούν μια κανονικοποίηση για κάθε κριτήριο στο διάστημα [0, 1] η οποία αναπαριστά τη χρησιμότητα της κάθε τιμής του κριτηρίου, καθώς και καθορίζουν τον τρόπο με τον οποίο ο αποφασίζων καθορίζει τις εναλλακτικές στο κάθε κριτήριο.



Σχήμα 2.5.1: Διάφορες συναρτήσεις μερικής χρησιμότητας [12]

Χρησιμοποιώντας το γινόμενο από τα αντίστοιχα βάρη των κριτηρίων επί των μερικών χρησιμοτήτων, υπολογίζεται η ολική χρησιμότητα των εναλλακτικών που κυμαίνονται στο διάστημα [0, 1] καθώς αποτελούν τον συνολικό δείκτη αξιολόγησης και χρησιμοποιούνται για την ταξινόμηση των εναλλακτικών στις προκαθορισμένες κατηγορίες. Όπως παρουσιάζεται στο Σχήμα 2.5.2, για την περίπτωση των δύο κατηγοριών, η ταξινόμηση των εναλλακτικών εφαρμόζεται συγκρίνοντας τις ολικές τους χρησιμότητες με ένα όριο, εναλλακτικές με ολικές χρησιμότητες μεγαλύτερες του ορίου αυτού τοποθετούνται στην πρώτη προκαθορισμένη κατηγορία, και αντίστοιχα μικρότερες του ορίου στη δεύτερη κατηγορία.



Σχήμα 2.5.2: Ταξινόμηση των εναλλακτικών δραστηριοτήτων [12]

Με την ύπαρξη *q* κατηγοριών, η ταξινόμηση των εναλλακτικών γίνεται βάσει των κανόνων:

$$
\begin{aligned}
&U(g(x_j)) \geq u_1 && \Rightarrow x_j \in C_1 \\
&u_2 \leq U(g(x_j)) < u_1 && \Rightarrow x_j \in C_2 \\
&\ldots\ldots\ldots\ldots\ldots\ldots && \ldots\ldots\ldots\ldots\ldots \\
&u_k \leq U(g(x_j)) < u_{k-1} && \Rightarrow x_j \in C_k \\
&\ldots\ldots\ldots\ldots\ldots\ldots && \ldots\ldots\ldots\ldots\ldots \\
&U(g(x_j)) < u_{q-1} && \Rightarrow x_j \in C_q
\end{aligned}
$$

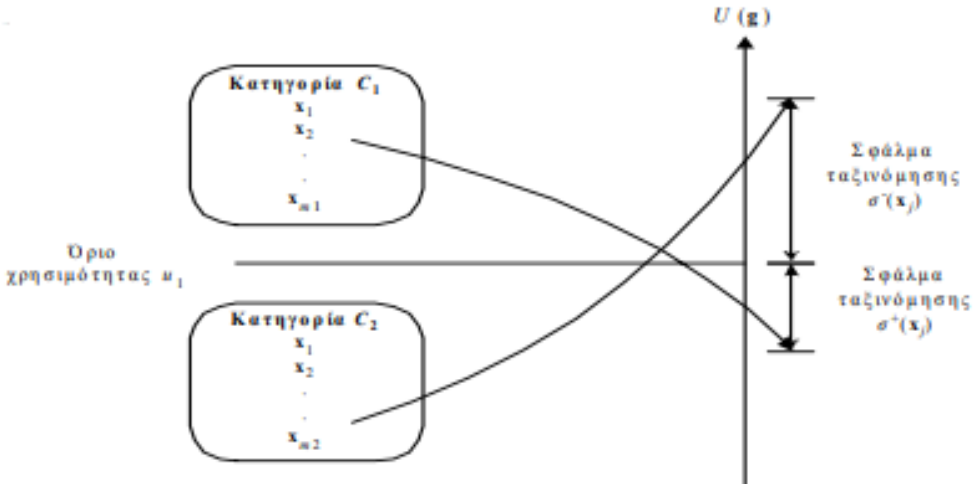
(2.5.1.2)

και $u_1, u_2, \ldots, u_{q-1}$ τα όρια χρησιμότητας τα οποία διαχωρίζουν τις υπάρχουσες κατηγορίες.

Τα δύο είδη σφαλμάτων που εμφανίζονται κατά την ταξινόμηση (Σχήμα 2.5.3):

α. Το σφάλμα υπερεκτίμησης $\sigma + (x_j)$ που σημαίνει ότι η εναλλακτική ταξινομήθηκε σε κατηγορία χαμηλότερη από ό,τι θα έπρεπε και έτσι για να διορθωθεί πρέπει να προστεθεί μία ποσότητα $\sigma(x)$ στην ολική χρησιμότητα της, ώστε να μπει στην κατηγορία που πρέπει.

β. Το σφάλμα υποεκτίμησης $\sigma - (x_j)$ που αντίστοιχα αντιπροσωπεύει ότι η εναλλακτική έχει μπει σε υψηλότερη κατηγορία από ό,τι θα έπρεπε και η ποσότητα $\sigma(x)$ πρέπει να αφαιρεθεί από την ολική χρησιμότητα της εναλλακτικής, ώστε να εισαχθεί στην σωστή κατηγορία



Σχήμα 2.5.3: Σφάλματα εκτίμησης [12]

Όπως φαίνεται δεν γίνεται να υπάρχουν παράλληλα και τα δύο σφάλματα καθώς πάντα το ένα από τα δύο είναι μηδέν, έτσι έχουμε μια αντιπροσωπευτική προσέγγιση του πραγματικού σφάλματος της ταξινόμησης.

Εν τέλει με την μοντελοποίηση αυτή, έχουμε μια ταξινόμηση που διαμορφώνεται με τη χρήση ενός προβλήματος (LP) και έχει στόχο την μείωση των σφαλμάτων που αναλύσαμε, με περιορισμούς την κανονικοποίηση του μοντέλου μεταξύ του 0 και 1, καθώς και τη διατήρηση της σωστής ταξινόμησης των εναλλακτικών.

2.6 Η μέθοδος TOPSIS

Η μέθοδος TOPSIS [13] χρησιμοποιεί την αρχή ότι η χρησιμότητα κάθε εναλλακτικής επιλογής πρέπει να έχει τη μικρότερη απόσταση από την Θετική Ιδεατή Λύση (Positive Ideal Solution - PIS) και τη μεγαλύτερη από την Αρνητική Ιδεατή Λύση (Negative Ideal Solution - NIS).

Υποθέτει την ύπαρξη ενός σύνολου *m* εναλλακτικών και *n* κριτηρίων, καθώς και ότι έχουμε την χρησιμότητα κάθε εναλλακτικής ως προς κάθε κριτήριο.

Ακολουθεί τα εξής:

- Έστω x_{ij} η χρησιμότητα της εναλλακτικής *i* ως προς το κριτήριο *j*. Δημιουργείται ένας πίνακας $X = (x_{ij})$ διαστάσεων *m* x *n*.
- Έστω *I* το σύνολο των μερικών χρησιμοτήτων των κριτηρίων (Κριτήρια που προτιμάμε τις υψηλές τιμές) .

- Έστω *J* το σύνολο αρνητικών κριτηρίων (Άντιστοιχα κριτήρια που προτιμάμε χαμηλές τιμές)

Έτσι η μέθοδος υλοποιείται από τα βήματα:

1) Υπολογισμός του κανονικοποιημένου πίνακα απόφασης με τιμή n_{ij} ς:

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \quad i = 1, \dots, m \quad , j = 1, \dots, n.$$

2) Κατασκευή του πίνακα των κανονικοποιημένων βαρών.

Έστω η ύπαρξη συνόλου των βαρών (χρησιμοτήτων) w_j για $j = 1, \dots, n$ και $\sum w_j = 1$ για $j = 1, \dots, n$, τότε η κανονικοποιημένη τιμή v_{ij} δημιουργείται:

$$v_{ij} = w_j n_{ij}, \qquad i = 1, \dots, m, \quad j = 1, \dots, n$$

3) Με θετική και αρνητική ιδεατή λύση:

$$A^+ = \{v_1^+, \dots, v_n^+\} = \{\left(\max_j v_{ij} \middle| i \in I\right), (\min_j v_{ij} | i \in J)\}$$

$$A^- = \{v_1^-, \dots, v_n^-\} = \{\left(\min_j v_{ij} \middle| i \in I\right), (\max_j v_{ij} | i \in J)\}$$

4) Με την χρήση της n – διάστατης Ευκλείδειας απόστασης αξιολογούμε κάθε εναλλακτική από την ιδεατή λύση ως εξής:

$$d_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^+)^2}\ , \qquad i = 1, \dots, m,$$

Αντίστοιχα, και την απόσταση από την αρνητική ιδεατή λύση ως:

$$d_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2}\ , \qquad i = 1, \dots, m,$$

5) Η σχετική απόσταση ως προς την ιδεατή λύση της εναλλακτικής A_i ως προς την A^+ προσδιορίζεται ώς:

$$R_i = \frac{d_i^-}{d_i^+ + d_i^-}, \quad i = 1, \dots, m.$$

Εφόσον $d_i -$ ≥ 0 και $d_i +$ ≥ 0, τοτε $R_i \in [0, 1]$.

6) Εν τέλει, κατατάσσουμε τις εναλλακτικές κατά φθίνουσα σειρά.

Κεφάλαιο 3 Μέθοδοι Μηχανικής Μάθησης

3.1 Εισαγωγή

Η μηχανική μάθηση προέρχεται από το συνδυασμό της επιστήμης υπολογιστών, της στατιστικής και της νευρο-επιστήμης. Ο εν λόγω τομέας προσπαθεί να κατασκευάσει συστήματα τα οποία μπορούν να εξελίσσονται αυτόματα με την εμπειρία που αποκτούν, καθώς και να καθορίσει τους θεμελιώδεις νόμους που διέπουν της διαδικασίας αυτής.

Υπάρχουν τρεις τρόποι μάθησης, ανάλογοι με τους τρόπους μάθησης του άνθρωπου:

- **Η Επιβλεπόμενη Μάθηση (Supervised Learning),** είναι η διαδικασία όπου ο αλγόριθμος διαθέτοντας τα σωστά αποτελέσματα σαν είσοδο μαθαίνει ένα γενικό κανόνα και δημιουργεί μια συνάρτηση προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα, και συνήθως βλέπουμε την χρήση της σε προβλήματα ταξινόμησης (Classification)

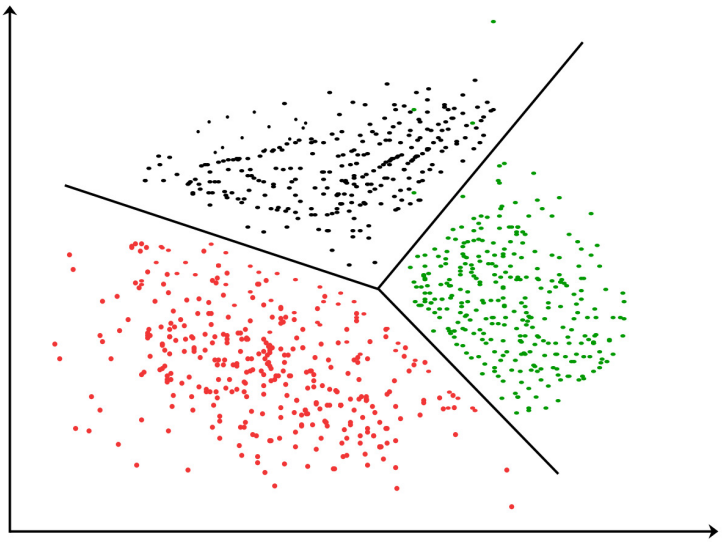
- **Η Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)**, χρησιμοποιείται συνήθως σε προβλήματα ομαδοποίησης (Clustering) και έχει στόχο να μοντελοποιησή σε μοτίβα την είσοδο, σε μορφή παρατηρήσεων χωρίς να έχει γνώση των επιθυμητών εξόδων.

- **Η Ενισχυτική Μάθηση (Reinforcement Learning),** μοντελοποιεί μια ακολουθία ενεργειών μέσα από αλληλεπίδραση με το περιβάλλον, και χρησιμοποιείται κυρίως σε προβλήματα Σχεδιασμού (Planning), όπως ο έλεγχος κίνησης ρομπότ και η αυτόματη οδήγηση ενός οχήματος.

3.2 Clustering - Συσταδοποίηση

Το πρόβλημα της συσταδοποίησης ή και ομαδοποίησης, το οποίο είναι επίμαχο εδώ, προσπαθεί να απευθυνθεί στην τμηματοποίηση (partitioning, clustering) ενός συνόλου δεδομένων σε όμοιες και σχετιζόμενες ομάδες (clusters).

Συνήθως υπάγεται στις μεθόδους μηχανικής μάθησης χωρίς επίβλεψη (Unsupervised Learning) και χρησιμοποιείται όπου υπάρχει η ανάγκη για ομαδοποίηση καθώς στόχος της, είναι ο διαχωρισμός των στοιχείων ενός συνόλου σε διαφορετικές συστάδες, όπως επίσης και η εύρεση των συστάδων που ανήκουν, έτσι ώστε τα στοιχεία σε μια συστάδα να σχετίζονται μεταξύ τους, όσο περισσότερο γίνεται, και αντίστοιχα να διαφοροποιούνται από άλλες συστάδες, όπως φαίνεται στο Σχήμα 3.2.1.



Σχήμα 3.2.1: Τμηματοποίηση μεταξύ στοιχείων-συστάδων

3.3 Αλγόριθμος k-means

Ο αλγόριθμος k-means[14] είναι η πιο γνωστή και απλή μέθοδος συσταδοποίησης και χρειάζεται ως είσοδο μόνο τον αριθμό των συστάδων (clusters) K .

Η μέθοδος ακολουθεί συνοπτικά τα εξής βήματα:

- Αρχικά γίνεται ο διαμερισμός των κέντρων των συστάδων σε K υποσύνολα.
- Ακολουθεί η ανάθεση των στοιχείων στα κοντινότερα κέντρα (centroids) συστάδων.
- Έπειτα γίνεται ξανά ο υπολογισμός του κέντρου της κάθε συστάδας.
- Επαναληπτικά, τα στοιχεία αλλάζουν θέση ανάμεσα στις συστάδες, έως ότου η ανάθεση ολοκληρωθεί.

Ο αλγόριθμος έχει στόχο να μείωση τη μέση τετραγωνική απόσταση των στοιχείων από τα κοντινότερα K κέντρα των συστάδων.

Αναλυτικότερα τα βήματα του αλγόριθμου k-means είναι:

Βήμα 1: Ανάθεση των αρχικών κέντρων v_i, i=1,2,…,c για τις c συστάδες. Για κάθε επανάληψη r=1,2,…,r_{max}.

Βήμα 2: Αξιολόγηση της απόστασης κάθε στοιχείου του συνόλου από το κέντρο κάθε συστάδας $d_{ki}=(x_k-v_i)$

Όπου k=1,2,…,n και i=1,2,…,c.

Βήμα 3: Αντιστοίχιση κάθε στοιχείου x_k στη συστάδα με την ελάχιστη απόσταση.

Βήμα 4: Υπολογισμός των νέων κέντρων των συστάδων

n_i ο αριθμός των στοιχείων που ανήκουν στην *i* συστάδα

$$m_i^{(r+1)} = \frac{\sum_{k=1}^{n_i} x_k}{n_i}$$

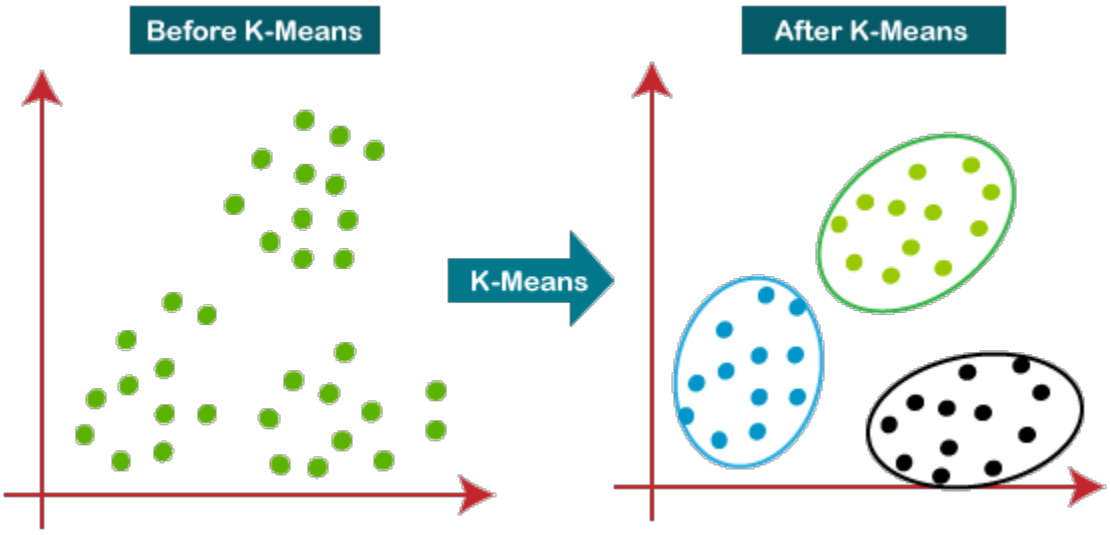
Βήμα 5: Εάν έχει γίνει ανάθεση όλων των στοιχείων, τότε γίνεται τερματισμός του αλγορίθμου (Σχήμα 3.3.1), διαφορετικά επανάληψη . Δηλαδή:

$$\text{if } \left\| m_i^{(r)} - m_i^{(r+1)} \right\| < \varepsilon \text{ then}$$

stop

else

$r = r+1$, goto 2



Σχήμα 3.3.1: Απεικόνιση του πριν και μετά τη εκτέλεση της μέθοδού k-means

Κεφάλαιο 4 Προτεινόμενη Μεθοδολογία

4.1 Εισαγωγή

Στο πιο σημαντικό κομμάτι της εργασίας θα αναλύσουμε την μεθοδολογία, στην οποία στηρίζεται το μοντέλο που προτείνουμε. Αρχικά να αναφέρουμε ότι χρησιμοποιούμε την Πολυκριτήρια Ανάλυση Αποφάσεων και πιο συγκεκριμένα τις μεθόδους UTASTAR, UTADIS και TOPSIS για την επίλυση των πολυκριτήριων προβλημάτων, όπως επίσης και η χρήση της μέθοδου συσταδοποίησης k-means για την επίλυση των προβλημάτων μηχανικής μάθησης. Ο συνδυασμός αυτών των μεθόδων, σκοπό έχει τη μείωση των διαστάσεων, την παρουσίαση συγκρίσεων μεταξύ των μεθόδων αυτών και την πρόταση ενός συνδυαστικού μοντέλου που θα προσφέρει την καλύτερη δυνατή ακρίβεια πρόβλεψης στο πρόβλημα των δεδομένων μας.

4.2 Υλοποίηση πολυκριτήριων μεθόδων και μεθόδων μηχανικής μάθησης

Οι τρεις πολυκριτήριες μεθοδολογίες του Κεφαλαίου 2 και η μέθοδος μηχανικής μάθησης υλοποιήθηκαν με την γλώσσα προγραμματισμού Python 3.7 [16] με χρήση του περιβάλλον Anaconda [17]. Η Python παρέχει την δυνατότητα ευέλικτου και κατανοητικού προγραμματισμού σε λίγες γραμμές κώδικα, καθώς είναι μια υψηλού επιπέδου γλώσσα προγραμματισμού. Έχει στην διάθεσή της πάρα πολλές βιβλιοθήκες για όλων των ειδών εφαρμογές, και κυρίως, τον τύπο δεδομένων Dataframe από την βιβλιοθήκη pandas [18], με τον οποίο δίνεται η δυνατότητα διαχείρισης μεγάλου όγκου δεδομένων σε μορφή πινάκων εύκολα και γρήγορα, όπως και την χρησιμοποιούμε στην παρούσα υλοποίηση, καθώς επίσης και τον επιλυτή lp_solve [19] που χρειάστηκε για την επίλυση των γραμμικών προβλημάτων.

Το σύστημα επιλύει προβλήματα με όλες τις μεθόδους και τα αποτελέσματα παρουσιάζονται σε γραφήματα και πίνακες. Για τη δημιουργία των γραφημάτων χρησιμοποιήθηκαν οι βιβλιοθήκες matplotlip [20].

Η υλοποίηση μεθόδων μηχανικής μάθησης και συγκεκριμένα της k-means έγινε με την χρήση της γνωστής βιβλιοθήκης Scikit-learn [21], η οποία περιέχει πρότυπα και διάφορα χρήσιμα εργαλεία για χρήση σε Supervised/Unsupervised learning algorithms, έτσι ώστε να προσφέρει ένα απλό και αποδοτικό τρόπο χρησιμοποίησης και λύσης πολύπλοκων προβλημάτων.

4.3 Η Μεθοδολογία κατασκευής του Μοντέλου

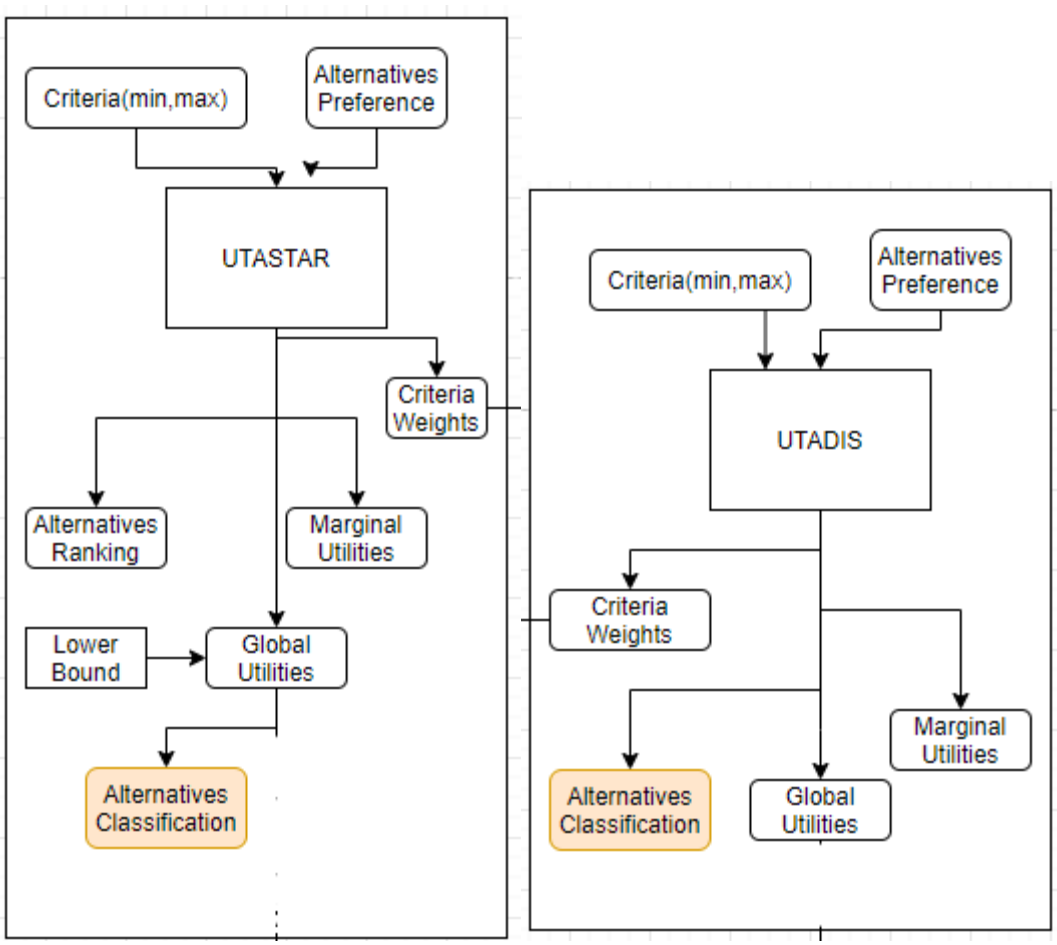
Στην προσπάθεια αύξησης της ακρίβειας ταξινόμησης στο κλάδο του Credit Risk Scoring γίνεται συνδυασμός μεθόδων από δύο τομείς: την **Θεωρία Αποφάσεων** που χρειάζεται και χρησιμοποιείται στην Διοίκηση και την **Στατιστική Μοντελοποίηση** που πλέον μπορούμε να αξιοποιήσουμε με την πληθώρα δεδομένων που είναι διαθέσιμα.

Όπως αναφέρθηκε στην Ενότητα 4.3, αρχικά υλοποιήσαμε τις τρεις πολυκριτήριες μεθόδους UTASTAR, UTADIS και TOPSIS στο περιβάλλον python, με σκοπό να μπορούμε να εξάγουμε βάρη για τα κριτήρια (διαστάσεις) των συνόλων δεδομένων και τις μερικές και ολικές χρησιμότητες για τις εναλλακτικές επιλογές.

Έπειτα ψάξαμε και βρήκαμε αξιόπιστα σύνολα δεδομένων διαφόρων τραπεζών που είχαν ως πρόβλημα την ταξινόμηση πελατών (εναλλακτικών επιλογών) σε αποδεκτούς και μη πελάτες για έγκριση δανείων, πιστωτικών καρτών, κλπ. Τα δεδομένα αυτά ήταν αναγκαίο να τα επεξεργαστούμε, έτσι ώστε να δημιουργήσουμε τους πολυκριτήριους πίνακες που χρειάζονται ως είσοδος στις μεθόδους, καθώς και να καθορίσουμε ή και να αναπροσαρμόσουμε την κατηγορία των κριτηρίων (ονομαστικές, κατηγορικές, κλπ.).

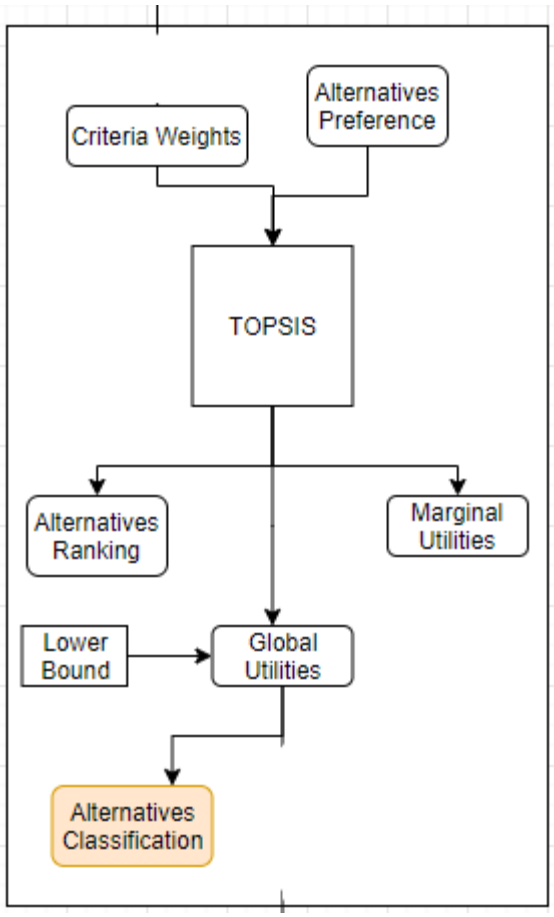
Συγκεκριμένα, χρειάστηκε ως είσοδο στις UTASTAR και UTADIS να καθορίσουμε προτίμηση στη σειρά κατάταξης εναλλακτικών επιλογών, καθώς και προτίμηση minimum ή maximum για το εύρος των τιμών κάθε κριτηρίου. Στην πρώτη περίπτωση, χρησιμοποιήσαμε ενδεικτικά την σωστή ταξινόμηση (class) κάθε εναλλακτικής επιλογής, και αντίστοιχα αποφασίσαμε και τους παράγοντες (min, max) κάθε κριτηρίου. Έτσι, τρέχοντας τις μεθόδους, έχουμε ως αποτέλεσμα τα βάρη των κριτηρίων (διαστάσεων), τις ολικές χρησιμότητες για κάθε εναλλακτική επιλογή και τις μερικές χρησιμότητες, δηλαδή την επιμέρους βαθμολογία της εναλλακτικής επιλογής σε κάθε κριτήριο. Αυτή η διαδικασία παρουσιάζεται στο Σχήμα 4.3.1.

Η UTASTAR έχει ως αποτέλεσμα την σειρά κατάταξης (ranking) των εναλλακτικών επιλογών και έτσι χρειάζεται να θέσουμε κατώτατο όριο για να ορίσουμε την ταξινόμηση. Αυτό το όριο το θέτουμε προσεγγιστικά, με την χρήση μιας απλής λογικής, κατά την οποία παίρνει την τιμή που βρίσκεται στην μέση του εύρους των ολικών χρησιμοτήτων (median) και εξετάζει ±10% των ορίων που βρίσκονται κοντά, έτσι ώστε να επιστρέψει το όριο με το οποίο προσφέρει την πιο μεγάλη ακρίβεια.



Σχήμα 4.3.1 Inputs and Outputs of UTASTAR, UTADIS

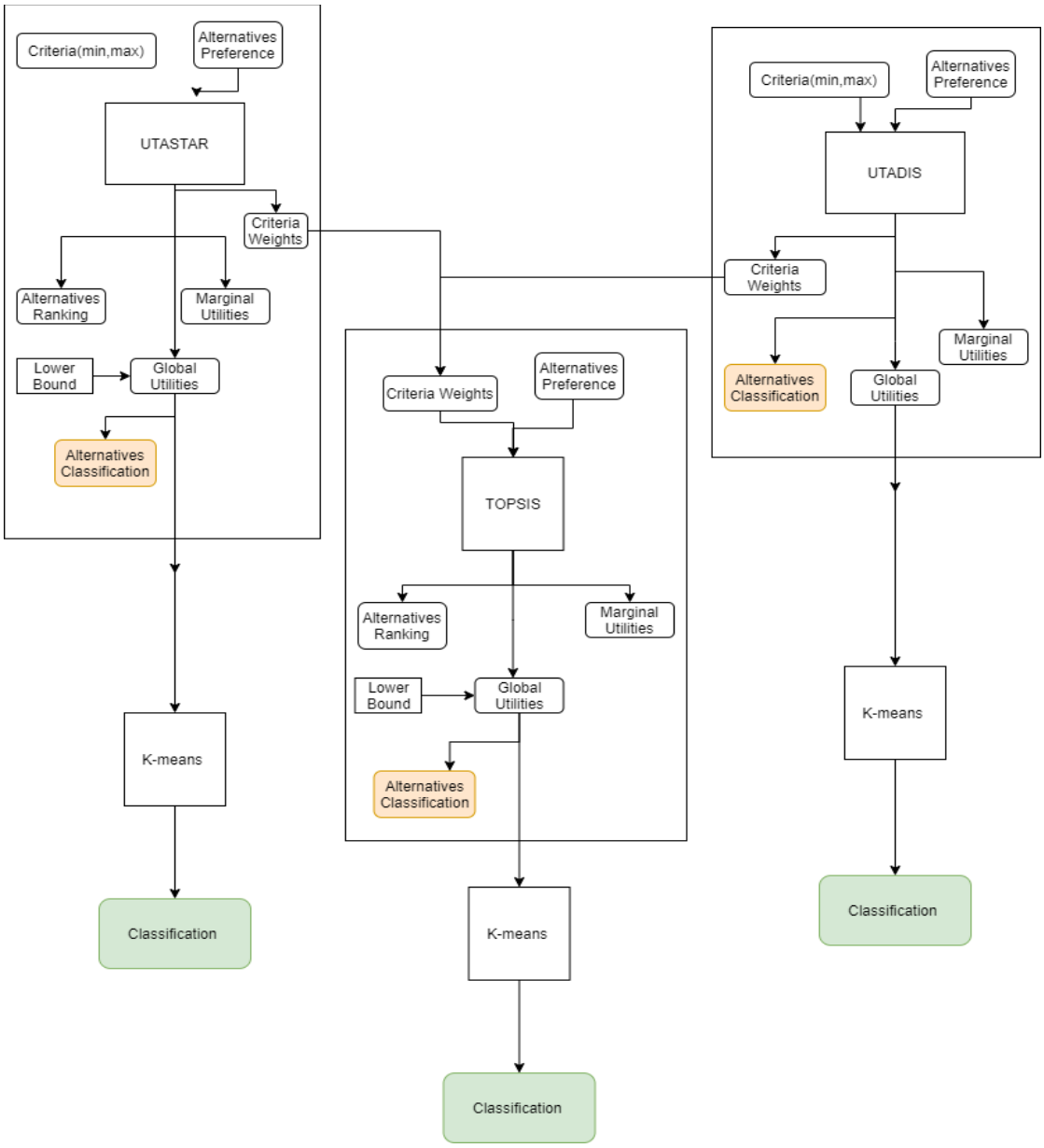
Αντίστοιχα, για να επαληθεύσουμε και να συγκρίνουμε αυτά τα αποτελέσματα, βλέπουμε την UTADIS στο Σχήμα 4.3.1, η οποία κάνει καθαρά ταξινόμηση και δημιουργεί τα όρια διαχωρισμού των εναλλακτικών επιλογών, μέσω του επιλυτή γραμμικού προγραμματισμού (LP), λαμβάνοντας υπόψη τα βάρη και τις χρησιμότητες που δημιουργεί.



Σχήμα 4.3.2 Inputs and Outputs of TOPSIS

Τέλος, η TOPSIS για την σειρά κατάταξης που επίσης παράγει, όχι με την χρήση LP, αλλά με τη χρήση Ιδεατών Λύσεων, χρειάζεται ως είσοδο βάρη για τα κριτήρια, όπως φαίνεται και στο Σχήμα 4.3.2, οπότε επιλέξαμε αντίστοιχα τα βάρη από UTASTAR και UTADIS και έπειτα καθορίσαμε το όριο της ταξινόμησης αντίστοιχα με πριν.

Έχοντας διάφορα αποτελέσματα από τις πολυκριτήριες μεθόδους όπως φαίνονται με πορτοκαλί στο Σχήμα 4.3.3 (UTASTAR, UTADIS, TOPSIS with UTASTAR Weights, TOPSIS with UTADIS Weights) εφαρμόζουμε k-means για ταξινόμηση σε συστάδες των δύο που φαίνεται με πράσινο στο Σχήμα 4.3.3 , θεωρώντας ότι μια συστάδα ισούται με μια κλάση και εξετάζουμε την βελτίωση στην ακρίβεια της πρόβλεψης που μπορούν να προσφέρουν τα αποτελέσματα που δίνονται από τις πολυκριτήριες μεθόδους.



Σχήμα 4.3.3 Complete model with all available results

Προχωράμε στην μείωση διαστάσεων (κριτηρίων) που μπορούμε να κάνουμε με τις πολυκριτήριες μεθόδους (UTASTAR, UTADIS). Αυτό επιτυγχάνεται εύκολα, καθώς τα βάρη των κριτηρίων αθροίζουν στο 1 (100%). Έτσι, μετά από καθοδήγηση ειδικών (experts), κρατήσαμε τα μεγαλύτερα κριτήρια που αθροίζουν στο 0.8 (80%).

Εν τέλει, με μειωμένες τις διαστάσεις, εφαρμόζουμε ξανά συσταδοποίηση k-means στα αποτελέσματα από τις πολυκριτήριες μεθόδους και εξετάζουμε και συγκρίνουμε την ακρίβεια αυτών των μοντέλων τα οποία και παρουσιάζουμε στη συνέχεια.

Κεφάλαιο 5 Σύγκριση και Παρουσίαση Αποτελεσμάτων

Στο κεφάλαιο αυτό παρουσιάζουμε τα πειραματικά αποτελέσματα των μεθόδων που υλοποιήσαμε. Χρησιμοποιήθηκαν τρία διαδεδομένα σύνολα δεδομένων από το αποθετήριο UCI [27] τα οποία περιέχουν δεδομένα για credit score από τρεις διαφορετικές τράπεζες και χώρες με διαφορετικά ποιοτικά χαρακτηριστικά, κριτήρια και μεγέθη, τα οποία αναλύουμε παρακάτω.

5.1 Πολυκριτήριοι Πίνακες Δεδομένων

UCI German Credit Score Dataset

Παρουσιάζουμε παρακάτω, στο Σχήμα 5.1.1, τον πολυκριτήριο πίνακα για το σύνολο δεδομένων UCI German Credit Score [23], όπου βλέπουμε τα στοιχεία από πελάτες μιας τράπεζας στην Γερμανία, που ήθελε να αξιολογήσει την διαδικασία έγκρισης δανείων ή πιστωτικών καρτών. Αναλύουμε παρακάτω τον τύπο, το εύρος και την εξήγηση του κάθε κριτηρίου που χρησιμοποιήθηκε.

Μέγεθος συνόλου δεδομένων: 480 γραμμές.

Criteria Name	Type	Range	Details
check_account	Ordinal	0 to 3	Status of existing checking account 1: ... < 0 DM (Deutsche Mark) 2: 0 <= ... < 200 DM 3: ... >= 200 DM /salary assignments for at least 1 year 0: no checking account
duration (months)	numerical	(4 - 72)	Duration in months
credit_history	Qualitative	(0 - 4)	Credit history

Purpose	Qualitative	1 - 11	1 = car (used) 2 = furniture/equipment 3 = radio/television 4 =domestic appliances 5 = repairs 6 = education 7 = vacation- 8 = retraining 9 = business 10 = others 11 = car (new)
Credit_amount	Numerical	409 - 14555	Credit amount
savings_account	Ordinal	0 - 4	Savings account/bonds 1: ... < 100 DM 2: 100 <= ... < 500 DM 3: 500 <= ... < 1000 DM 4: >= 1000 DM 0: unknown/ no savings account
employment	Ordinal	0 - 4	Present employment since 0: unemployed 1: ... < 1 year 2: 1 <= ... < 4 year 3: 4 <= ... < 7 years 4: >= 7 years
disposable income	Numerical	1 - 4	Installment rate in percentage of disposable income
Status_sex	qualitative	1 - 5	Personal status and sex 1: male: divorced/separated 2: female: divorced/separated/married 3: male: single 4: male: married/widowed 5: female: single
Debtos_guarantors	Qualitative	0 - 2	Other debtors / guarantors 0: none 1: co-applicant 2: guarantor
Residence_since	Numerical	1 - 4	Present residence since

Property	qualitative	0 - 3	Property 1: real estate 2: otherwise, building society savings agreement / life insurance 3: otherwise, car or other, not in attribute "Savings account/bonds" 0: unknown / no property
Age	numerical	19 - 75	Age in years
Installment_plans	qualitative	0 - 2	Other installment plans 1: bank 2: stores 0: none
Housing	qualitative	1 - 3	Housing 1: rent 2: own 3: for free
Num_credits_bank	ordinal	1 - 4	Number of existing credits at this bank
Job	numerical	0 - 3	Job 0: unemployed/ unskilled - non-resident 1: unskilled - resident 2: skilled employee / official 3: management / self-employed / highly-qualified employee / officer
Providor_num	Numerical	1 - 2	Number of people being liable to provide maintenance for
Telephone	Qualitative	0 - 1	Telephone 0: none 1: yes, registered under the customer's name
Foreign	qualitative	1 - 2	foreign worker 1: yes 2: no
Class		1, 2	Approved or not

Σχήμα 5.1.1: Πολυκριτήριος Πίνακας UCI German Credit Score Dataset

UCI Taiwan Credit Score Dataset

Αντίστοιχα, στο Σχήμα 5.1.2, βλέπουμε το σύνολο δεδομένων για το UCI Taiwan Credit Score Dataset [24] με τις αντίστοιχες αναλύσεις, όπου η συγκεκριμένη τράπεζα στην Ταϊλάνδη ήθελε να αξιολογήσει την ικανότητα πληρωμής της επόμενης δόσης ή την ικανότητα έναρξης ενός νέου δανείου.

Μέγεθος συνόλου δεδομένων: 30000 γραμμές.

Column Name	**Data Attributes**	**Range**	**Details**
Limit_Bal	Ordinal	10k - 1mil	Amount of the given credit (NT dollar): It includes both the individual consumer credit and his/her family (supplementary) credit.
Sex	Nominal	(1, 2)	1 = male 2 = female
Education	Nominal	(0 to 4)	0 = no education 1 = graduate school 2 = university 3 = high school 4 = others
Marriage	Nominal	(1, 2)	1 = married 2 = single
Age	Ordinal	(21 to 79)	Age in years
Pay_0	Nominal	(0 to 8)	History of past payment We tracked the past monthly payment records (from April to September 2005) as follows: 0 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; 3 = payment delay for three months; … 6 = payment delay for six months;
Pay_2			
Pay_3			
Pay_4			
Pay_5			
Pay_6			
Bill_AMT1	Ordinal	(-300k to 700k)	Amount of bill statement (NT dollar) X12 = amount of bill statement in September 2005;
Bill_AMT2			

Bill_AMT3			… X16 = amount of bill statement in November 2005; X17 = amount of bill statement in December 2005.
Bill_AMT4			
Bill_AMT5			
Bill_AMT6			
Pay_AMT1	Ordinal	(0 to 2mil)	Amount of previous payment (NT dollar) X18 = amount paid in September 2005; …. X22 = amount paid in November 2005; X23 = amount paid in December 2005.
Pay_AMT1			
Pay_AMT1			
Pay_AMT1			
Pay_AMT1			
Pay_AMT1			
Y		Class (0-1)	

Σχήμα 5.1.2 : Πολυκριτήριος πίνακας UCI Taiwan Credit Score Dataset

UCI Australian Credit Score Dataset

Παρακάτω, βλέπουμε και το τελευταίο σύνολο δεδομένων που χρησιμοποιήθηκε, το UCI Australian Credit Score [25], που είχε σκοπό την αξιολόγηση της αποδοχής για πίστωση σε ένα πελάτη, είτε για κάρτα, είτε για Δάνειο. Έτσι, παρατηρούμε την σχετική ανάλυση στο Σχήμα 5.1.3 με την ιδιαιτερότητα ότι όλες οι ονομασίες κριτηρίων έχουν αλλάξει σε ουδέτερα σύμβολα για λόγους ασφάλειας προσωπικών δεδομένων.

Μέγεθος συνόλου δεδομένων: 690 γραμμές.

Column Name	Data Attributes	Data Qualities	Range
A1	Nominal	Categorical	(0-1)
A2	Ordinal	Continuous	(16-80)
A3	Ordinal	Continuous	(0-9)
A4	Nominal	Categorical	(1-3)

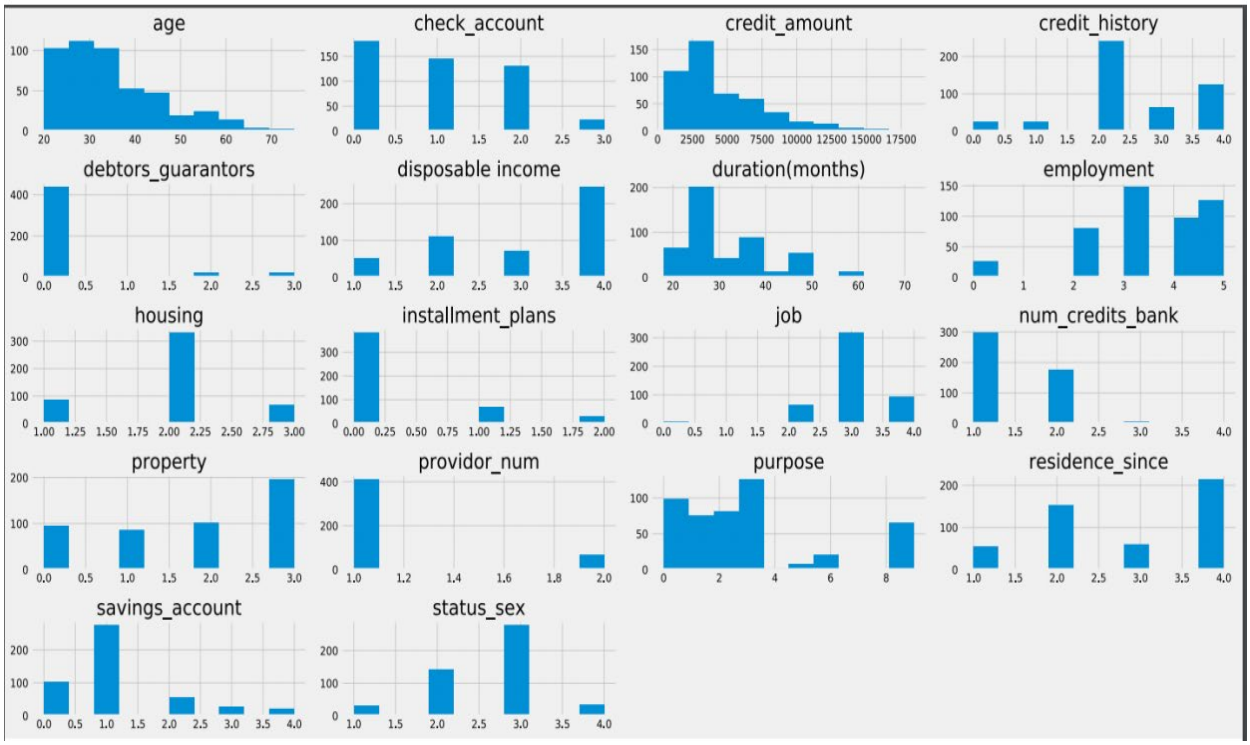
A5	Nominal	Categorical	(1-14)
A6	Nominal	Categorical	(1-9)
A7	Ordinal	Continuous	(0-9)
A8	Nominal	Categorical	(0-1)
A9	Ordinal	Categorical	(1-0)
A10	Ordinal	Continuous	(0-67)
A11	Nominal	Categorical	(1-0)
A12	Nominal	Categorical	(1-3)
A13	Ordinal	Continuous	(0-2000)
A14	Ordinal	Continuous	(1-1000001)

Σχήμα 5.1.2: Πολυκριτήριος πίνακας UCI Australian Credit Score Dataset

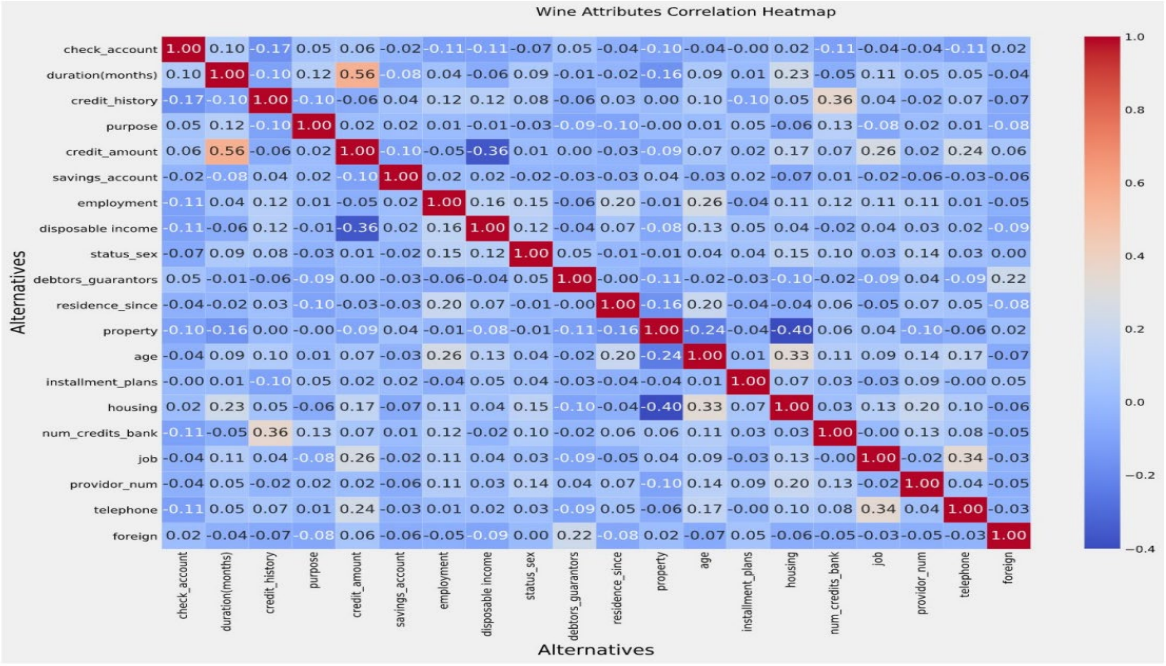
5.2 Αρχική Ανάλυση Διαστάσεων

Για να μπορέσουμε να χρησιμοποιήσουμε τα σύνολα δεδομένων, απαιτείται η κατάλληλη προ-επεξεργασία. Στην προσπάθεια αυτή, απεικονίζουμε όλα τα κριτήρια, όπως φαίνονται στα Σχήματα 5.2.2, 5.2.4 και 5.2.6, για να βρούμε συσχετίσεις μεταξύ τους, που ίσως έχουν μεγάλη επιρροή στο αποτέλεσμα, καθώς και διορθώσεις στις μεταβλητές (πραγματικά 0 σε ονομαστικές μεταβλητές, κλπ.) που χρειάστηκαν οι πολυκριτήριες μέθοδοι και διακρίνονται εύκολα με την χρήση ιστογραμμάτων, τα οποία φαίνονται στα Σχήματα 5.2.1, 5.2.3 και 5.2.5.

UCI German Credit Score



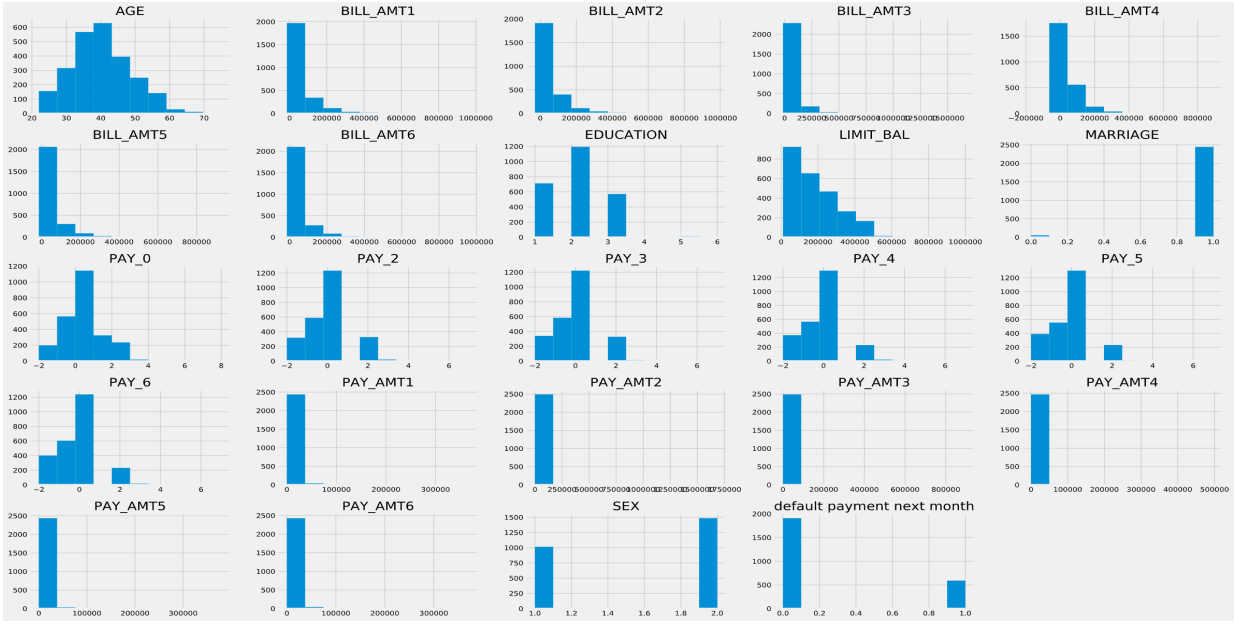
Σχήμα 5.2.1: UCI German Credit Score Criteria Histogram Distribution



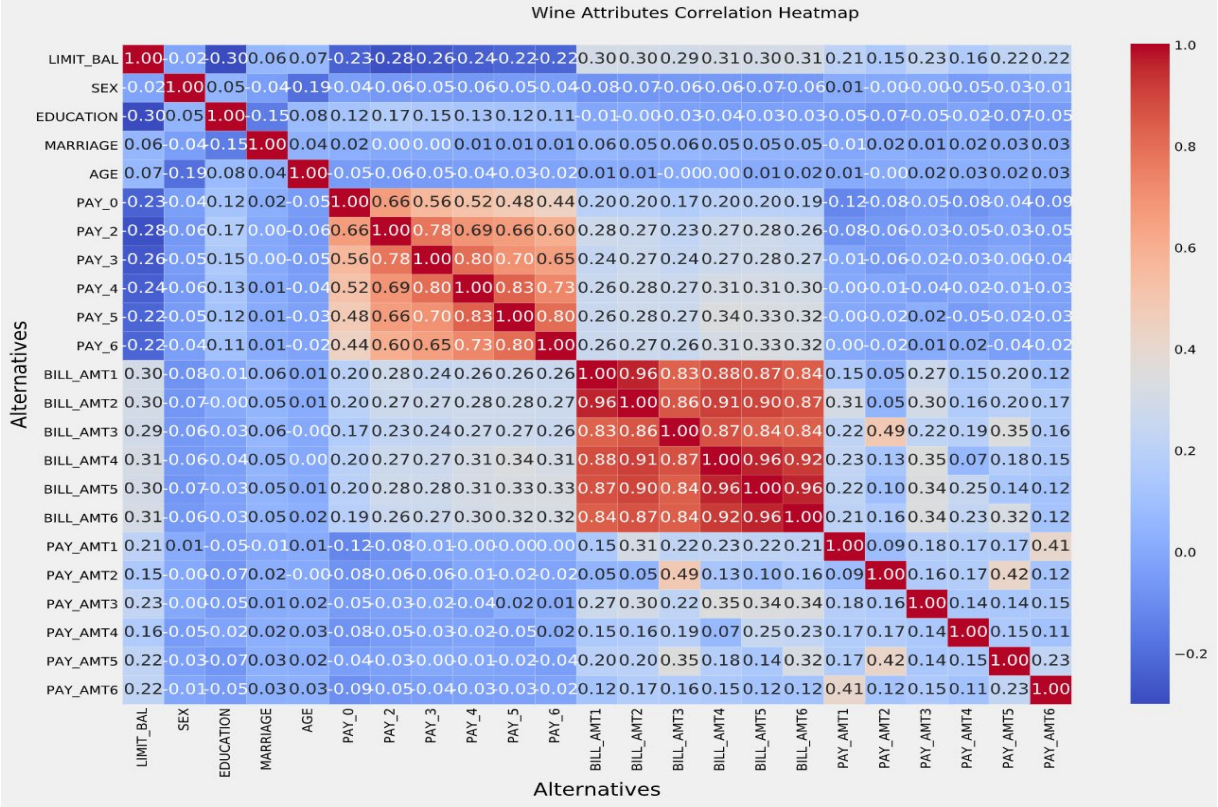
Σχήμα 5.2.2 : UCI German Credit Score Heatmap

Παρατηρούμε μια σχετικά ομοιόμορφη κατανομή των τιμών σε όλες τις διαστάσεις, χωρίς να υπάρχουν ισχυρές συσχετίσεις μεταξύ των διαστάσεων.

UCI Taiwan Dataset



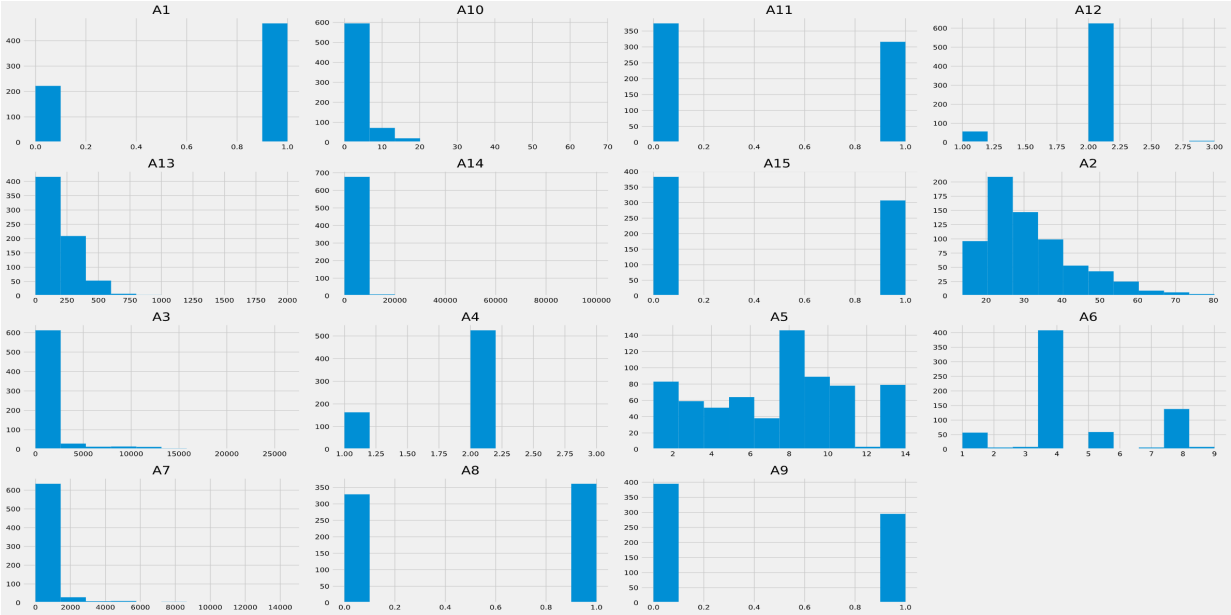
Σχήμα 5.2.3. UCI Taiwan Credit Score Criteria Histogram Distribution



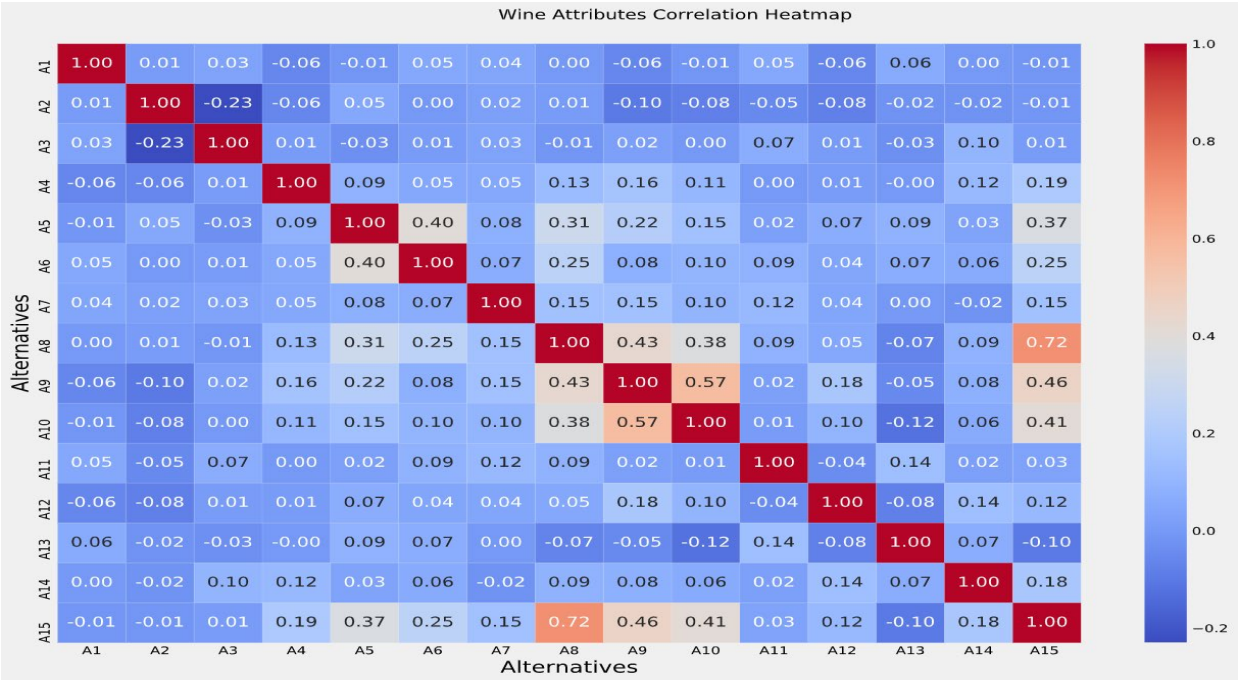
Σχήμα 5.2.4 UCI Taiwan Credit Score Criteria Heatmap

Παρατηρήσαμε σε αυτό το σύνολο δεδομένων μια έντονη ανισορροπία στα κριτήρια που περιέχουν ποσοτικές τιμές, όπως επίσης και μια μικρή διασπορά στα υπόλοιπα. Θα εξισορροπήσουμε την διασπορά με μερικές αναδιατάξεις των τιμών και, εφόσον χρειαστεί, μετά τις δοκιμές των μεθόδων που αναφέρονται, ίσως να ταξινομήσουμε τις ποσοτικές τιμές προς τον σκοπό καλύτερης διαχείρισης.

UCI Australian Dataset



Σχήμα 5.2.5 : UCI Australia Credit Score Histogram Distribution



Σχήμα 5.2.6: UCI Australian Credit Score Criteria Heatmap

Στο τελευταίο σύνολο δεδομένων που χρησιμοποιήσαμε παρατηρούμε και πάλι μικρή διασπορά των κατηγορικών μεταβλητών και μεγάλη ανισορροπία στις ποσοτικές που θα διορθώσουμε, εάν χρειαστεί στην πορεία.

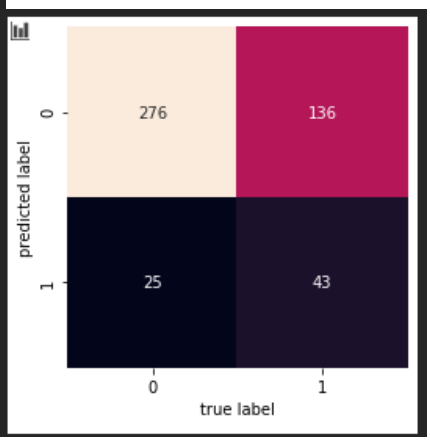
5.3 Πολυκριτήριες Μέθοδοι και Αποτελέσματα στα Αρχικά Δεδομένα

5.3.1 UTASTAR Weights and Classification Accuracy

Παρατηρούμε στα Σχήματα 5.3.1 έως και 5.3.6, τα βάρη που δημιούργησε η UTASTAR και η UTADIS αντίστοιχα για τα κριτήρια για κάθε σύνολο δεδομένων, καθώς και την μεγαλύτερη και μικρότερη ολική χρησιμότητα που είχαμε στις εναλλακτικές επιλογές. Επίσης, βλέπουμε και την ολική χρησιμότητα που τέθηκε ως κατώτατο όριο για να ταξινομήσει τις εναλλακτικές επιλογές. Τέλος, στα δεξιά φαίνεται η ακρίβεια της συγκεκριμένης ταξινόμησης αντίστοιχα, η οποία είναι αρκετά χαμηλή λόγω των πολλών διαστάσεων.

UCI German Credit Score

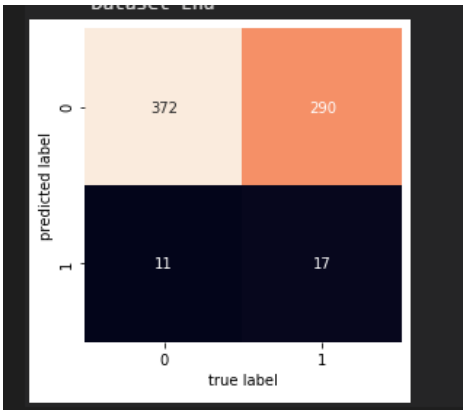
```
---Utastar---
Utastar Max Overall Value --> 0.804999816089358
Utastar Min Overall Value --> 0.4848492825189469
Utastar Classification Value Bound--> 0.6664757786062049
Utastar Accuracy--> 0.6645833333333333
                    ValueFunc
check_account            0.91
duration(months)        27.03
credit_history           8.30
purpose                  0.08
credit_amount           11.81
savings_account          1.93
employment               1.66
disposable income        0.01
status_sex               6.88
debtors_guarantors       1.39
residence_since          0.90
property                 6.95
age                      1.96
installment_plans        0.62
housing                  7.73
num_credits_bank         1.02
job                      9.75
providor_num             0.38
telephone                0.21
foreign                 10.49
```



Σχήμα 5.3.1 UTASTAR βάρη κριτηρίων, χαμηλότερο όριο και ακρίβεια ταξινόμησης (66%).

UCI Australian Credit Score

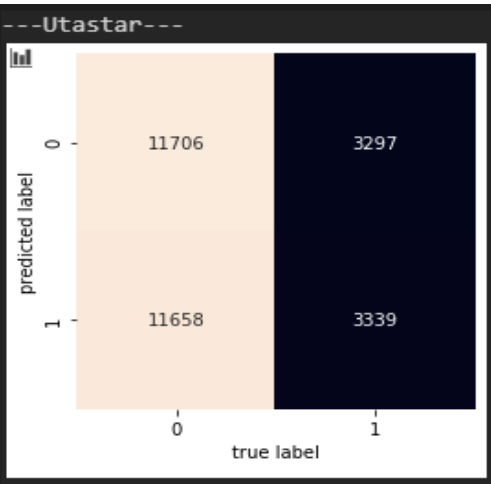
```
---Utastar---
Utastar Max Overall Value --> 0.7999277048160685
Utastar Min Overall Value --> 0.30055325243215014
Utastar Classification Value Bound--> 0.31378615886152084
Utastar Accuracy--> 0.563768115942029
     ValueFunc
A1       10.14
A2        7.91
A3        8.49
A4        7.54
A5        4.14
A6       10.77
A7        4.63
A8        3.12
A9       10.47
A10      11.30
A11       6.40
A12       6.59
A13       2.61
A14       5.89
```



Σχήμα 5.3.2 UTASTAR βάρη κριτηρίων, χαμηλότερο όριο και ακρίβεια ταξινόμησης (56%).

UCI Taiwan dataset

```
              ValueFunc
LIMIT_BAL          5.83
SEX                8.22
EDUCATION          8.10
MARRIAGE           3.03
AGE                2.48
PAY_0              2.67
PAY_2              6.95
PAY_3              4.13
PAY_4              6.42
PAY_5              7.53
PAY_6              8.04
BILL_AMT1          6.18
BILL_AMT2          4.35
BILL_AMT3          0.83
BILL_AMT4          3.56
BILL_AMT5          2.56
BILL_AMT6          0.51
PAY_AMT1           2.91
PAY_AMT2           6.00
PAY_AMT3           2.39
PAY_AMT4           1.14
PAY_AMT5           2.95
PAY_AMT6           3.22
```



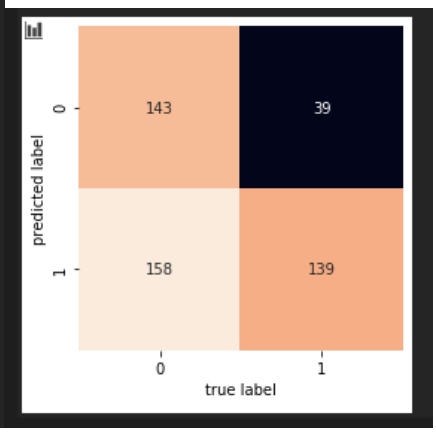

```
---Utastar---
Utastar Max Overall Value --> 0.7999788423271226
Utastar Min Overall Value --> 0.3000136845982127
Utastar Classification Value Bound--> [0.55009256]
Utastar Accuracy--> 0.5015
```

Σχήμα 5.3.3 UTASTAR βάρη κριτηρίων, χαμηλότερο όριο και ακρίβεια ταξινόμησης (50%).

5.3.2 UTADIS Weights and Classification Accuracy

UCI German Credit score dataset

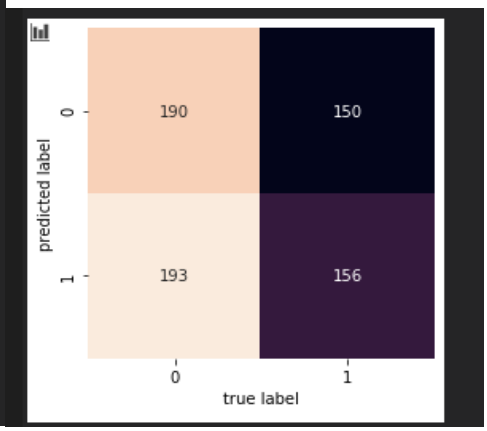
```
---Utadis---
Utadis Max Overall Value --> 0.7616962493682139
Utadis Min Overall Value --> 0.37536596915451836
Utadis Classification Value Bound--> [0.60172544]
Utadis Accuracy--> 0.5887265135699373
                     ValueFunc
check_account             1.67
duration(months)         24.02
credit_history            9.22
purpose                   0.21
credit_amount             6.06
savings_account           9.02
employment                5.73
disposable income         0.11
status_sex                6.54
debtors_guarantors        3.28
residence_since           3.68
property                  6.61
age                       2.73
installment_plans         3.19
housing                  10.45
num_credits_bank          0.11
job                       0.15
providor_num              1.52
telephone                 1.79
foreign                   3.90
```



Σχήμα 5.3.4 UTADIS βάρη κριτηρίων, χαμηλότερο όριο και ακρίβεια ταξινόμησης (58%)

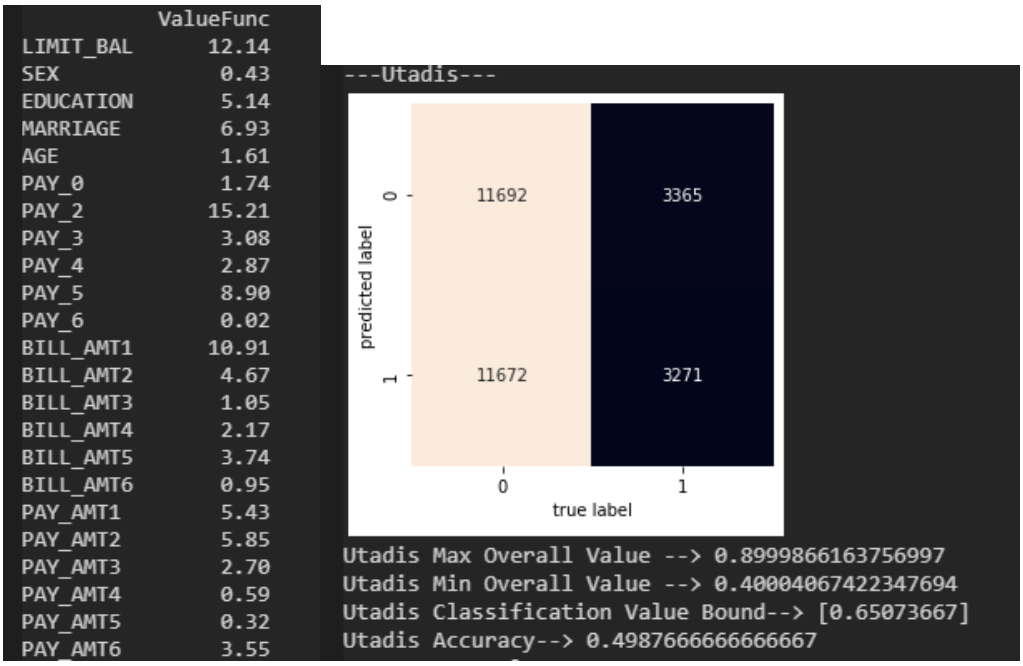
UCI Australian dataset

```
---Utadis---
Utadis Max Overall Value --> 0.8994730928174337
Utadis Min Overall Value --> 0.40194257843391007
Utadis Classification Value Bound--> [0.64511231]
Utadis Accuracy--> 0.502177068214804
     ValueFunc
A1        7.30
A2       13.28
A3        1.73
A4       12.78
A5        3.38
A6       13.27
A7        8.92
A8        1.02
A9        0.72
A10      12.77
A11       8.90
A12       9.53
A13       5.25
A14       1.17
```



Σχήμα 5.3.5 UTADIS βάρη κριτηρίων, χαμηλότερο όριο και ακρίβεια ταξινόμησης (50%).

UCI Taiwan dataset



Σχήμα 5.3.6 UTADIS βάρη κριτηρίων, χαμηλότερο όριο και ακρίβεια ταξινόμησης (49%).

Παρατηρήσεις:

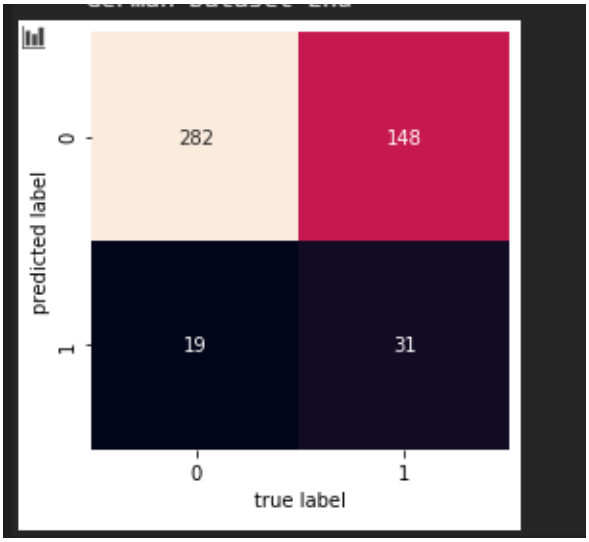
- Από τα αποτελέσματα των UTADIS και UTASTAR παρατηρούμε μια ακρίβεια στην ταξινόμηση γύρω στο 50-60% που θεωρείται τυχαία και ανακριβής. Είναι επίσης κρίσιμο να αναφερθεί ότι, λόγω και της φύσης του προβλήματος, δηλαδή λόγω του ότι προσπαθούμε να κάνουμε μια δυαδική ταξινόμηση, είναι ακόμη πιο δύσκολο να συμπεράνουμε αν η ακρίβεια είναι τυχαία ή όχι. Αυτό προσπαθήσαμε να το λύσουμε χρησιμοποιώντας διάφορα σύνολα δεδομένων για να συγκρίνουμε.

5.3.3 TOPSIS with UTASTAR weights and classification

UCI German dataset

Χρησιμοποιώντας λοιπόν τα παραγόμενα βάρη, δοκιμάζουμε να τα αξιολογήσουμε με την μέθοδο TOPSIS και την ταξινόμηση που μπορούμε να προσφέρουμε με την χρήση της, όπως βλέπουμε στα Σχήματα 5.3.7 και 5.3.12.

```
--Topsis with Utastar Weights--
Topsis Max Overall Value --> 0.7670902092105188
Topsis Min Overall Value --> 0.30153501361565976
Topsis Classification Value Bound--> 0.47854790117274865
Topsis Accuracy--> 0.6520833333333333
```

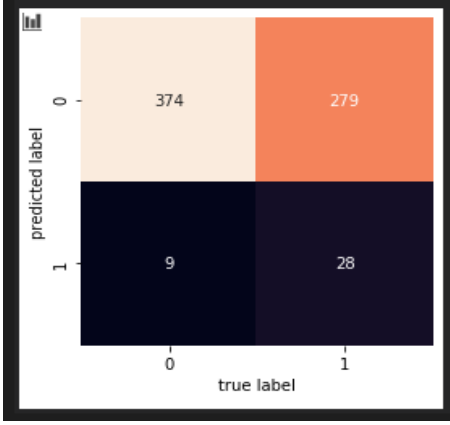


Σχήμα 5.3.7 Ακρίβεια TOPSIS με βάρη από UTASTAR (65%)

UCI Australian dataset

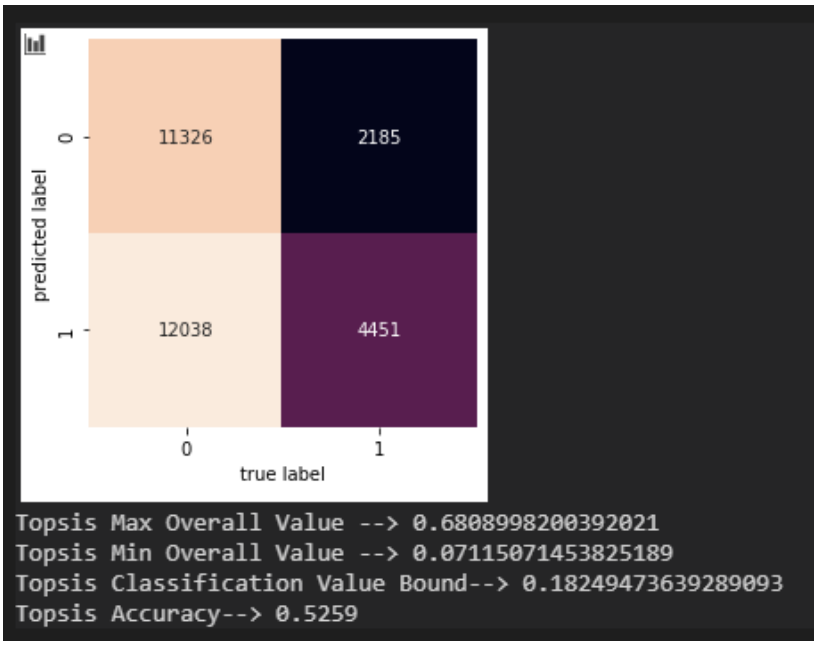
```
--Topsis with Utastar Weights--
Topsis Max Overall Value --> 0.9375502510493013
Topsis Min Overall Value --> 0.492976073737483
Topsis Classification Value Bound--> 0.82539071147063
Topsis Accuracy--> 0.5826086956521739

----- Dataset End----------
```



Σχήμα 5.3.8 Ακρίβεια TOPSIS με βάρη από UTASTAR (58%)

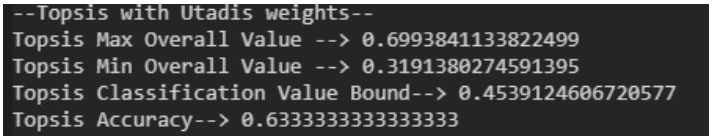
UCI Taiwan dataset

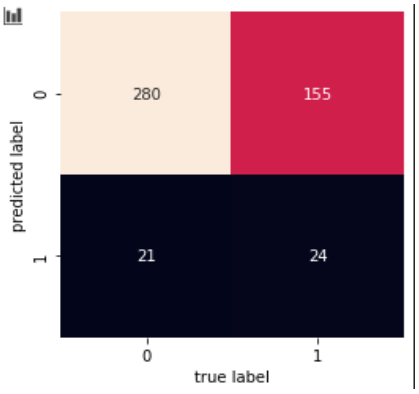


Σχήμα 5.3.9 Ακρίβεια TOPSIS με βάρη από UTASTAR (52%)

5.3.4 TOPSIS with UTADIS weights and classification

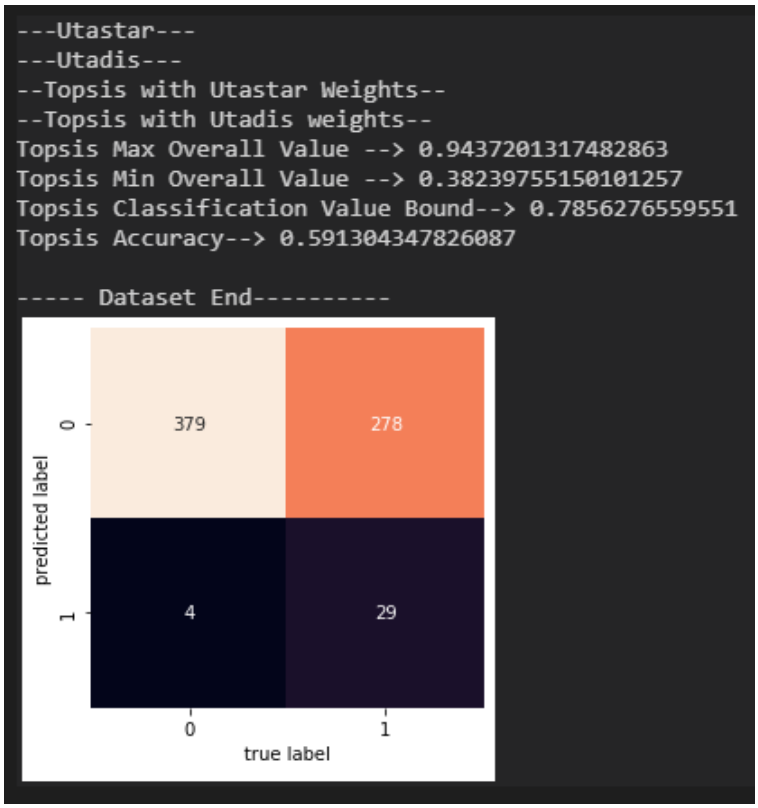
UCI German dataset





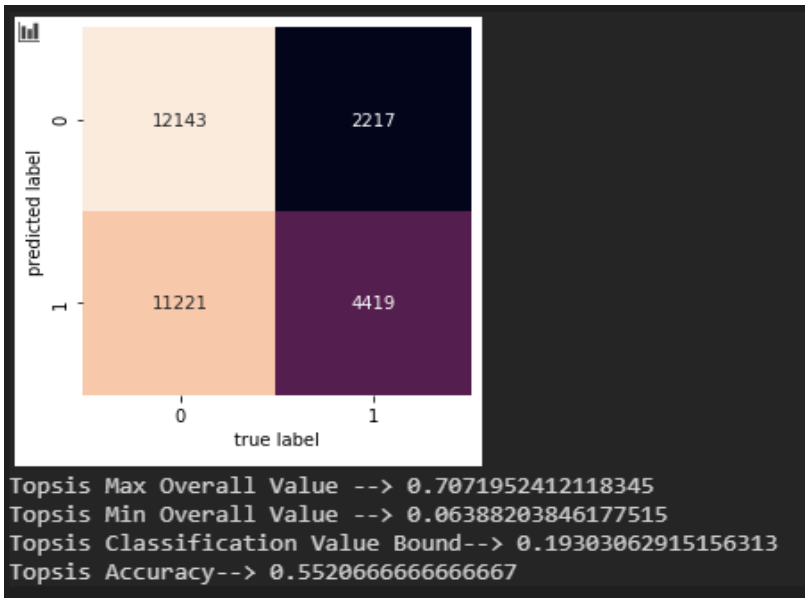
Σχήμα 5.3.10 Ακρίβεια TOPSIS με βάρη από UTADIS (63%).

UCI Australian dataset



Σχήμα 5.3.11 Ακρίβεια TOPSIS με βάρη από UTADIS (59%).

UCI Taiwan dataset



Σχήμα 5.3.12 Ακρίβεια TOPSIS με βάρη από UTADIS (55%).

Παρατηρήσεις:

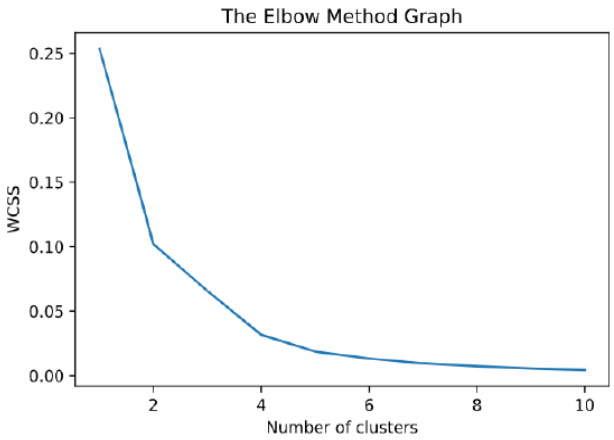
- Προσπαθήσαμε με την TOPSIS να συγκρίνουμε και την απόδοση που έχουν για το πρόβλημά μας οι πολυκριτήριες μέθοδοι, καθώς είναι βασισμένες στην Αναλυτική Συνθετική Προσέγγιση, ενώ η TOPSIS είναι βασισμένη στις Ιδεατές Λύσεις και βλέπουμε ότι έχει παρόμοια αποτελέσματα.

5.4 Μέθοδοι Μηχανικής Μάθησης και Αποτελέσματα

Σε αυτό το σημείο, θα εξεταστεί η απόδοση που έχει μόνη της η μέθοδος k-means στα αρχικά δεδομένα, για να χρησιμοποιηθεί ως σημείο σύγκρισης.

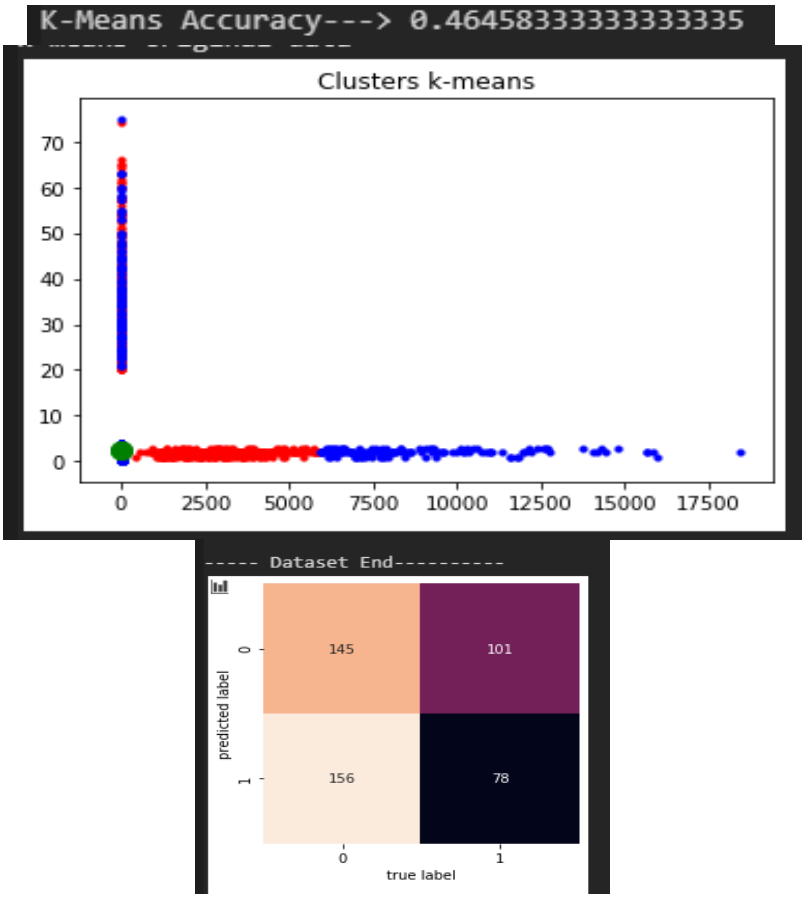
5.4.1 Ταξινόμηση k-means

UCI German Credit Score



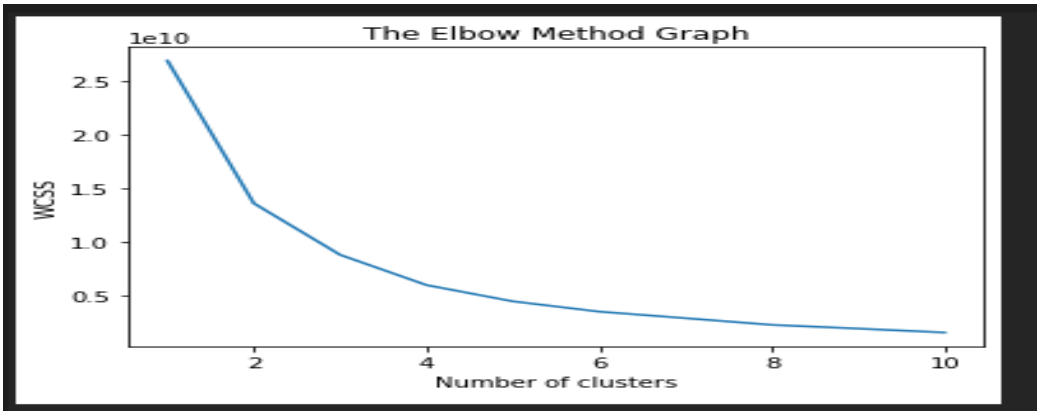
Σχήμα 5.4.1 Ανάλυση αριθμών cluster.

Στο Σχήμα 5.4.1 βλέπουμε την γνωστή ανάλυση του Elbow Method που προσπαθεί να εκτιμήσει τον βέλτιστο αριθμό συστάδων για το σύνολο δεδομένων, απλά για να επιβεβαιώσουμε ότι κινούμαστε προς την επιθυμητή ταξινόμηση. Στο Σχήμα 5.4.2 φαίνονται τα αποτελέσματα της ταξινόμησης με k-means. Αντίστοιχα σχήματα ακολουθούν και για τα άλλα datasets.

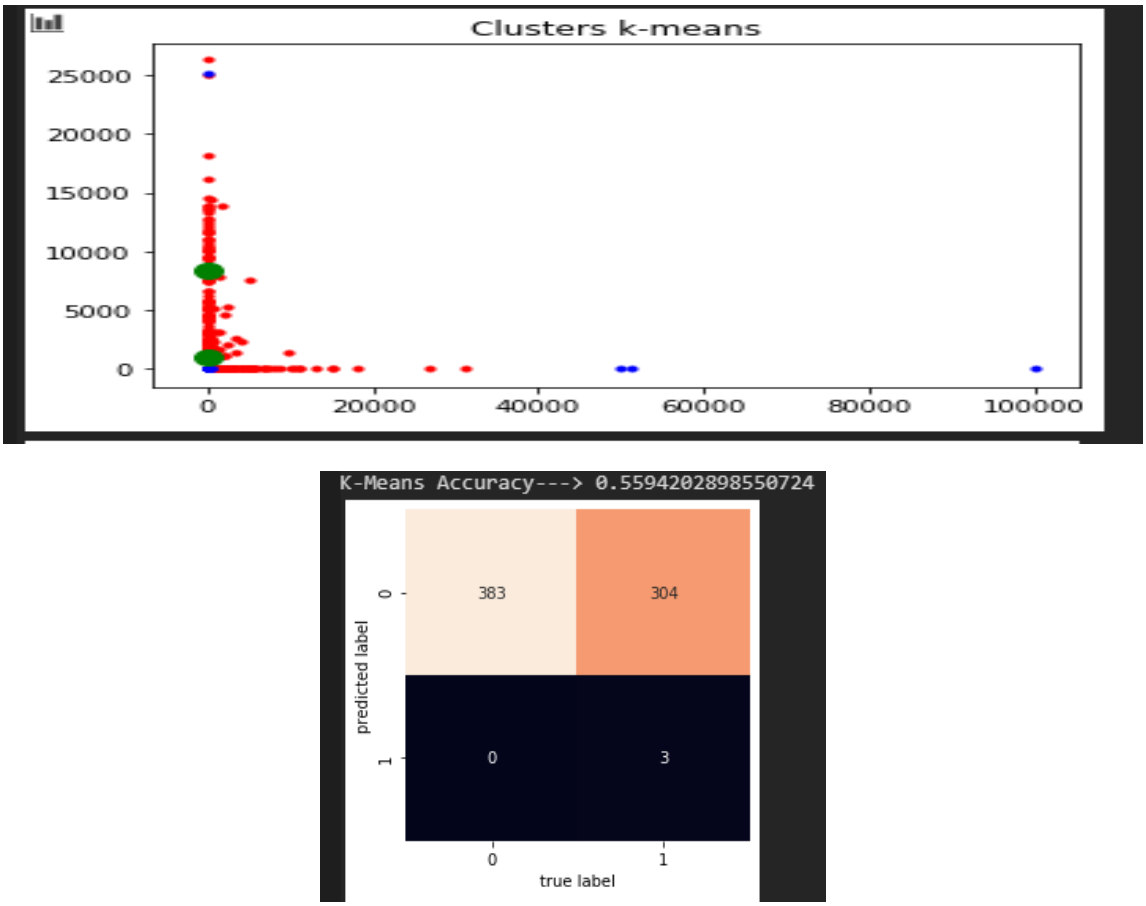


Σχήμα 5.4.2 Ακρίβεια clustering με k-means (46%).

UCI Australian Dataset

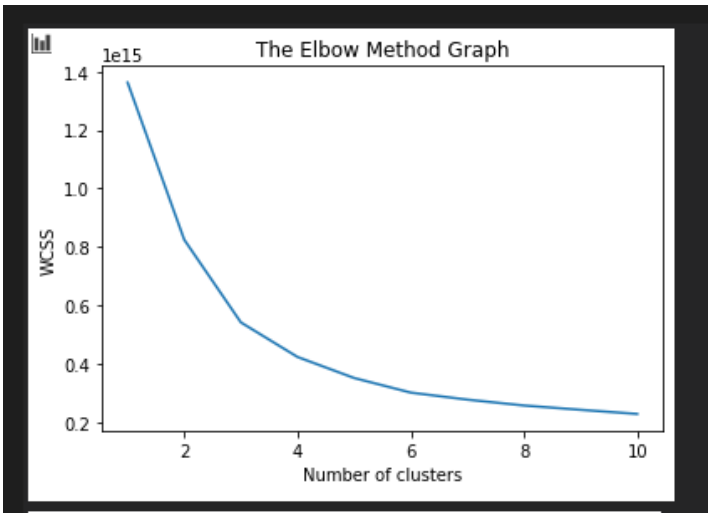


Σχήμα 5.4.3 Ανάλυση αριθμού cluster.

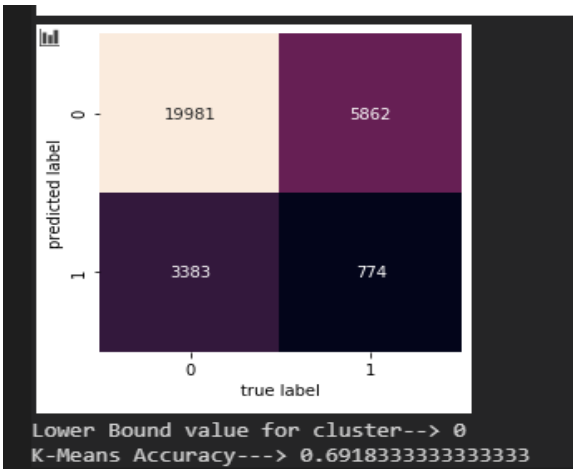


Σχήμα 5.4.4 Ακρίβεια clustering με k-means (55%).

UCI Taiwan Dataset



Σχήμα 5.4.5 Ανάλυση Αριθμού cluster.



Σχήμα 5.4.6 Ακρίβεια clustering με k-means (69%).

Παρατηρήσεις:

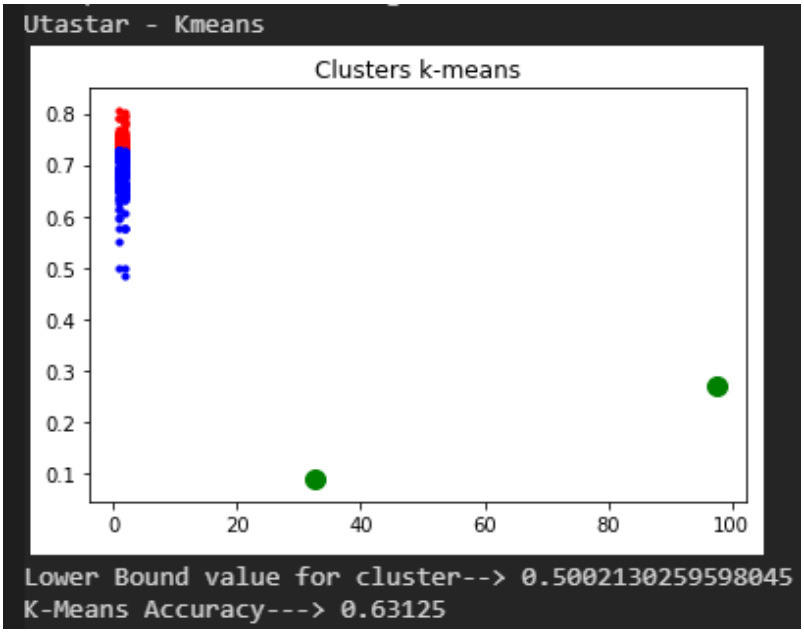
- Διαπιστώνουμε ότι με τα μικρά σύνολα δεδομένων η μέθοδος είχε παρόμοια χαμηλά αποτελέσματα με τις πολυκριτήριες μεθόδους, αλλά στην περίπτωση του UCI Taiwan, που είχε αρκετά δεδομένα μπόρεσε να βελτιωθεί και να πετύχει ακρίβεια 69%.

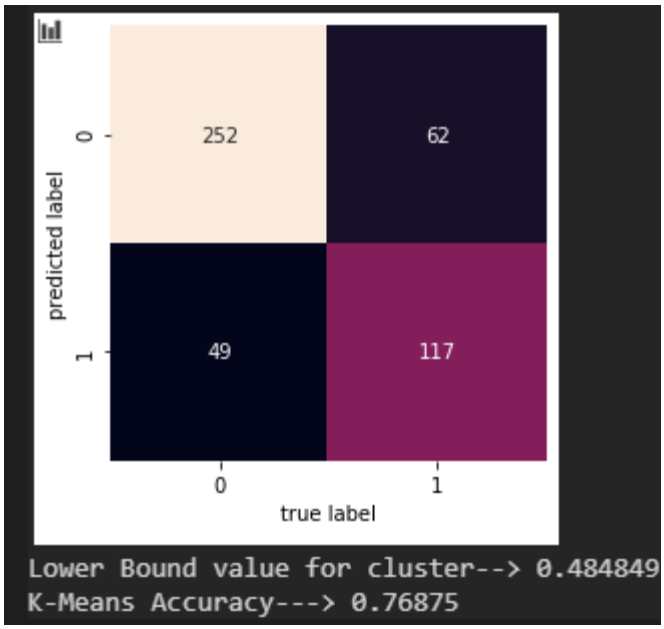
5.5 Συνδυασμός Μεθόδων και Αποτελέσματα

Στο επίκεντρο της παρούσας διπλωματικής εργασίας τοποθετείται η εξέταση του τρόπου με τον οποίο οι μερικές χρησιμότητες που δημιουργούνται από τις πολυκριτήριες μεθόδους θα βοηθήσουν στην αύξηση ακρίβειας στην κατανομή σε συστάδες από k-means, καθώς θεωρούμε ότι οι μερικές χρησιμότητες είναι πολύ καλύτεροι δείκτες από μια απλή κανονικοποίηση.

5.5.1 UTASTAR & k-means

UCI German Dataset



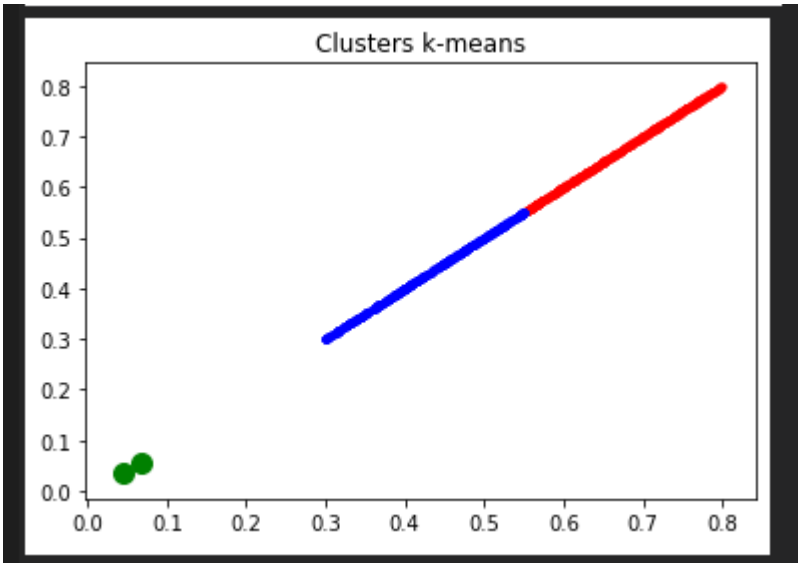


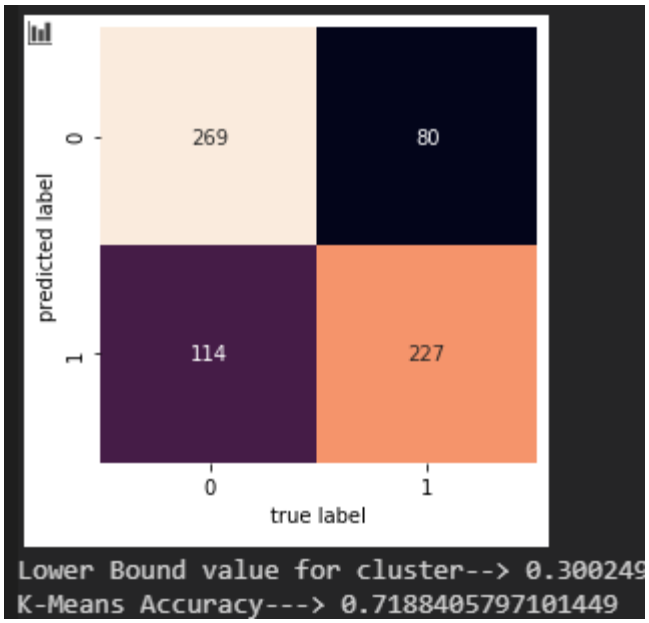
Σχήμα 5.5.1 Ακρίβεια συνδυαστικής μεθόδου UTASTAR & k-means (76%).

Το 0.4848 είναι το κατώτατο όριο της βαθμολογίας, ή αλλιώς ολικής χρησιμότητας, στο οποίο διαχωρίζει η k-means με το score της UTASTAR τις δύο συστάδες/κλάσεις .

Τα σημεία παρουσιάζονται σε δισδιάστατο επίπεδο. Τα δυο πράσινα σημεία είναι το κέντρο των συστάδων, όπου τα μπλε είναι η μια κλάση και τα κόκκινα η άλλη.

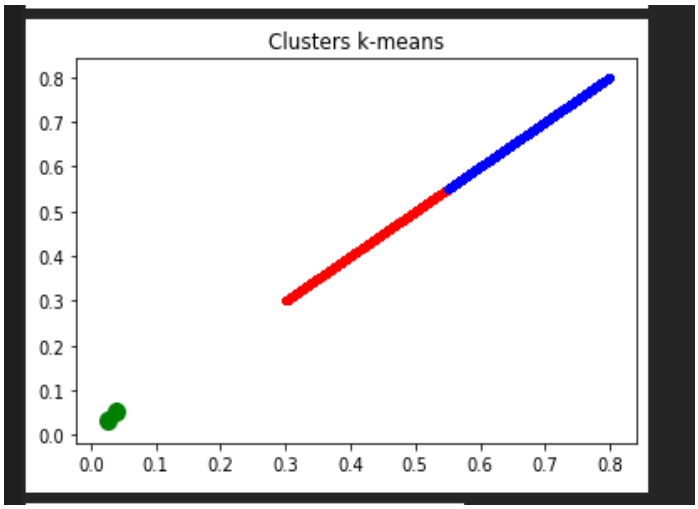
UCI Australian Dataset

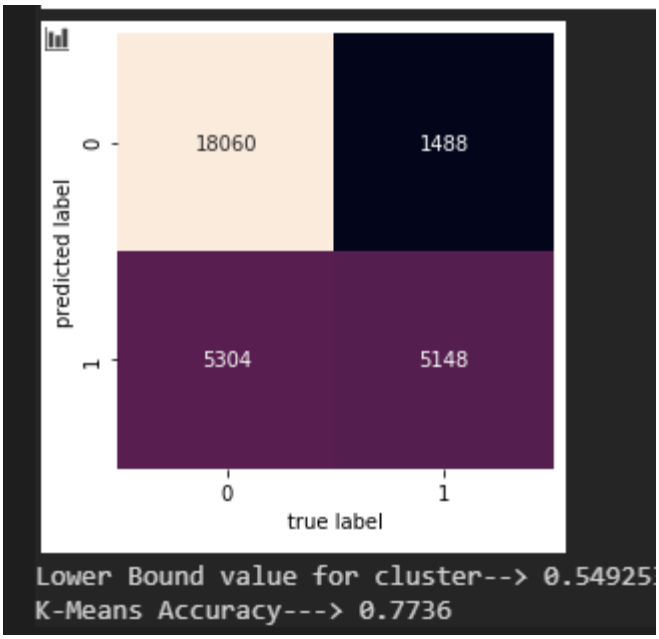




Σχήμα 5.5.2 Ακρίβεια συνδυαστικής μεθόδου UTASTAR & k-means (70%).

UCI Taiwan Dataset

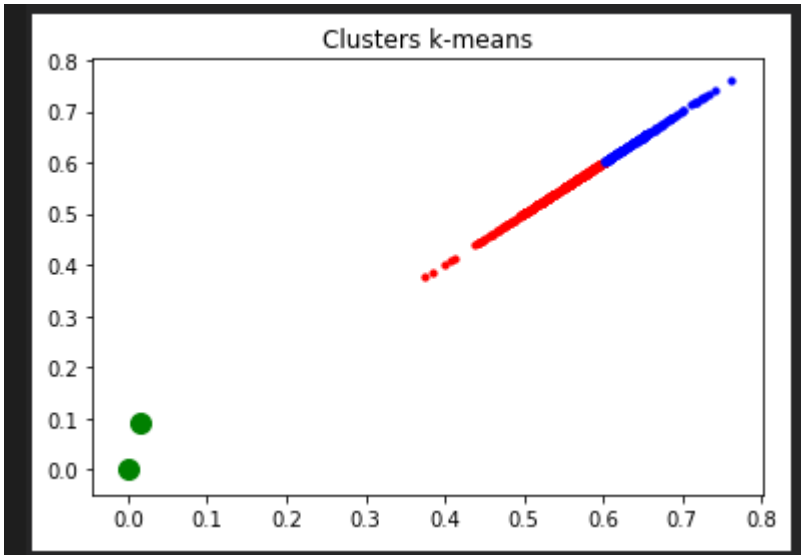


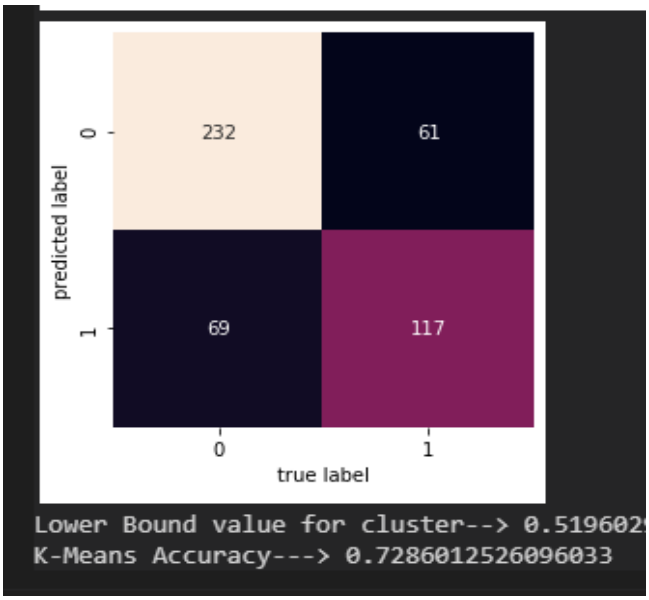


Σχήμα 5.5.3 Ακρίβεια συνδυαστικής μεθοδου UTASTAR & k-means (77%).

5.5.2 UTADIS & k-means

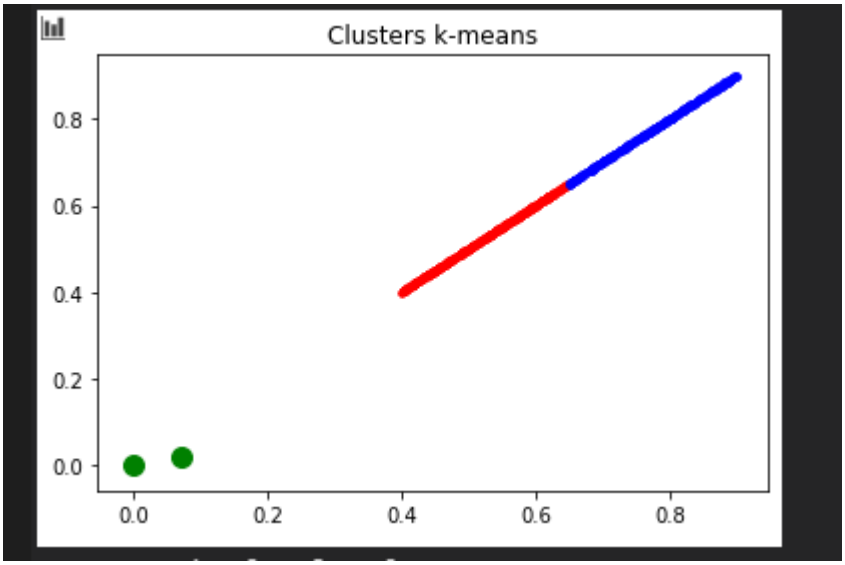
UCI German Dataset

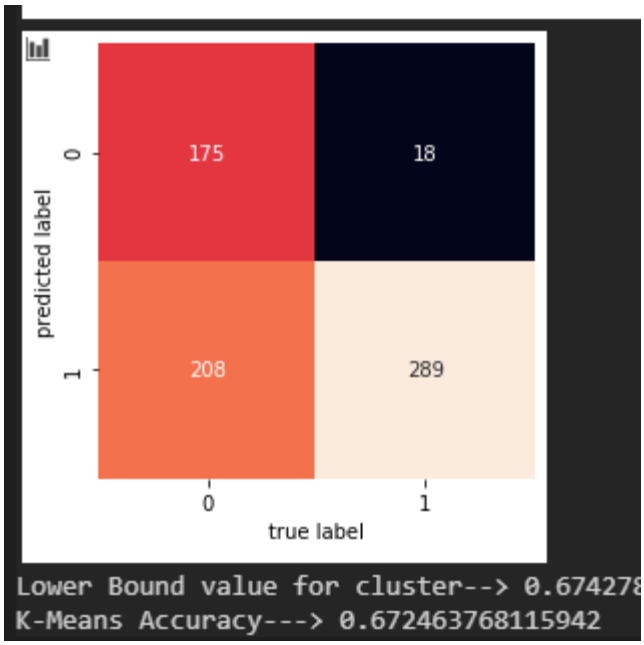




Σχήμα 5.5.4 Ακρίβεια συνδυαστικής μεθόδου UTADIS & k-means (72%).

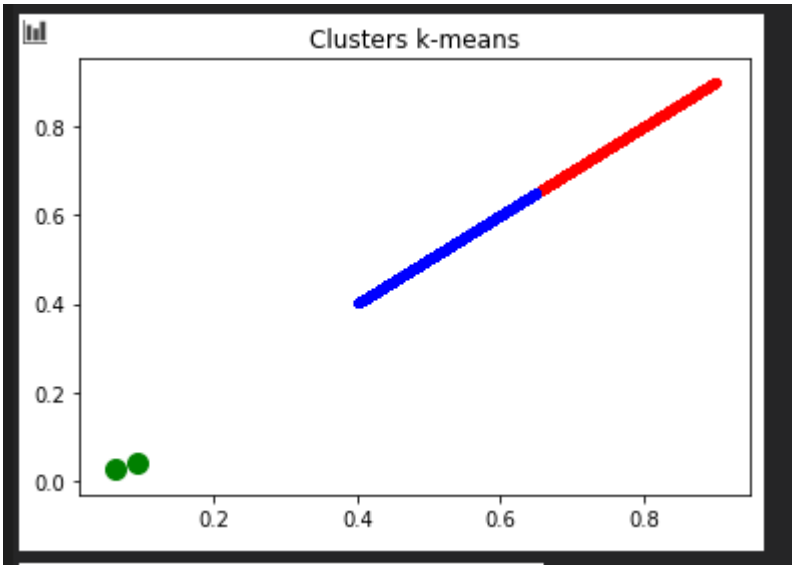
UCI Australian Dataset

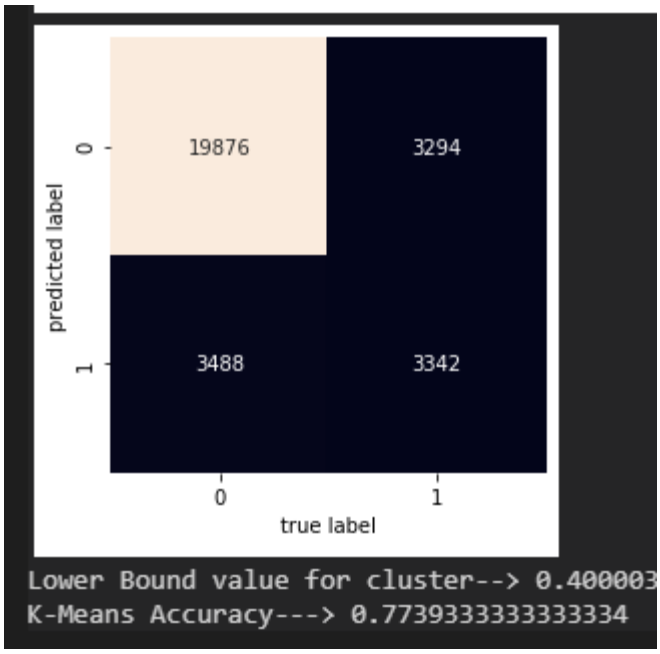




Σχήμα 5.5.5 Ακρίβεια συνδυαστικής μεθόδου UTADIS & k-means (67%).

UCI Taiwan Dataset

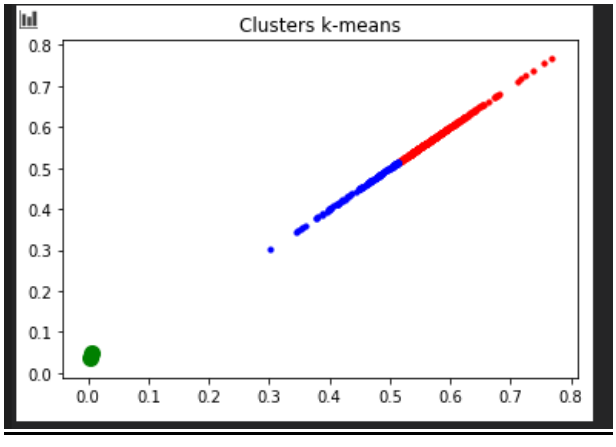


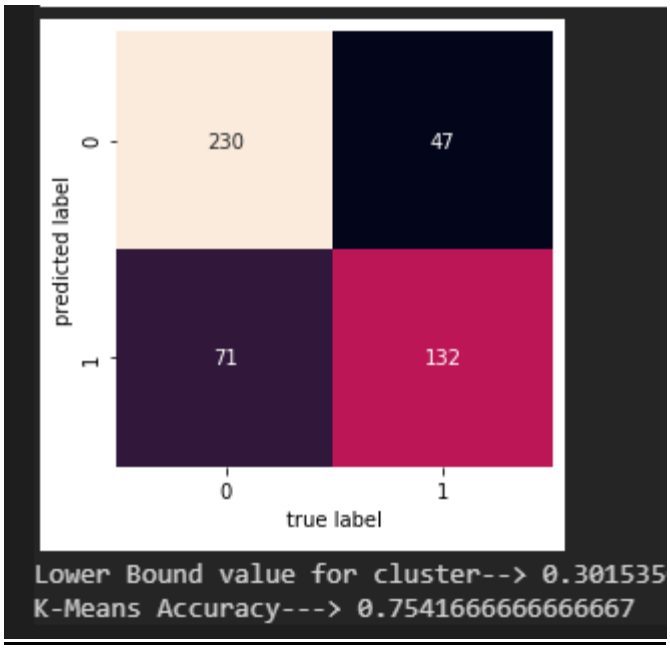


Σχήμα 5.5.6 Ακρίβεια συνδυαστικής μεθόδου UTADIS & k-means (40%).

5.5.3 TOPSIS with UTASTAR Weights & k-means

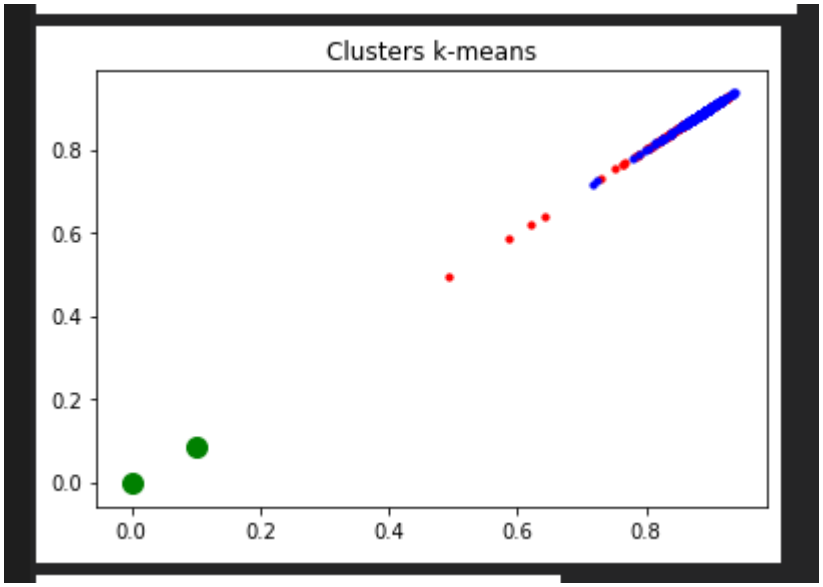
UCI German Dataset

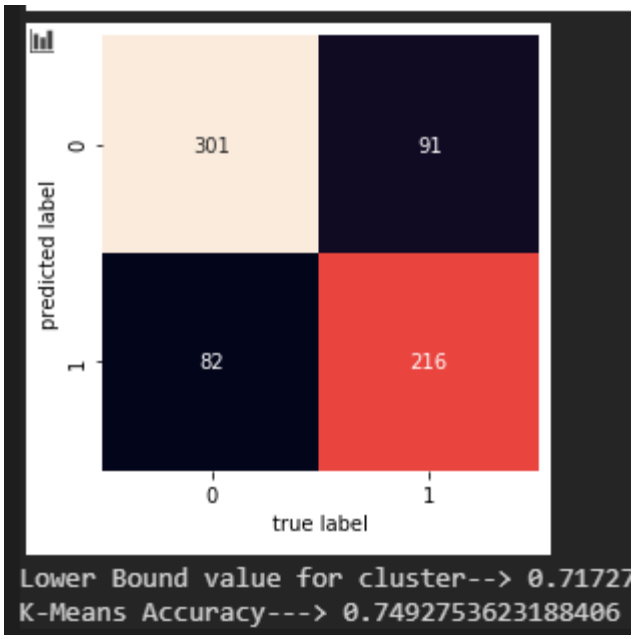




Σχήμα 5.5.7 Ακρίβεια συνδυαστικής μεθόδου TOPSIS with UTASTAR Weights & k-means (75%).

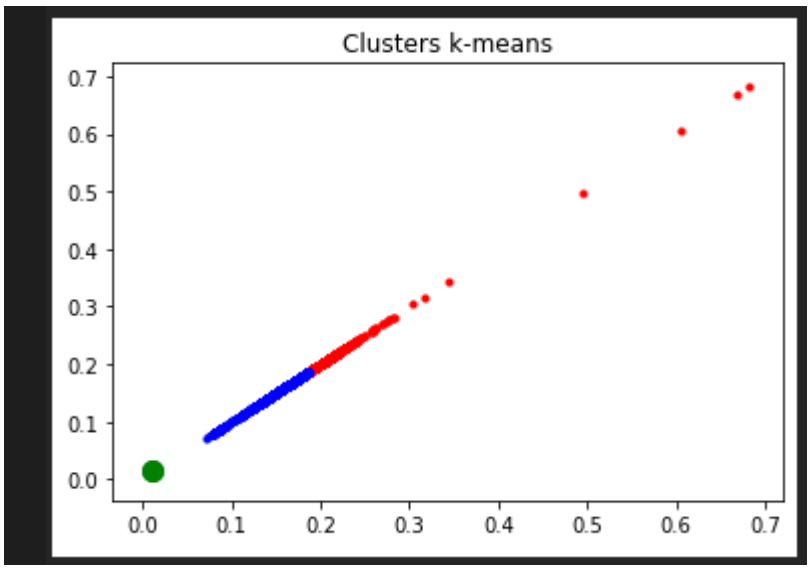
UCI Australian Dataset

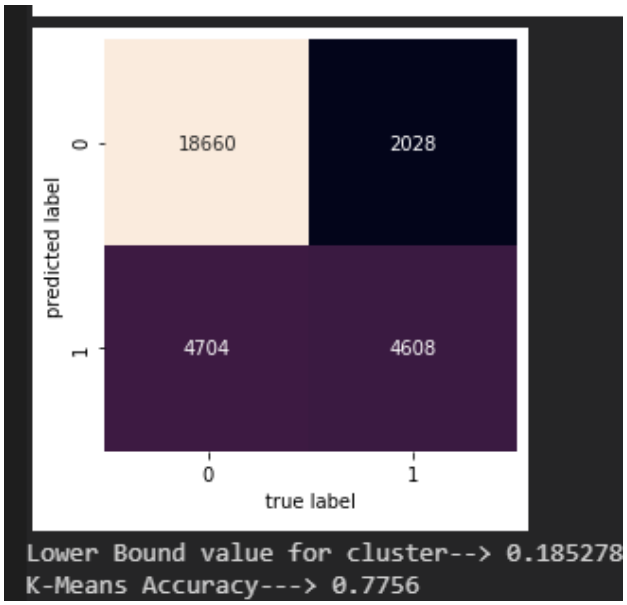




Σχήμα 5.5.8 Ακρίβεια συνδυαστικής μεθόδου TOPSIS with UTASTAR Weights & k-means (74%).

UCI Taiwan Dataset

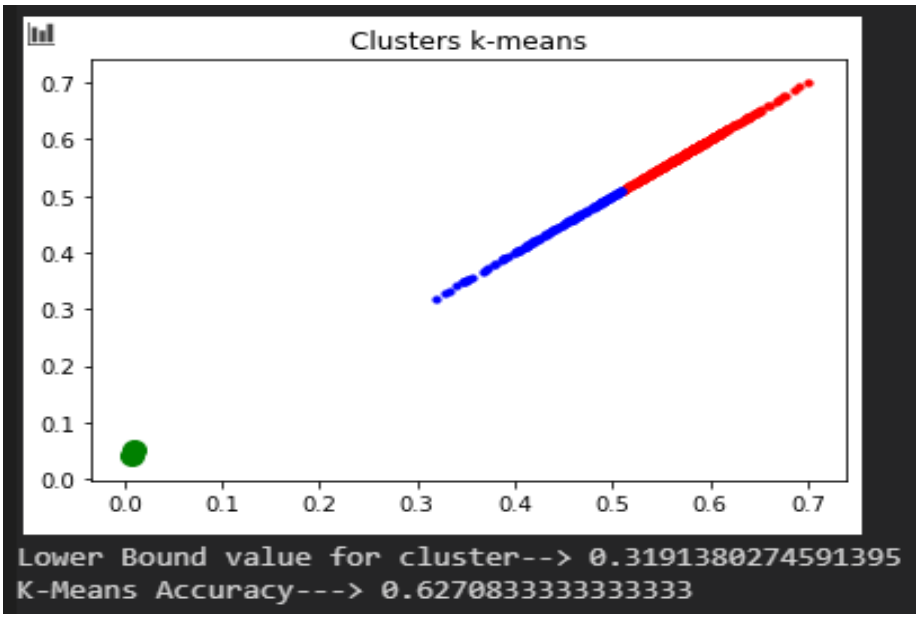


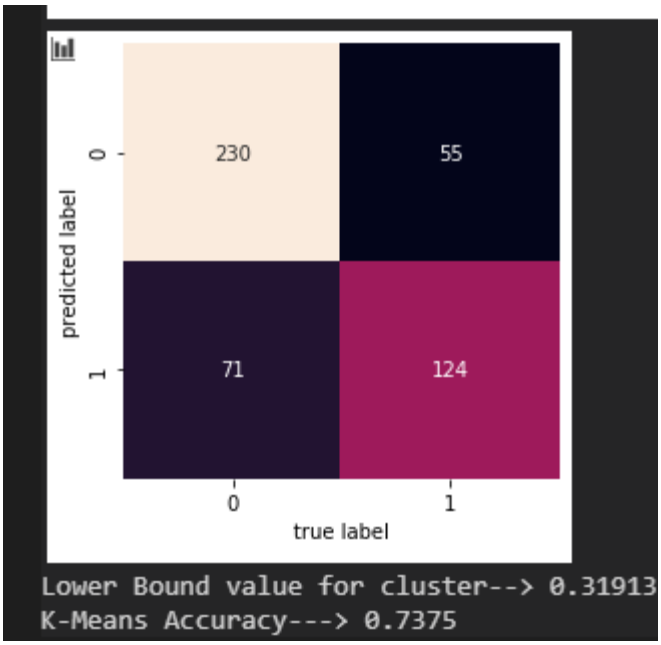


Σχήμα 5.5.9 Ακρίβεια συνδυαστικής μεθόδου TOPSIS with UTASTAR Weights & k-means (77%).

5.5.4 TOPSIS with UTADIS Weights & k-means

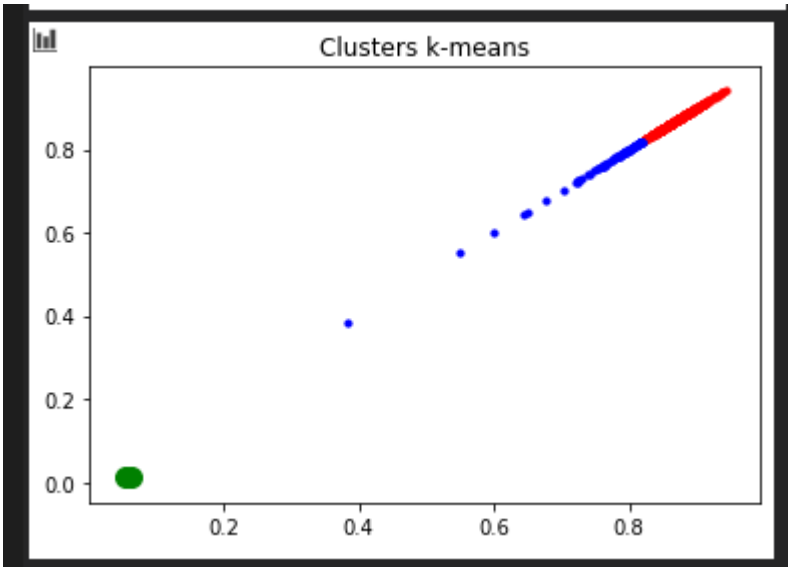
UCI German Dataset

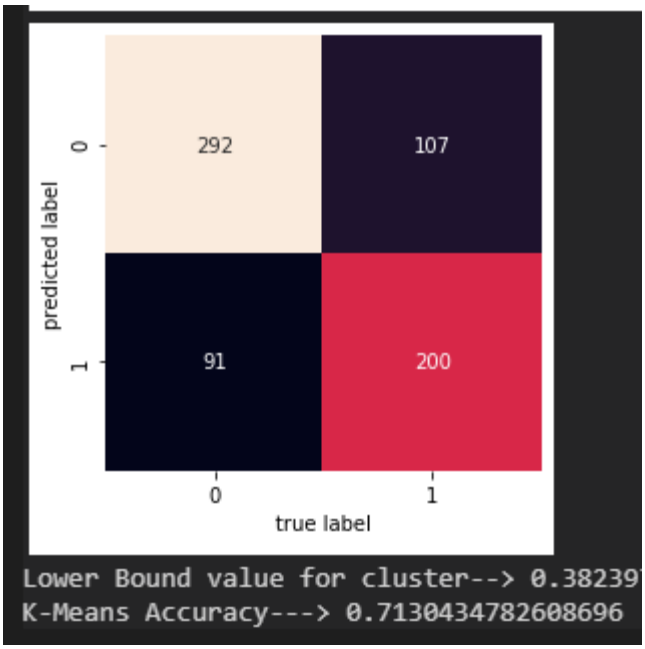




Σχήμα 5.5.10 Ακρίβεια συνδυαστικής μεθόδου TOPSIS with UTADIS Weights & k-means (73%).

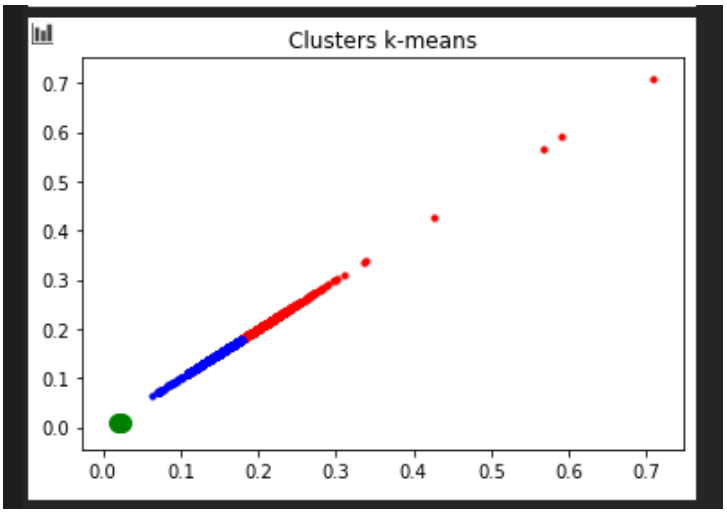
UCI Australian Dataset

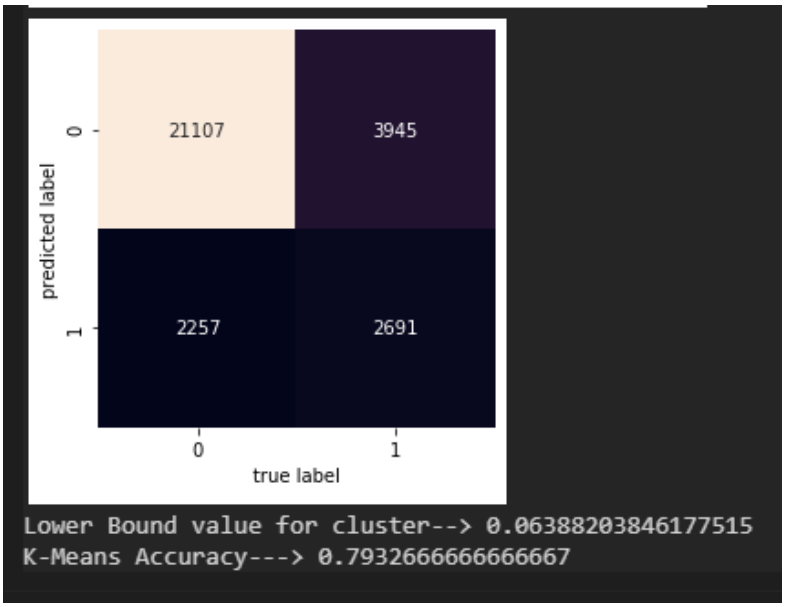




Σχήμα 5.5.11 Ακρίβεια συνδυαστικής μεθόδου TOPSIS with UTADIS Weights & k-means (71%).

UCI Taiwan Dataset





Σχήμα 5.5.12 Ακρίβεια συνδυαστικής μεθόδου TOPSIS with UTADIS Weights & k-means (79%).

Παρατηρήσεις:

- Διαπιστώνουμε επιτυχώς μια ελαφρά αύξηση στην ακρίβεια, όπως περιμέναμε, αλλά όχι αρκετή για να έχουμε επικυρωμένα αποτελέσματα.

5.6 Μείωση διαστάσεων Πολυκριτήριων Μεθόδων

Πραγματοποιήθηκε μείωση κριτηρίων (διαστάσεων), με τη χρήση του ενδεικτικού αθροίσματος στο (0.8-0.9). Οι διαστάσεις αυτές μπήκαν σε σειρά κατάταξης, αναλόγως του βάρους που τους καθορίστηκε από την αντίστοιχη μέθοδο και διατηρήθηκαν αυτές που τηρούν τον παραπάνω κανόνα.

5.6.1 UTASTAR Feature Reduction

UCI German Credit Score

Dimensions Before

	ValueFunc
check_account	0.009145
duration(months)	0.270280
credit_history	0.082980
purpose	0.000750
credit_amount	0.118079
savings_account	0.019279
employment	0.016570
disposable income	0.000132
status_sex	0.068752
debtors_guarantors	0.013889
residence_since	0.009045
property	0.069533
age	0.019629
installment_plans	0.006167
housing	0.077329
num_credits_bank	0.010175
job	0.097490
providor_num	0.003790
telephone	0.002122
foreign	0.104864

Value Functions weight SUM: [0.88930674]

Dimensions after Reduction

	ValueFunc
duration(months)	0.270280
credit_amount	0.118079
foreign	0.104864
job	0.097490
credit_history	0.082980
housing	0.077329
property	0.069533
status_sex	0.068752

Σχήμα 5.6.1 Μείωση διαστάσεων με UTASTAR.

UCI Australian Dataset

```
Dimensions Before
          ValueFunc
A1         0.101439
A2         0.079097
A3         0.084852
A4         0.075381
A5         0.041357
A6         0.107697
A7         0.046333
A8         0.031249
A9         0.104688
A10        0.112951
A11        0.064024
A12        0.065910
A13        0.026111
A14        0.058910
```

```
Value Functions weight SUM: [0.85494941]
Dimensions after Reduction
      ValueFunc
A10    0.112951
A6     0.107697
A9     0.104688
A1     0.101439
A3     0.084852
A2     0.079097
A4     0.075381
A12    0.065910
A11    0.064024
A14    0.058910
```

Σχήμα 5.6.2 Μείωση διαστάσεων με UTASTAR.

UCI Taiwan dataset

```
Dimensions Before
             ValueFunc
LIMIT_BAL    0.058347
SEX          0.082238
EDUCATION    0.081046
MARRIAGE     0.030273
AGE          0.024751
PAY_0        0.026671
PAY_2        0.069513
PAY_3        0.041283
PAY_4        0.064165
PAY_5        0.075347
PAY_6        0.080389
BILL_AMT1    0.061784
BILL_AMT2    0.043499
BILL_AMT3    0.008296
BILL_AMT4    0.035611
BILL_AMT5    0.025596
BILL_AMT6    0.005115
PAY_AMT1     0.029068
PAY_AMT2     0.060027
PAY_AMT3     0.023865
PAY_AMT4     0.011352
PAY_AMT5     0.029524
PAY_AMT6     0.032242
```

```
Value Functions weight SUM: [0.87435478]
Dimensions after Reduction
             ValueFunc
SEX          0.082238
EDUCATION    0.081046
PAY_6        0.080389
PAY_5        0.075347
PAY_2        0.069513
PAY_4        0.064165
BILL_AMT1    0.061784
PAY_AMT2     0.060027
LIMIT_BAL    0.058347
BILL_AMT2    0.043499
PAY_3        0.041283
BILL_AMT4    0.035611
PAY_AMT6     0.032242
MARRIAGE     0.030273
PAY_AMT5     0.029524
PAY_AMT1     0.029068
```

Σχήμα 5.6.3 Μείωση διαστάσεων με UTASTAR.

5.6.2 UTADIS Feature Reduction

UCI German Credit Score

```
Dimensions Before
                    ValueFunc
check_account       0.016744
duration(months)    0.240232
credit_history      0.092162
purpose             0.002083
credit_amount       0.060583
savings_account     0.090179
employment          0.057304
disposable income   0.001057
status_sex          0.065426
debtors_guarantors  0.032817
residence_since     0.036848
property            0.066131
age                 0.027318
installment_plans   0.031895
housing             0.104529
num_credits_bank    0.001150
job                 0.001483
providor_num        0.015193
telephone           0.017888
foreign             0.038979
```

```
Value Functions weight SUM: [0.8851882]
Dimensions after Reduction
                    ValueFunc
duration(months)    0.240232
housing             0.104529
credit_history      0.092162
savings_account     0.090179
property            0.066131
status_sex          0.065426
credit_amount       0.060583
employment          0.057304
foreign             0.038979
residence_since     0.036848
debtors_guarantors  0.032817
```

Σχήμα 5.6.4 Μείωση διαστάσεων με UTADIS.

UCI Australian Dataset

```
Dimensions Before
     ValueFunc
A1    0.072982
A2    0.132779
A3    0.017280
A4    0.127776
A5    0.033840
A6    0.132694
A7    0.089226
A8    0.010168
A9    0.007161
A10   0.127674
A11   0.088975
A12   0.095261
A13   0.052465
A14   0.011721
```

```
Value Functions weight SUM: [0.86736613]
Dimensions after Reduction
     ValueFunc
A2    0.132779
A6    0.132694
A4    0.127776
A10   0.127674
A12   0.095261
A7    0.089226
A11   0.088975
A1    0.072982
```

Σχήμα 5.6.5 Μείωση διαστάσεων με UTADIS.

UCI Taiwan dataset

```
Dimensions Before
            ValueFunc
LIMIT_BAL   0.121390
SEX         0.004334
EDUCATION   0.051408
MARRIAGE    0.069309
AGE         0.016053
PAY_0       0.017375
PAY_2       0.152075
PAY_3       0.030813
PAY_4       0.028720
PAY_5       0.089030
PAY_6       0.000213
BILL_AMT1   0.109062
BILL_AMT2   0.046717
BILL_AMT3   0.010488
BILL_AMT4   0.021670
BILL_AMT5   0.037437
BILL_AMT6   0.009482
PAY_AMT1    0.054336
PAY_AMT2    0.058539
PAY_AMT3    0.027017
PAY_AMT4    0.005854
PAY_AMT5    0.003214
PAY_AMT6    0.035466
```

```
Value Functions weight SUM: [0.88430057]
Dimensions after Reduction
            ValueFunc
PAY_2       0.152075
LIMIT_BAL   0.121390
BILL_AMT1   0.109062
PAY_5       0.089030
MARRIAGE    0.069309
PAY_AMT2    0.058539
PAY_AMT1    0.054336
EDUCATION   0.051408
BILL_AMT2   0.046717
BILL_AMT5   0.037437
PAY_AMT6    0.035466
PAY_3       0.030813
PAY_4       0.028720
```

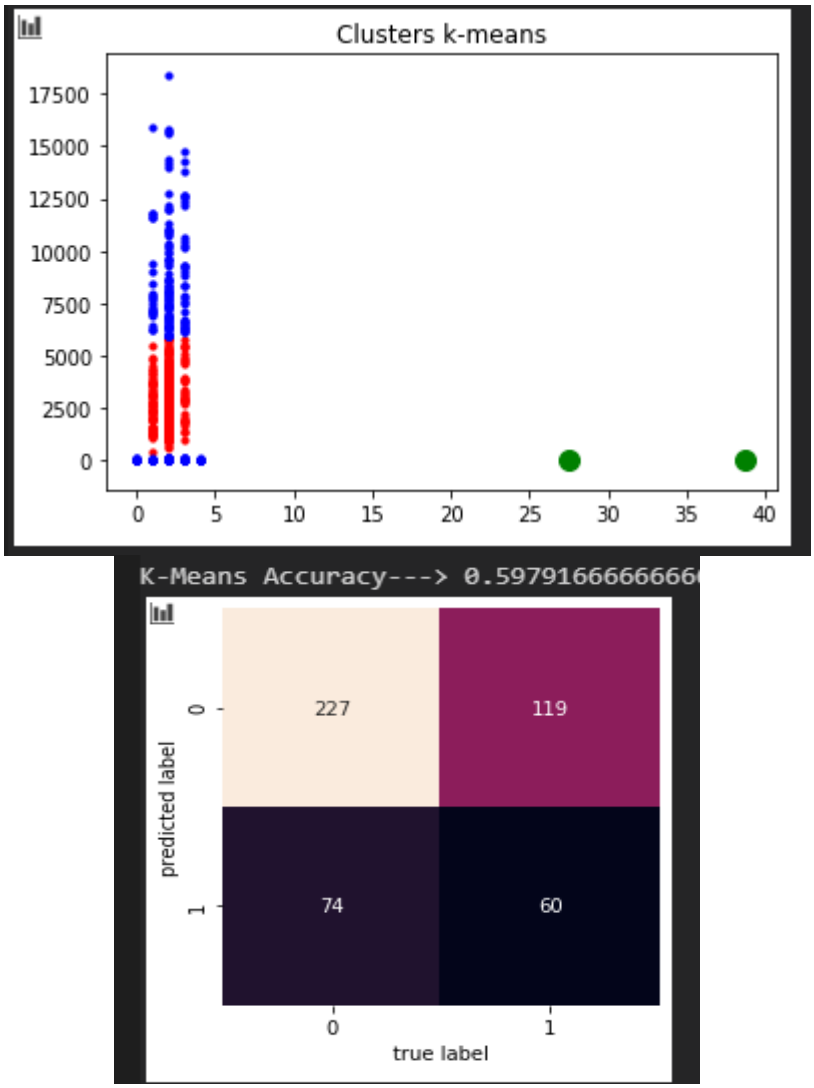
Σχήμα 5.6.6 Μείωση διαστάσεων με UTADIS.

5.7 Συνδυασμός Μεθόδων και Τελικά Αποτελέσματα με Μειωμένες Διαστάσεις

Εν τέλει, μετά από την μείωση διαστάσεων, αναμένεται η μέγιστη ακρίβεια που είναι δυνατόν να επιτευχθεί, συνδυάζοντας όλες τις μεθόδους.

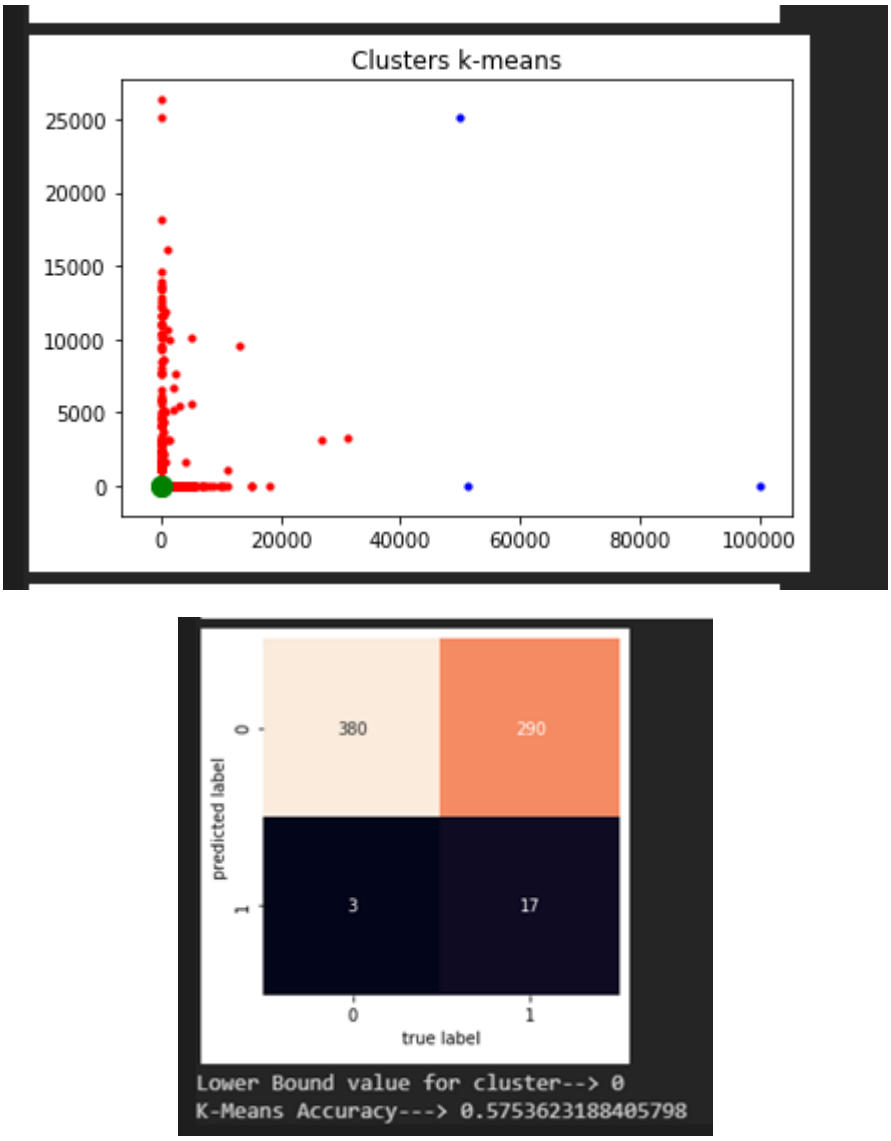
5.7.1 k-means with original data and feature reduction

UCI German Credit Score



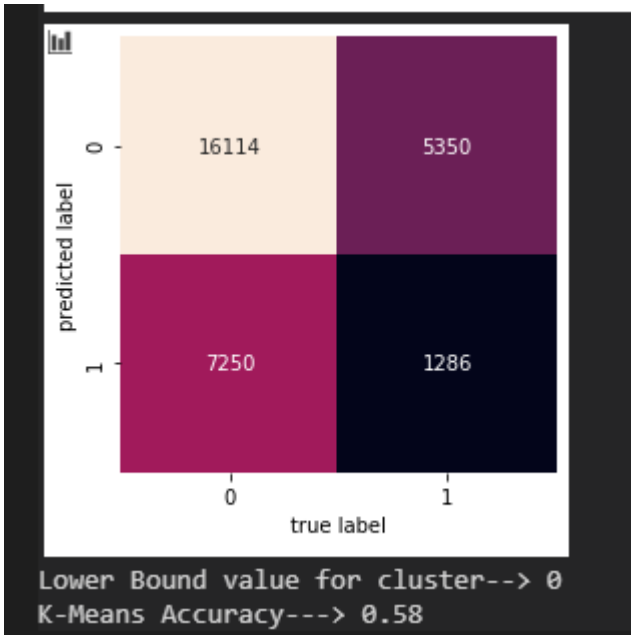
Σχήμα 5.7.1 Ακρίβεια ταξινόμησης k-means (59%).

UCI Australian Dataset



Σχήμα 5.7.2 Ακρίβεια ταξινόμησης k-means (57%).

UCI Taiwan Dataset



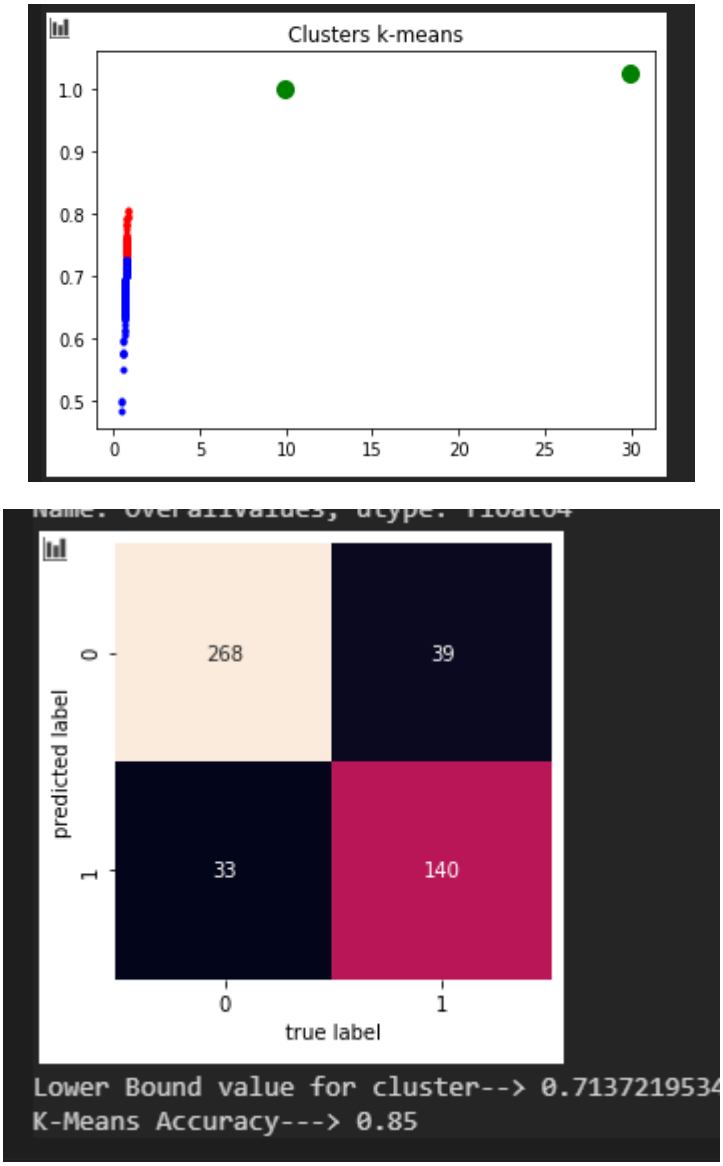
Σχήμα 5.7.3 Ακρίβεια ταξινόμησης k-means (59%).

Παρατηρήσεις:

- Ως σημείο αναφοράς εξετάζουμε και πάλι την ακρίβεια της k-means με μειωμένες διαστάσεις με αποτέλεσμα να μην παρατηρούμε σημαντική βελτίωση.

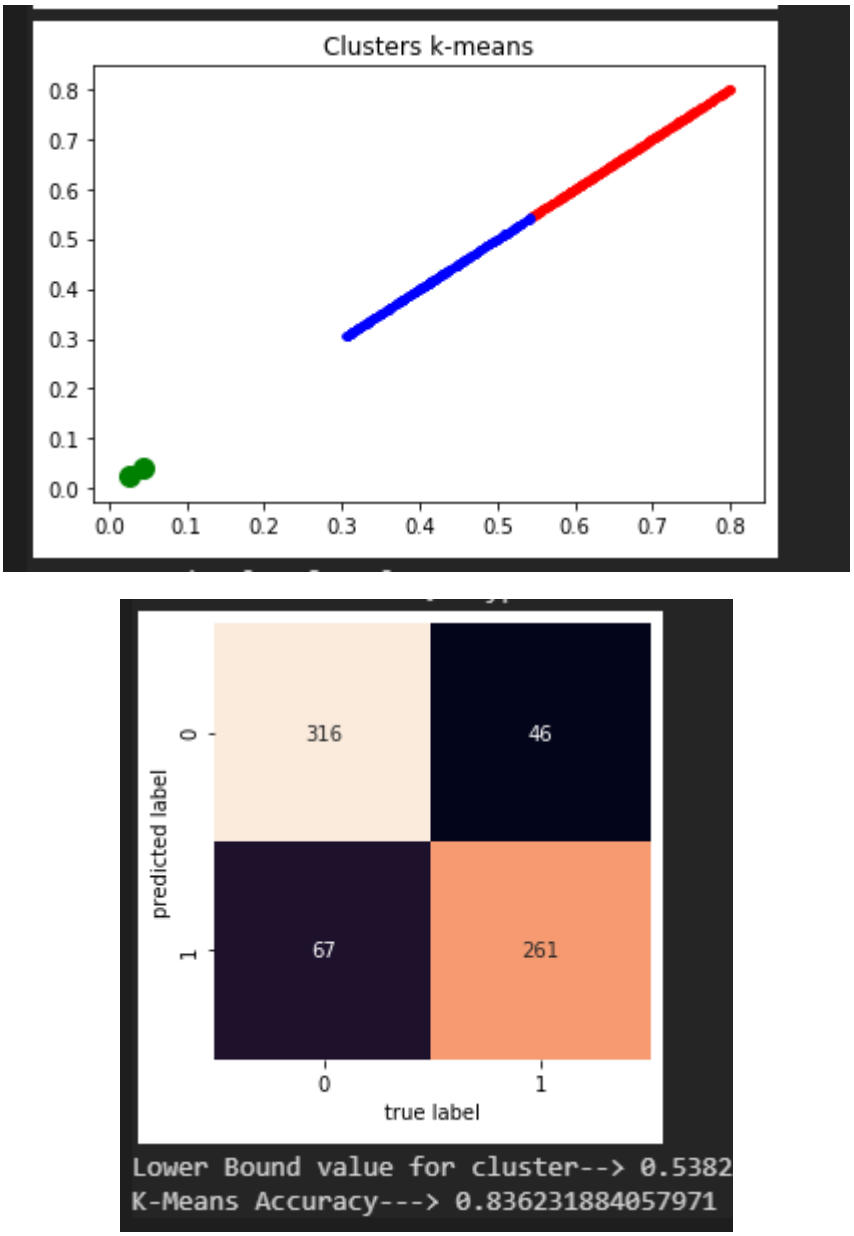
5.7.2 UTASTAR & k-means

UCI German Credit Score



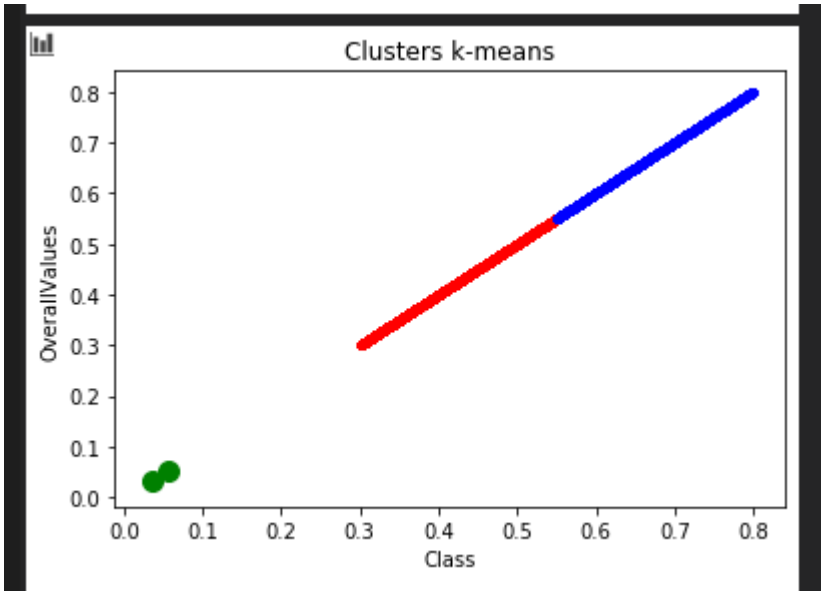
Σχήμα 5.7.4 Ακρίβεια ταξινόμησης UTASTAR & k-means (88%).

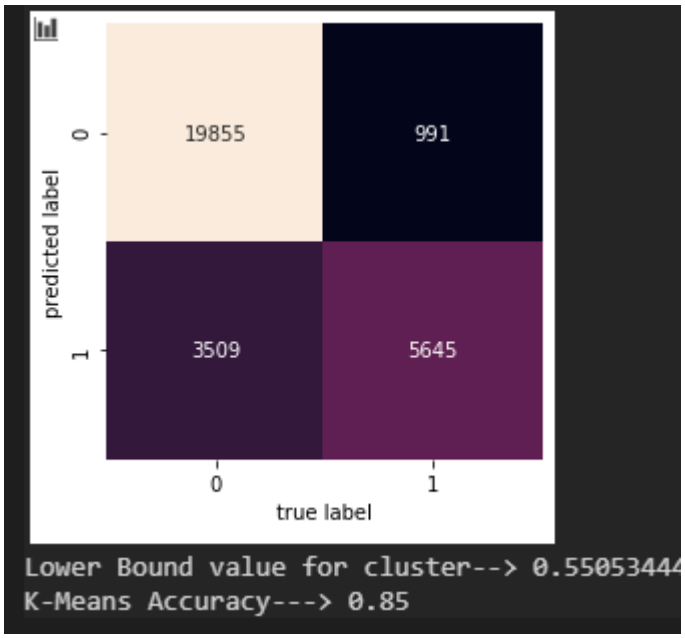
UCI Australian Dataset



Σχήμα 5.7.5 Ακρίβεια ταξινόμησης UTASTAR & k-means (83%).

UCI Taiwan Dataset

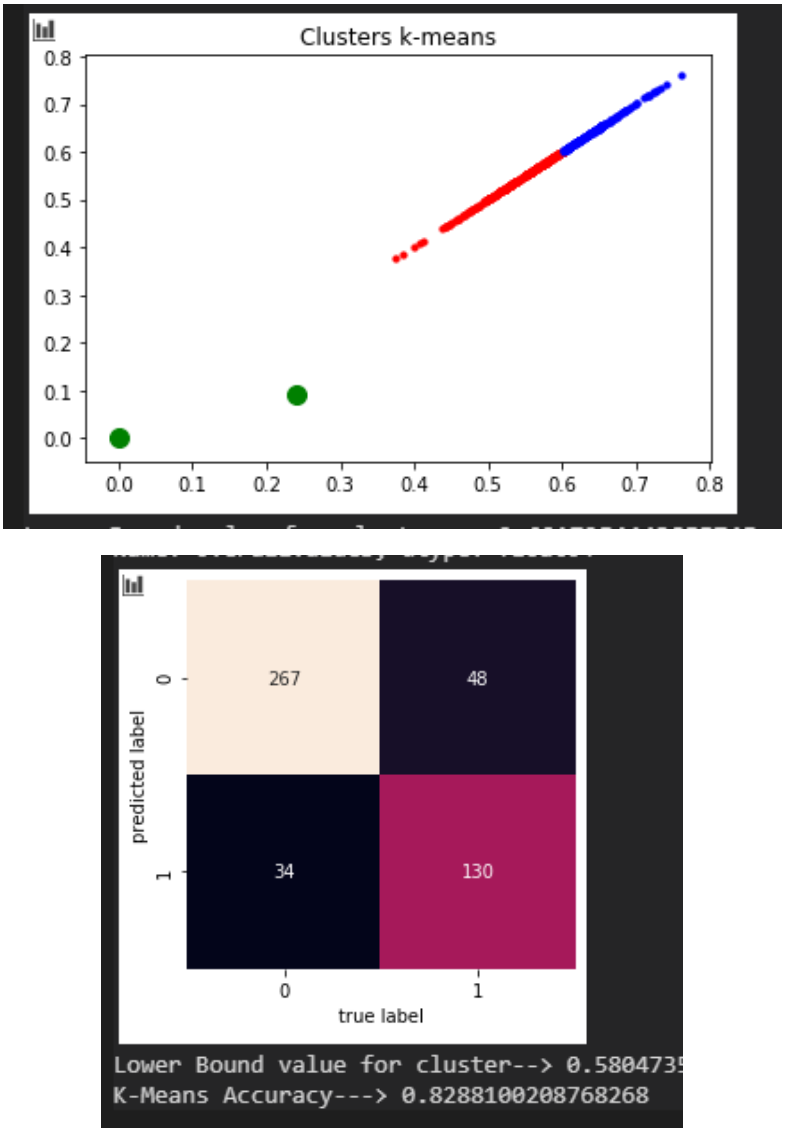




Σχήμα 5.7.6 Ακρίβεια ταξινόμησης UTASTAR & k-means (84%).

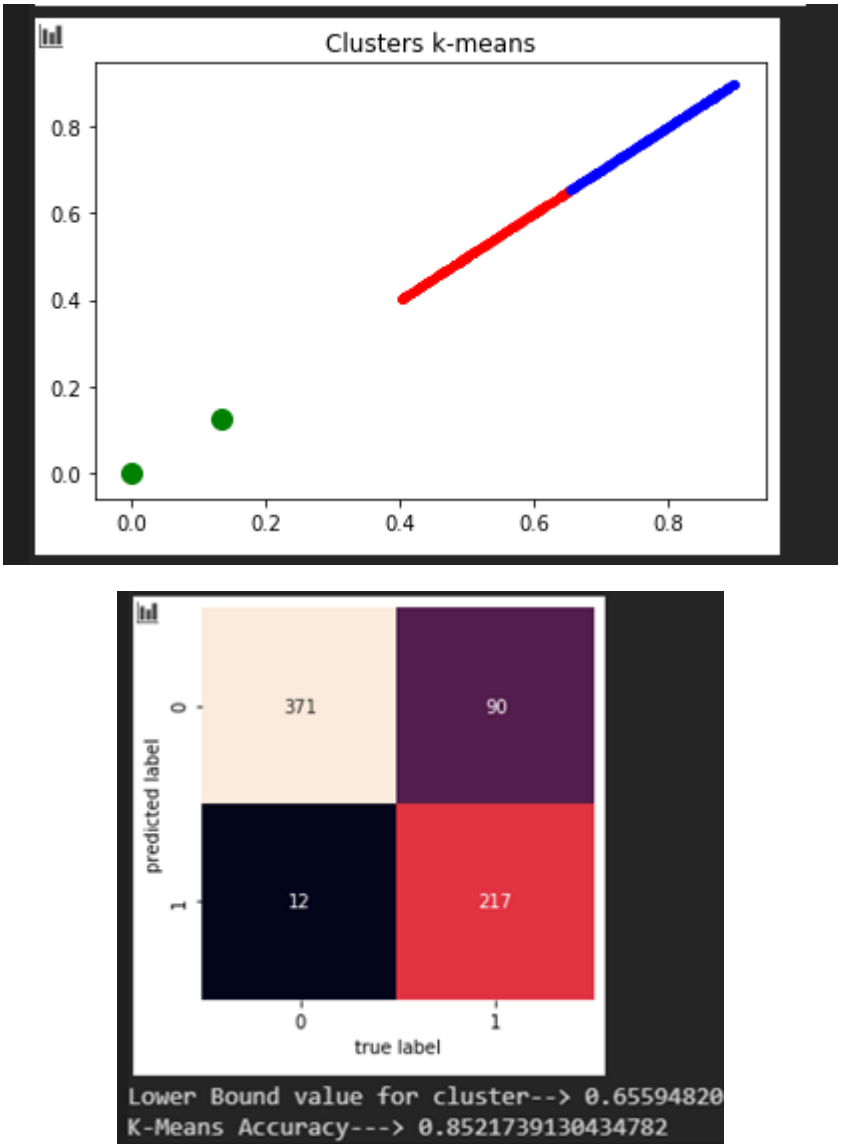
5.7.3 UTADIS & k-means

UCI German Credit Score



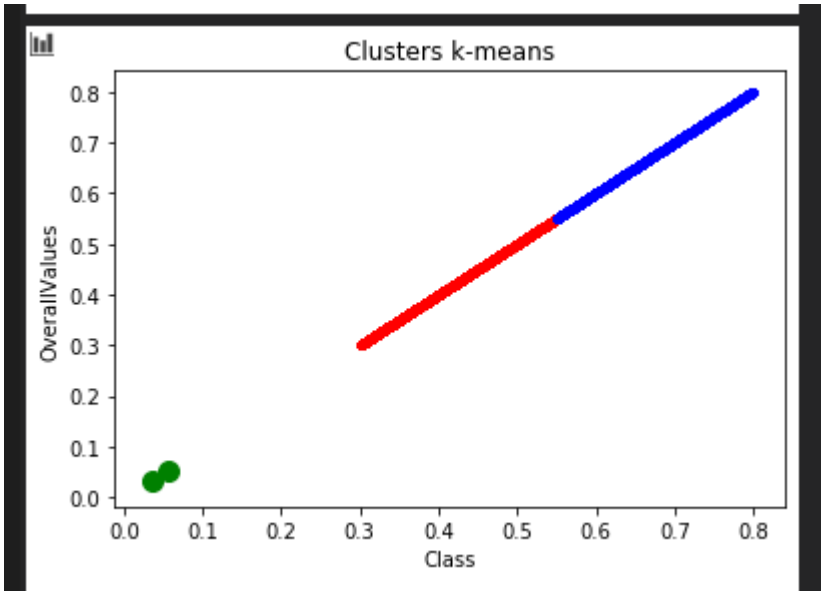
Σχήμα 5.7.7 Ακρίβεια ταξινόμησης UTADIS & k-means (82%).

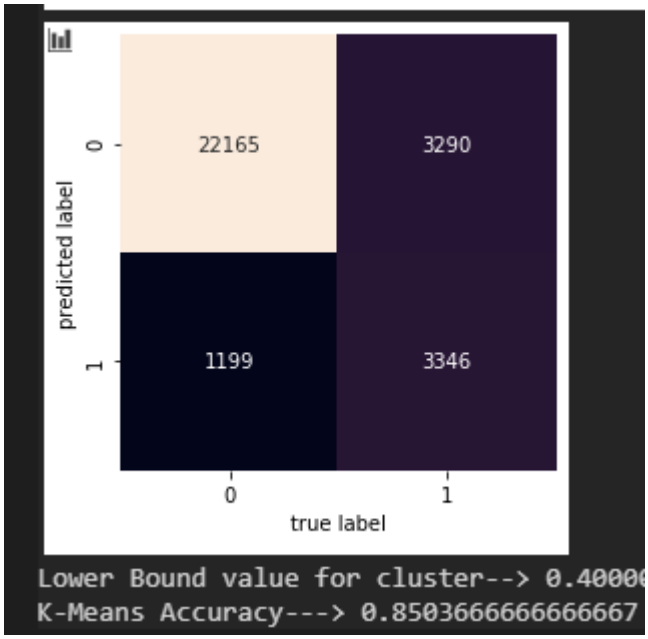
UCI Australian Dataset



Σχήμα 5.7.8 Ακρίβεια ταξινόμησης UTADIS & k-means (84%).

UCI Taiwan Dataset

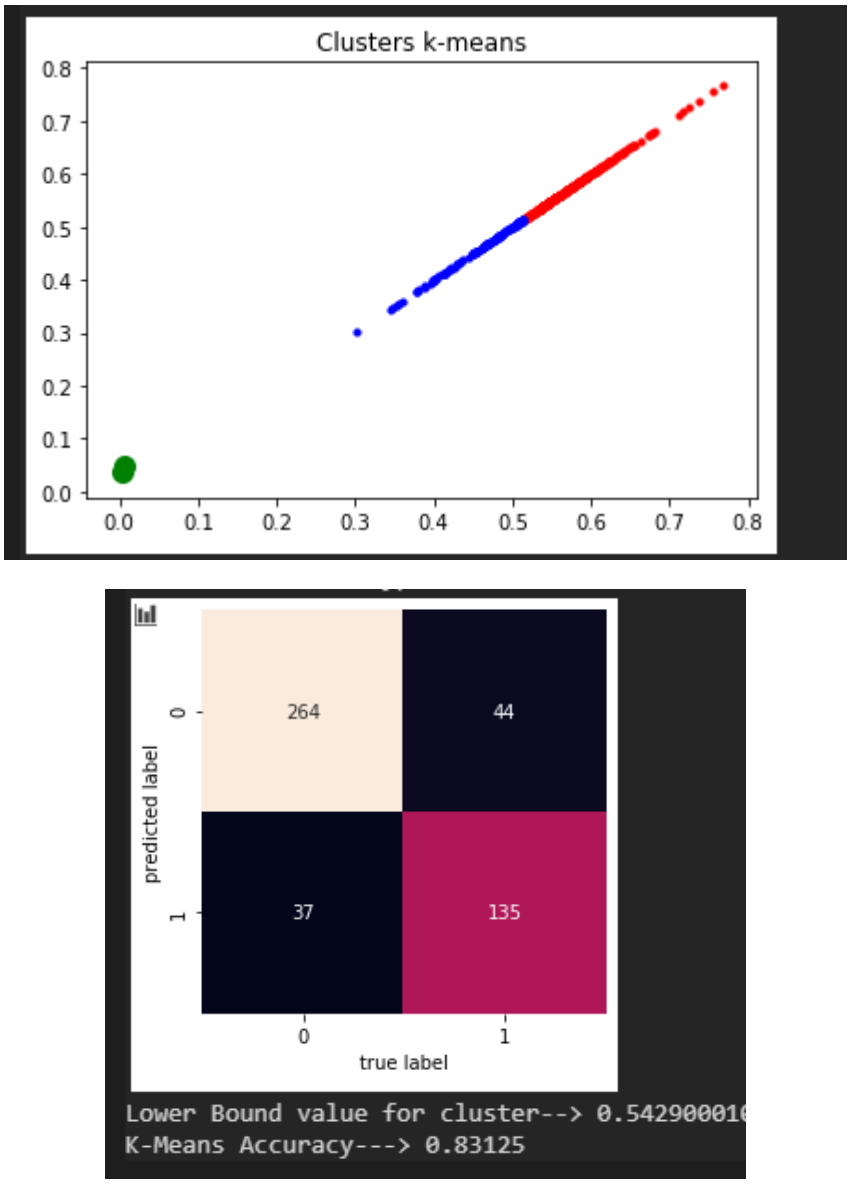




Σχήμα 5.7.9 Ακρίβεια ταξινόμησης UTADIS & k-means (85%).

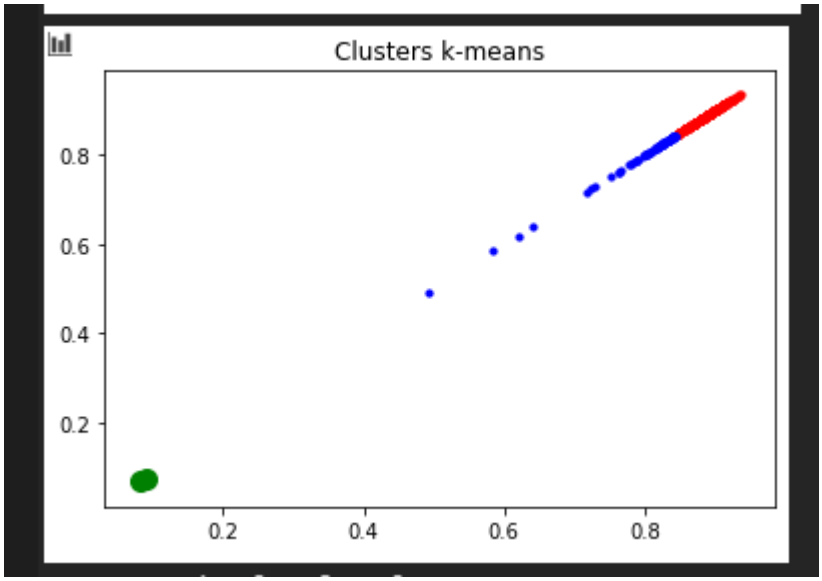
5.7.4 TOPSIS with UTASTAR weights & k-means

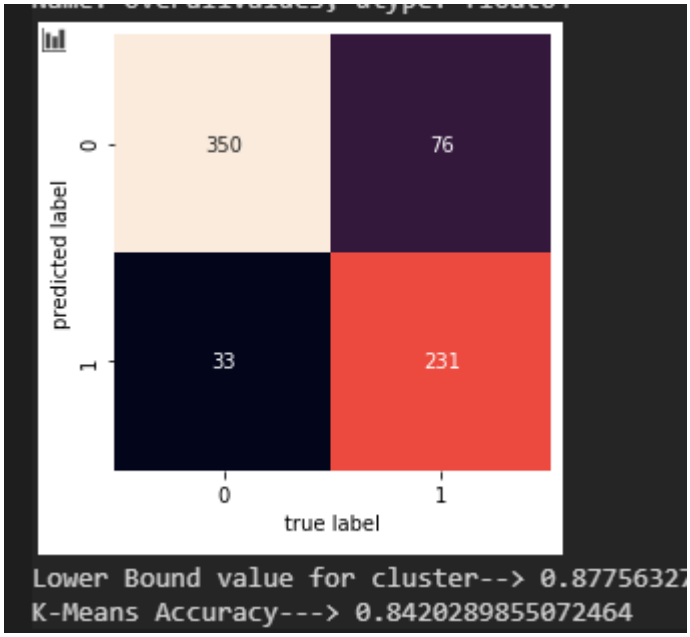
UCI German Credit Score



Σχήμα 5.7.10 Ακρίβεια ταξινόμησης TOPSIS with UTASTAR weights & k-means (83%).

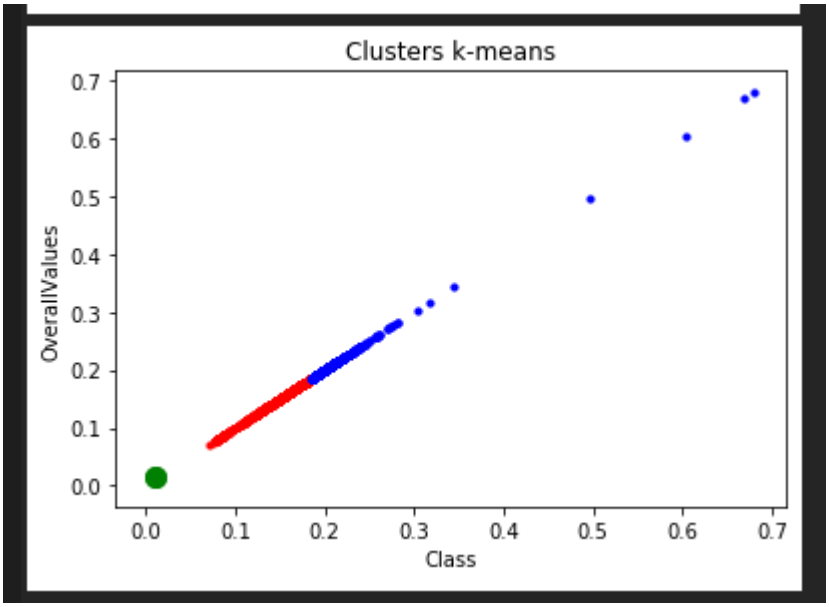
UCI Australian Dataset

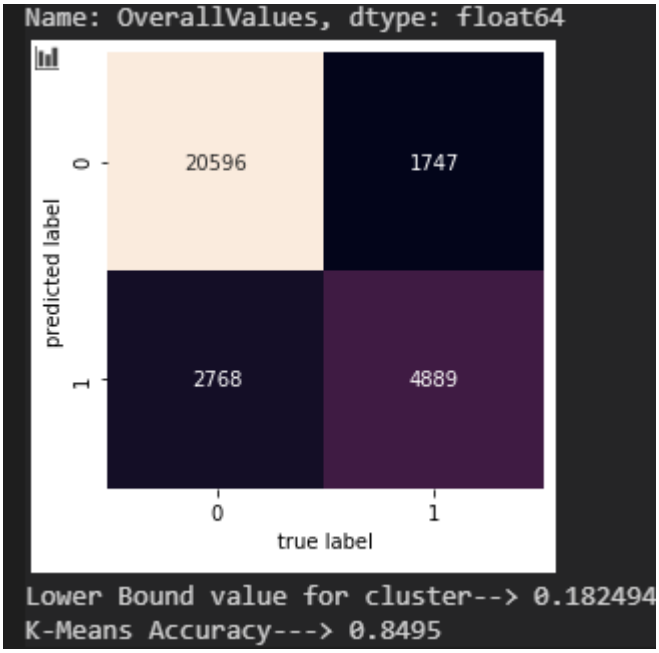




Σχήμα 5.7.11 Ακρίβεια ταξινόμησης TOPSIS with UTASTAR weights & k-means (84%).

UCI Taiwan Dataset

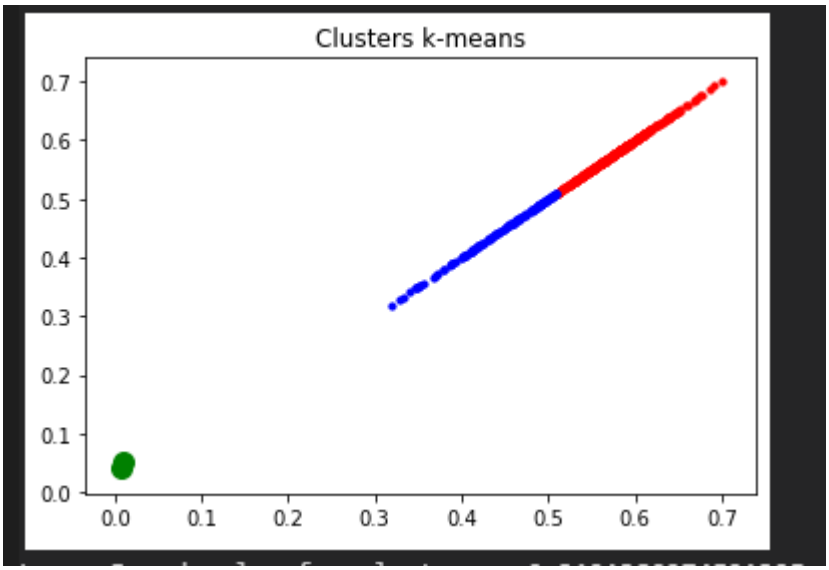


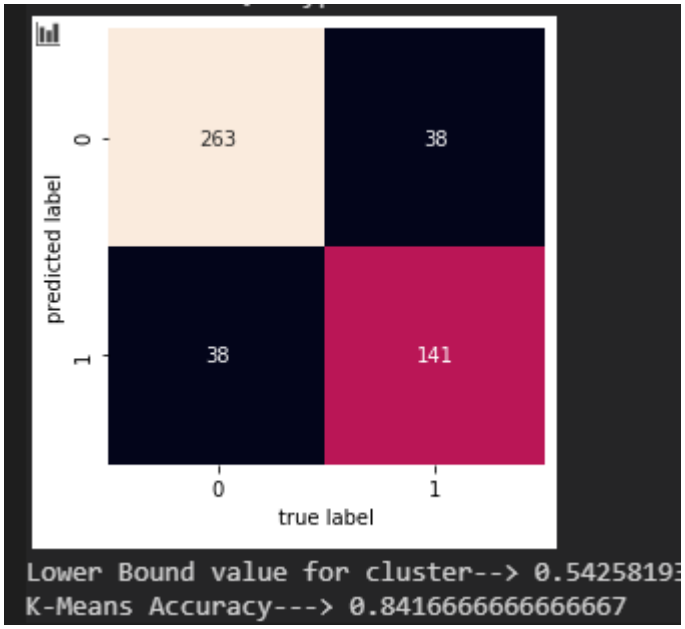


Σχήμα 5.7.12 Ακρίβεια ταξινόμησης TOPSIS with UTASTAR weights & k-means (84%).

5.7.5 TOPSIS with UTADIS weights & k-means

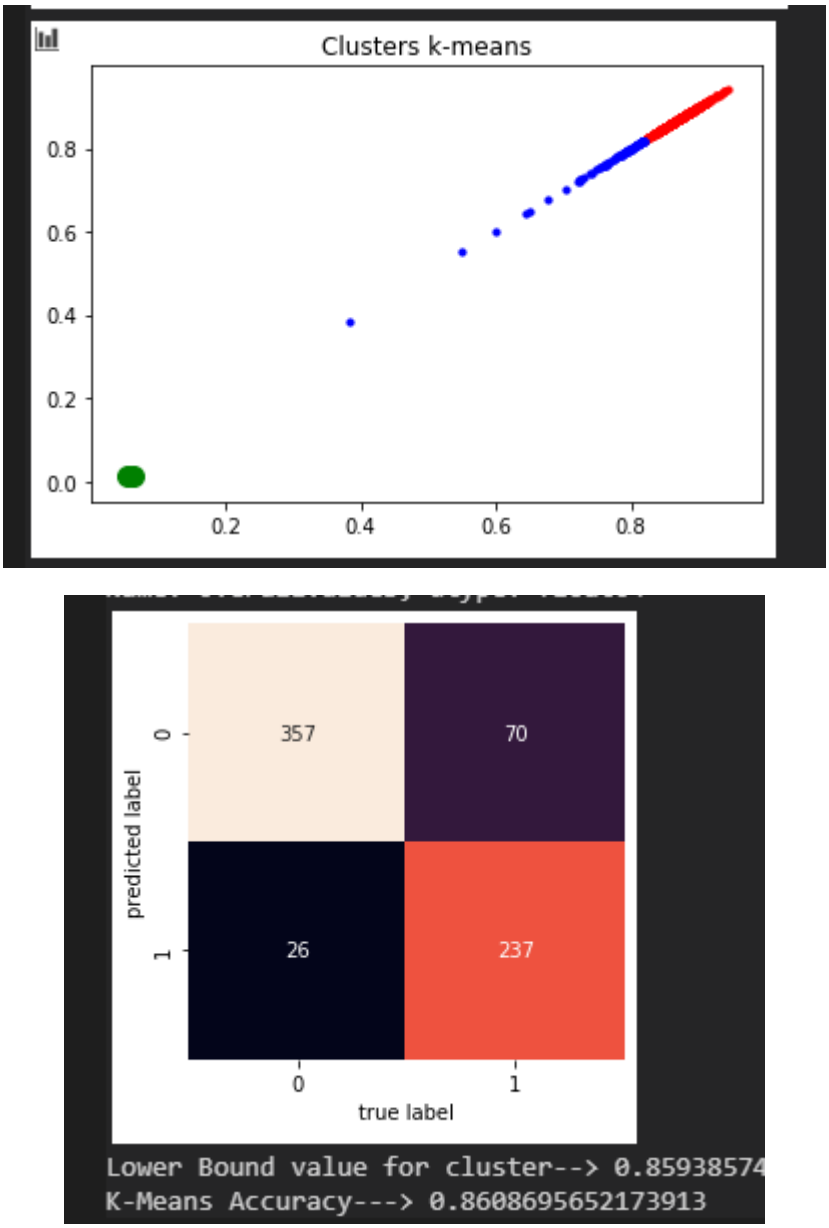
UCI German Credit Score





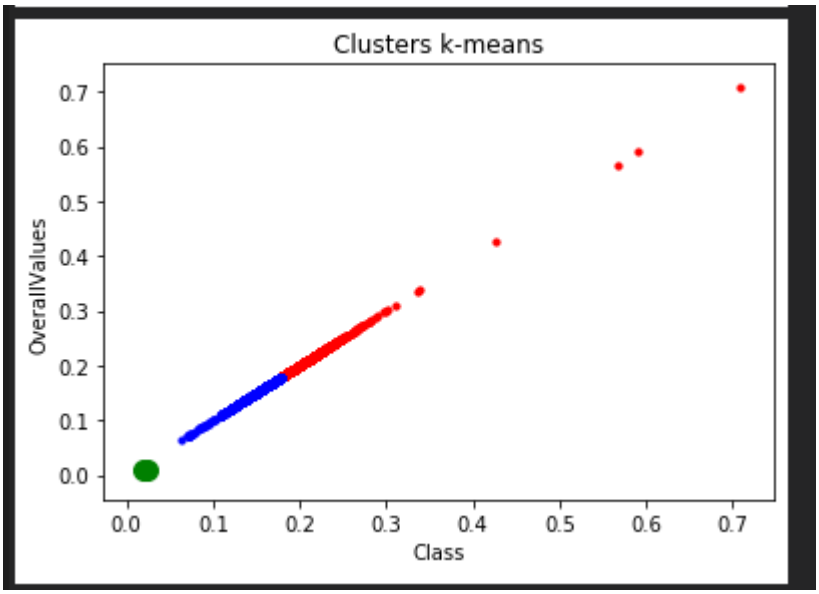
Σχήμα 5.7.13 Ακρίβεια ταξινόμησης TOPSIS with UTADIS weights & k-means (84%).

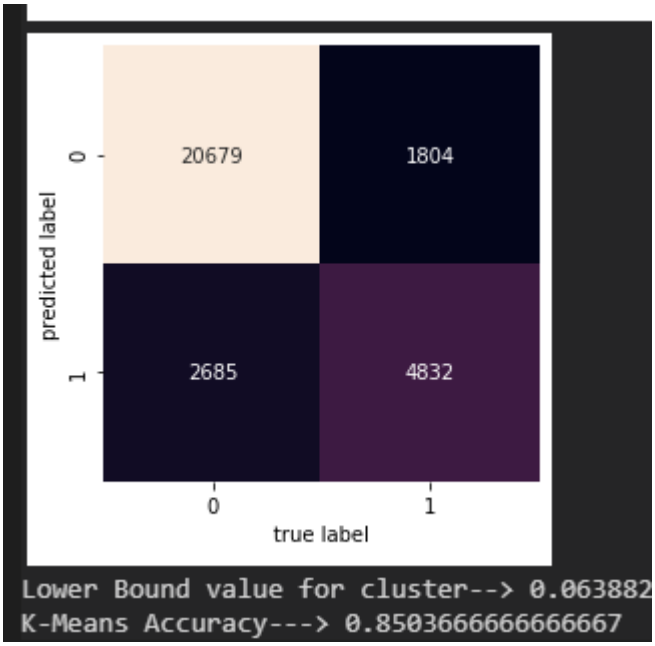
UCI Australian Dataset



Σχήμα 5.7.14 Ακρίβεια ταξινόμησης TOPSIS with UTADIS weights & k-means (86%).

UCI Taiwan Dataset

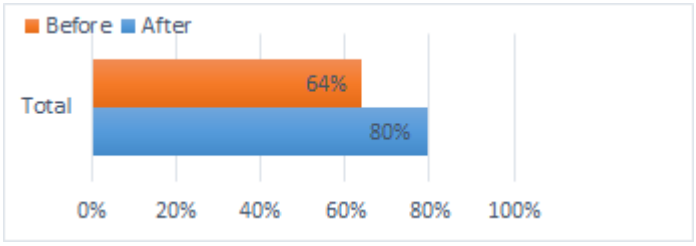




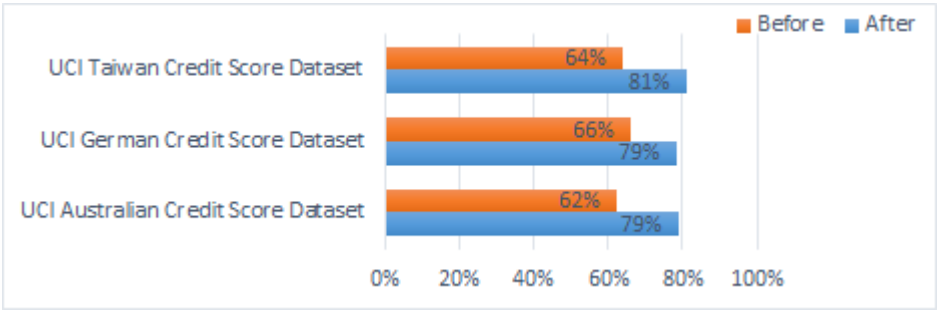
Σχήμα 5.7.15 Ακρίβεια ταξινόμησης TOPSIS with UTADIS weights & k-means (85%).

5.8 Συγκρίσεις και Ολικά αποτελέσματα

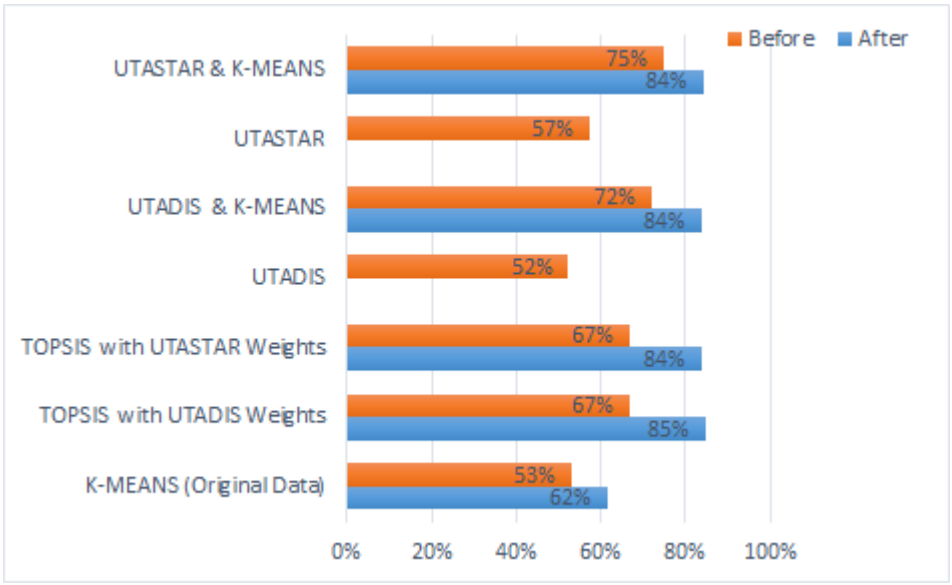
Παρουσιάζουμε τώρα συγκεντρωμένα όλα τα αποτελέσματα της ερευνάς μας.



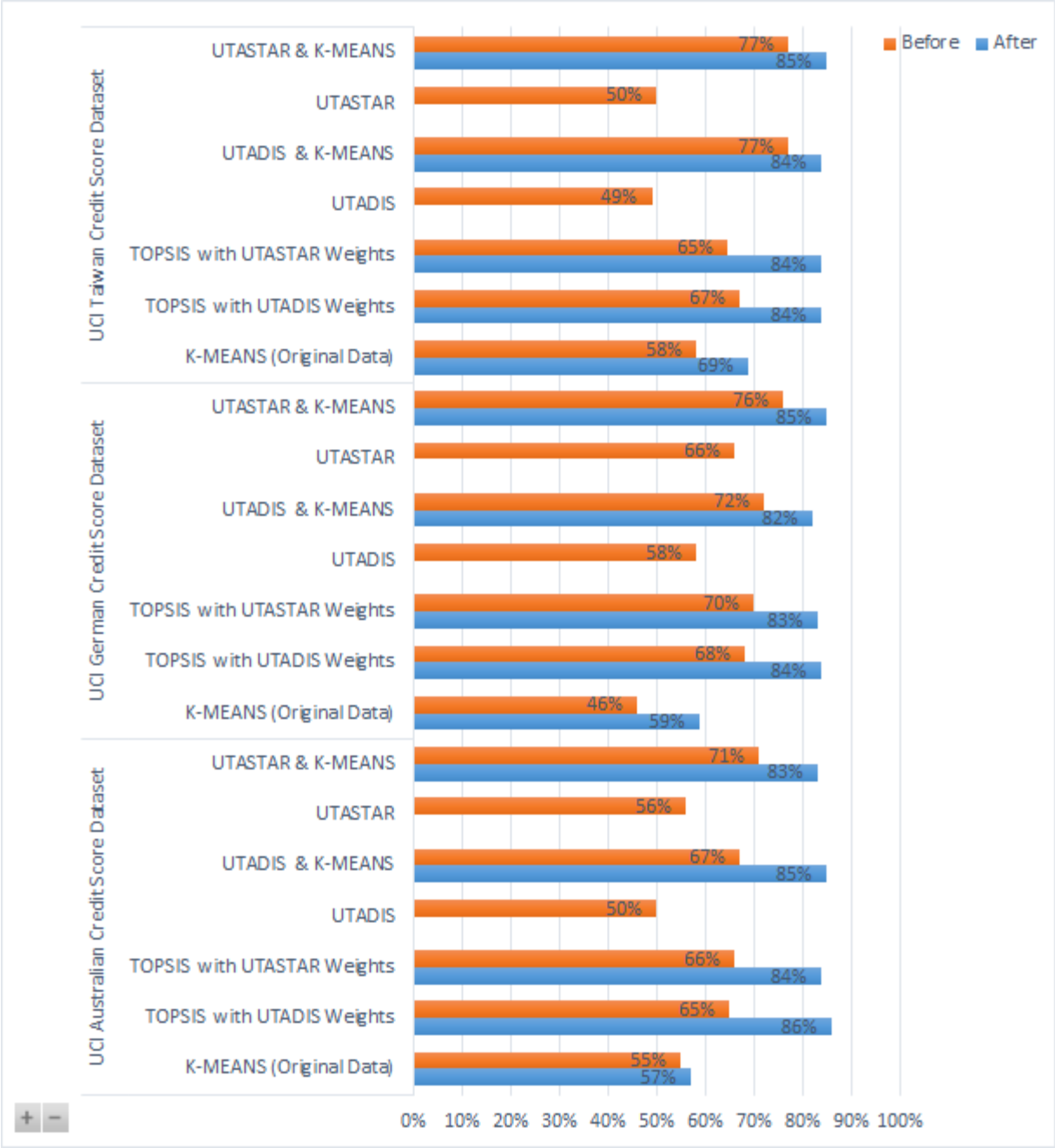
Σχήμα 5.8.1 Μέση Αύξηση Ακρίβειας Με Μείωση Διαστάσεων



Σχήμα 5.8.2 Μέση Αύξηση Ακρίβειας Ανά Σύνολο Δεδομένων



Σχήμα 5.8.3 Μέση Αύξηση Ακρίβειας Ανά Μέθοδο η Συνδυασμό Μεθόδων



Σχήμα 5.8.4 Μέση αύξηση ακρίβειας ανά σύνολο δεδομένων και συνδυασμό μέθοδών

Παρατηρήσεις:

Ως τελικά αποτελέσματα, έχουμε μια αρκετά καλή ακρίβεια ταξινόμησης σε όλες τις μεθόδους, καθώς και στα τρία σύνολα δεδομένων, φαινόμενο το οποίο καθιστά αρκετά αξιόπιστα τα αποτελέσματά μας, τα οποία και δίνουν μια καλή βάση για περαιτέρω έρευνα.

Κεφάλαιο 6 Συμπεράσματα και Μελλοντικές Προεκτάσεις

6.1 Συμπεράσματα αποτελεσμάτων

Εξερευνώντας λοιπόν μερικές από τις Πολυκριτήριες Μεθόδους Ανάλυσης Αποφάσεων και Μεθόδους Μηχανικής Μάθησης στον τομέα της αξιολόγησης χρηματο-οικονομικής πίστωσης (credit scoring), κάναμε την αρχή με θετικά αποτελέσματα για την ανάπτυξη ενός συστήματος που θα υποστηρίζει την συνδυαστική εφαρμογή τους, έτσι ώστε να βελτιστοποιήσουμε την εξόρυξη γνώσης στην Ανάλυση Μεγάλων Δεδομένων (Big Data Analysis).

Είδαμε ότι οι πολυκριτήριες μέθοδοι μάς έδωσαν εύκολα και γρήγορα μια βαθύτερη κατανόηση στα δεδομένα μας, αναθέτοντας βάρη στα κριτήρια (διαστάσεις), καθορίζοντας αυτόματα την σημαντικότητά τους και αξιολογώντας τις εναλλακτικές επιλογές με τις ολικές χρησιμότητες. Έτσι, μπορέσαμε να οριοθετήσουμε ακόμη καλύτερα την ταξινόμηση των δεδομένων με την χρήση της συσταδοποίησης και άρα να έχουμε ένα αρκετά αξιόπιστο μοντέλο ταξινόμησης πελατών.

Επιβεβαιώσαμε επίσης τα αποτελέσματά μας, συγκρίνοντας τις πολυκριτήριες μεθόδους μεταξύ τους και συνδυαστικά, στην προσπάθεια να εξαγάγουμε περισσότερες πληροφορίες και να έχουμε ακόμη καλύτερα αποτελέσματα.

6.2 Μελλοντικές προεκτάσεις

Μελλοντικά, η παρούσα διπλωματική εργασία θα μπορούσε να εμπλουτιστεί και να αξιοποιηθεί δοκιμάζοντας τους προτεινόμενους συνδυασμούς μεθόδων και σε προβλήματα εκτός της χρηματο-οικονομικής πίστωσης με ανάγκη για περισσότερες ταξινομήσεις. Επίσης, μπορεί να γίνει μια βελτιστοποίηση στις πολυκριτήριες μεθόδους που χρησιμοποιούν γραμμικό προγραμματισμό, καθώς για μεγάλα σύνολα δεδομένων έχουμε μεγάλους χρόνους ανταπόκρισης.

Επιπρόσθετα, μπορεί να εξεταστεί και η χρήση άλλων μεθόδων μηχανικής μάθησης, όπως k-nearest neighbors, global k -means ή ακόμη μεθόδων εκτός συσταδοποίησης. Τέλος, επίσης δυνατή είναι και η σύγκριση με τη χρήση μείωσης διαστάσεων, που χρησιμοποιείται σε συνδυασμό και με άλλες έρευνες ή μεθόδους.

Βιβλιογραφία

[1] Jasmina Novaković, Perica Strbac, Dusan Bulatović, “Toward Optimal Feature Selection Using Ranking Methods And Classification Algorithms”, Yugoslav Journal of Operations Research, 2011.

[2] Amir Hossein Azadnia, Pezhman Ghadimib, Mohammad Molani-Aghdama, “A Hybrid Model of Data Mining and MCDM Methods for Estimating Customer Lifetime Value”, Proceedings of the 41st International Conference on Computers and Industrial Engineering, 2011.

[3] Abdollah Nazari, Mohammadreza Mehreganb, Reza Tehranic, “Using the Hybrid Model for Credit Scoring (Case Study: Credit Clients of Microloans, Bank RefahKargaran of Zanjan, Iran)”, Journal of Optimization in Industrial Engineering, 2019.

[4] Yang Ruicheng, Guo Rongrong, Shen Qing, “Detecting Fraudulent Financial Data Using Multicriteria Decision Aid Method”, Third International Conference on Information Science and Control Engineering, 2016.

[5] Xiaoqian Zhu, Jianping Li, Dengsheng Wu, Haiyan Wang, Changzhi Liang, “Balancing accuracy, complexity and interpretability in consumer credit decision making: A C-TOPSIS classification approach”, Knowledge-Based Systems, 52, 2013, Pages 258-267.

[6] Desheng Wu, University of Toronto, Canada David L. Olson, A TOPSIS Data Mining Demonstration and Application to Credit Scoring, International Journal of Data Warehousing and Mining, 2006.

[7] Majid Esmaelian, Hadi Shahmoradi and Fateme Nemati, P-UTADIS: A Multi Criteria Classification Method, Article in Computers and Industrial Engineering, 2016.

[8] Y. Siskos, E. Grigoroudis, and N. F. Matsatsinis, “UTA Methods,” in Multiple Criteria Decision Analysis: State of the Art Surveys, Springer-Verlag, 2005, pp. 297–334.

[9] E. Jacquet-Lagreze and J. Siskos, “Assessing a set of additive utility functions for multicriteria decision-making, the UTA method,” European Journal of Operations Research, vol. 10, no. 2, pp. 151–164, June 1982.

[10] Y. Siskos and D. Yannacopoulos, “UTASTAR - an ordinal regression method for building additive value functions”, Investigaçao Operacional, 5(1), pp. 39–53, 1985.

[11] D. Despotis, D. Yannacopoulos, and C. Zopounidis, “A review of the UTA multicriteria method and some improvements”, Foundations of Computing and Decision Sciences, 15(2), 63–76 , 1990.

[12] Devaud, J.M., Groussaud, G. and Jacquet-Lagrèze, “UTADIS: A method of constructing additive utility functions accounting for global judgments”, European Working Group on Multicriteria Decision Aid, 1980.

[13] Hwang and Yoon “Technique for Order Preference by Similarity to Ideal Solution”, 4th International Conference on Process Engineering and Advanced Materials, 1981.

[14] Aristidis Likas, Nikos Vlassis, Jakob J., Verbeek, “The global k-means clustering algorithm”, Pattern Recognition, Volume 36, Issue 2, February 2003, Pages 451-461.

[15] Peter J. Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”, Journal of Computational and Applied Mathematics, 1987.

[16] Python Core Team, “Python: A dynamic, open source programming language.” Python Software Foundation, 2015.

[17] “Anaconda.” https://www.continuum.io/.

[18] “Pandas.” http://pandas.pydata.org/.

[19] “lp_solve 5.5.2.0.” http://web.mit.edu/lpsolve/doc/

[20] https://matplotlib.org/

[21] https://scikit-learn.org/stable/

[22] Ματσατσίνης Νικόλαος, Συστήματα Υποστήριξης Αποφάσεων, Εκδόσεις Νέων Τεχνολογιών, 2010.

[23] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[24] http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#

[25] http://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29

[26] https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

[27] https://archive.ics.uci.edu/ml/index.php