ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Πολιτικών Μηχανικών

Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

Ανάπτυξη Μεθοδολογίας Αξιολόγησης Έργων Έρευνας και Ανάπτυξης με Περιβάλλουσα Ανάλυση Δεδομένων και Πρότυπα Μηχανικής Μάθησης

Διπλωματική Εργασία

Ανδρέας Τσαμπούλας

Επιβλέπουσα Καθηγήτρια : Ελένη Βλαχογιάννη,

Αναπληρώτρια Καθηγήτρια Σχολής Πολιτικών Μηχανικών ΕΜΠ

Αθήνα, Ιούλιος 2020

Αφιερώνεται στον πατέρα μου.

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Πολιτικών Μηχανικών

Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής

Ανάπτυξη Μεθοδολογίας Αξιολόγησης Έργων Έρευνας και Ανάπτυξης με Περιβάλλουσα Ανάλυση Δεδομένων και Πρότυπα Μηχανικής Μάθησης

Ανδρέας Τσαμπούλας

Επιβλέπουσα Καθηγήτρια : Ελένη Βλαχογιάννη

Αθήνα, 2020.

National Technical University of Athens

School of Civil Engineering

Department of Transportation Planning and Engineering,

Development of Evaluation Methodology for Research and Development Projects Using Data Envelopment Analysis and Machine Learning

Thesis Author: Andreas Tsampoulas

Supervising Professor: Eleni Vlahogianni

Athens, 2020.

Copyright © Ανδρέας Τσαμπούλας

Με επιφύλαξη παντός δικαιώματος

Απαγορεύεται η αντιγραφή, αποθήκευση σε αρχείο πληροφοριών, διανομή, αναπαραγωγή, μετάφραση ή μετάδοση της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό, υπό οποιαδήποτε μορφή και με οποιοδήποτε μέσο επικοινωνίας, ηλεκτρονικό ή μηχανικό, χωρίς την προηγούμενη έγγραφη άδεια της συγγραφέως. Επιτρέπεται η αναπαραγωγή, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν στη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Η έγκριση της διπλωματικής εργασίας από τη Σχολή Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέως (Ν. 5343/1932, Άρθρο 202).

Copyright © Andreas Tsampoulas, 2020

All Rights Reserved

Neither the whole nor any part of this diploma thesis may be copied, stored in a retrieval system, distributed, reproduced, translated, or transmitted for commercial purposes, in any form or by any means now or hereafter known, electronic or mechanical, without the written permission from the author. Reproducing, storing and distributing this thesis for non-profitable, educational or research purposes is allowed, without prejudice to reference to its source and to inclusion of the present text. Any queries in relation to the use of the present thesis for commercial purposes must be addressed to its author.

Approval of this diploma thesis by the School of Civil Engineering of the National Technical University of Athens (NTUA) does not constitute in any way an acceptance of the views of the author contained herein by the said academic organization (L. 5343/1932, art. 202).

Ευχαριστίες

Ολοκληρώνοντας την διπλωματική μου εργασία, θα ήθελα να ευχαριστήσω θερμά την επιβλέπουσα καθηγήτρια κα Ελένη Βλαχογιάννη, Αναπληρώτρια Καθηγήτρια στον Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής της Σχολής Πολιτικών Μηχανικών, για την καθοδήγηση και τις συμβουλές που μου προσέφερε για την εκπόνηση της διπλωματικής μου εργασίας.

Θα ήθελα, επίσης, να ευχαριστήσω ολόψυχα την οικογένεια μου, τους φίλους μου και τα κοντινά μου πρόσωπα για τη συνεχή στήριξη και την έμπνευση που μου προσέφεραν.

Τίτλος: Ανάπτυξη Μεθοδολογίας Αξιολόγησης Έργων Έρευνας και Ανάπτυξης με Περιβάλλουσα Ανάλυση Δεδομένων και Πρότυπα Μηχανικής Μάθησης

Συγγραφέας Διπλωματικής Εργασίας: Ανδρέας Τσαμπούλας

Επιβλέπουσα Καθηγήτρια: Ελένη Βλαχογιάννη

Σύνοψη

Κατά την αξιολόγηση έργων Έρευνας και Ανάπτυξης (Ε&Α), η ποιότητα των αποτελεσμάτων εξαρτάται τόσο από την αναλυτική εμπειρογνωμοσύνη του οργανισμού όσο και από την ίδια τη μέθοδο αξιολόγησης. Στη παρούσα εργασία, αναπτύσσεται μια μεθοδολογία αξιολόγησης που εφαρμόζεται σε 2232 προγράμματα Ε&Α που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση τα τελευταία έξι χρόνια. Επιπλέον, προτείνονται στρατηγικές μέθοδοι για τη σταδιακή βελτιστοποίηση έργων Ε&Α που είναι ήδη σε εξέλιξη, αλλά και για τη βελτιστοποίηση μελλοντικών έργων, ανάλογα με το επίπεδο αποδοτικότητας που προβλέπεται να έχουν, από το μοντέλο. Συνδυάζοντας τη μέθοδο Περιβάλλουσας Ανάλυσης Δεδομένων με μοντέλα μηχανικής μάθησης, στο προγραμματιστικό περιβάλλον της R, αναλύονται και προτεραιοποιούνται τα προγράμματα ως προς το επίπεδο αποδοτικότητας τους. Επιπρόσθετα, αναπτύσσεται μια μεθοδολογία με βάση δένδρα ταξινόμησης και πρότυπα ομαδοποίησης για την πρόβλεψη του επιπέδου αποδοτικότητας και τον προσδιορισμό της καλύτερης διαδρομής για τη σταδιακή βελτιστοποίηση των μη αποδοτικών προγραμμάτων. Η εφαρμογή της προτεινόμενης μεθοδολογίας οδηγεί σε σημαντικά συμπεράσματα ως προς την αποδοτικότητα των έργων σε σχέση με τον αριθμό των οργανισμών και των διαφορετικών χωρών που συμμετέχουν. Τέλος, παρουσιάζονται οι προεκτάσεις που θα μπορούσε να έχει η εφαρμογή της προτεινόμενης μεθοδολογίας σε διαχειριστικό επίπεδο.

Λέξεις Κλειδιά: Ε&Α, Αξιολόγηση Ε&Α, Αποδοτικότητα, Περιβάλλουσα Ανάλυση Δεδομένων, Μηχανική μάθηση, C4.5, K-μέσων

Title: Development of Evaluation Methodology for Research and Development Projects Using Data Envelopment Analysis and Machine Learning

Thesis Author: Andreas Tsampoulas

Supervising Professor: Eleni Vlahogianni

Abstract

When evaluating Research and Development (R&D) projects, the quality of the results depends both on the analytical expertise of the organization and on the evaluation method itself. In the present work, an evaluation methodology is developed that is applied to 2232 R&D projects funded by the European Union in the last six years. In addition, strategic methods are proposed for the gradual optimization of ongoing R&D projects, but also for the optimization of future projects, depending on the predicted level of efficiency. Combining the Data Envelopment Analysis method with machine learning models, in the R programming environment, projects are analyzed and prioritized in terms of their efficiency level. Moreover, a methodology based on classification trees and clustering algorithms is developed to better predict the efficiency level and efficiency path for the gradual optimization of inefficient projects. The application of the proposed methodology leads to important conclusions regarding the efficiency of the projects in relation to the number of organizations and the different countries that are participating. Finally, the extensions that the application of the proposed methodology could have at management level are presented.

Keywords: R&D, R&D evaluation, Efficiency, Data envelopment analysis, Machine learning, C4.5, K-means

Περίληψη

Οι επενδύσεις Ε&Α είναι ένα από τα πιο κρίσιμα στοιχεία για την προώθηση της επιστημονικής και τεχνολογικής προόδου. Οι εταιρείες και οι κυβερνήσεις, που ασχολούνται με έργα έρευνας και ανάπτυξης - Ε&Α (Research and Development, R&D), αντιμετωπίζουν προβλήματα κατά την αξιολόγηση έργων Ε&Α. Τα αποτελέσματα των αξιολογήσεων, επηρεάζουν τις αποφάσεις που λαμβάνονται και άρα έχουν αντίκτυπο στην επιβίωση και την ανάπτυξη ενός τεχνολογικού οργανισμού. Κατά την αξιολόγηση έργων Ε&Α, μια σημαντική απόφαση που πρέπει να ληφθεί είναι η μέθοδος αξιολόγησης που θα χρησιμοποιηθεί. Πολύ συχνά, η ποιότητα των αποτελεσμάτων εξαρτάται τόσο από την αναλυτική εμπειρογνωμοσύνη της εταιρείας όσο και από την ίδια τη μέθοδο αξιολόγησης. Παρά τις δυσκολίες που εμφανίζονται, τα έργα πρέπει να αξιολογηθούν και να δοθούν προτεραιότητες, καθώς ανταγωνίζονται για πόρους.

Στόχος της διπλωματικής εργασίας είναι η ανάπτυξη μιας μεθοδολογίας αξιολόγησης έργων Ε&Α που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση τα τελευταία έξι χρόνια. Επιπλέον, προτείνονται στρατηγικές μέθοδοι για τη σταδιακή βελτιστοποίηση έργων Ε&Α που είναι ήδη σε εξέλιξη, αλλά και για τη βελτιστοποίηση μελλοντικών έργων, ανάλογα με το επίπεδο αποδοτικότητας που προβλέπεται να έχουν από το μοντέλο. Η μεθοδολογία που αναπτύσσεται στη παρούσα διπλωματική, προσπαθεί να ανταποκριθεί στις προκλήσεις της σημερινής εποχής, συνδυάζοντας τη μέθοδο Περιβάλλουσας Ανάλυσης Δεδομένων (Data Envelopment Analysis, DEA) με μοντέλα μηχανικής μάθησης (machine learning), στο προγραμματιστικό περιβάλλον της R.

Η DEA είναι μια τεχνική που διερευνήθηκε από τους Charnes et al. (1978) και που βασίζεται σε γραμμικό προγραμματισμό για τη μέτρηση της σχετικής απόδοσης των οργανισμών ή των μονάδων λήψης αποφάσεων (DMUs) όπου η παρουσία πολλαπλών εισόδων και εξόδων καθιστά τις συγκρίσεις δύσκολες. Η DEA αναπτύχθηκε σε επιχειρησιακές έρευνες και οικονομικές μελέτες ως μέθοδος για την αξιολόγηση της αποτελεσματικότητας των μονάδων δραστηριότητας, κάνοντας τις ελάχιστες δυνατές παραδοχές σχετικά με τη λειτουργική μορφή της υποκείμενης λειτουργίας παραγωγής. Η DEA έχει χρησιμοποιηθεί εκτενώς για την αξιολόγηση της σχετικής αποτελεσματικότητας των μονάδων δραστηριότητας μη κερδοσκοπικού χαρακτήρα (π.χ. σχολεία, νοσοκομεία) και κερδοσκοπικούς οργανισμούς (π.χ. τράπεζες, αεροπορικές εταιρείες).

Η μεθοδολογία που αναπτύχτηκε, εφαρμόστηκε σε 2232 προγράμματα Ε&Α που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση τα τελευταία έξι χρόνια (Ιανουάριος του 2014 μέχρι Δεκέμβριος του 2019). Τα δεδομένα συλλέχτηκαν και επεξεργάστηκαν από το CORDIS (Community Research and Development Information Service -

Κοινοτική Υπηρεσία Πληροφοριών Έρευνας και Ανάπτυξης). Το CORDIS είναι το κύριο δημόσιο αποθετήριο και «πύλη» της Ευρωπαϊκής Επιτροπής για τη διάδοση πληροφοριών για όλα τα ερευνητικά έργα που χρηματοδοτούνται από την ΕΕ και τα αποτελέσματά τους.

Αρχικά, εφαρμόστηκε η μέθοδος DEA στα 2232 προγράμματα E&A. Ως εισροές για το μοντέλο της DEA τέθηκαν: το συνολικό κόστος του έργου, ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο, το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο, η συνολική διάρκεια του κάθε έργου. Ως εκροές για το μοντέλο της DEA τέθηκε το άθροισμα των παραδοτέων με τις δημοσιεύσεις. Για το συγκεκριμένο μοντέλο επιλέχτηκε προσανατολισμός προς τα δεδομένα εισόδου (input-oriented) και μεταβαλλόμενη απόδοση κλίμακας (Variable Returns to Scale, VRS). Μέσα από αυτή τη μέθοδο προέκυψε ένα σύνολο προγραμμάτων που έχουν δείκτη αποδοτικότητας ίσο με 1 και άρα είναι αποδοτικά. Αυτό το σύνολο αποκαλείται Tier 1 (Επίπεδο 1). Μόνο τα 123 από τα 2232 (περίπου το 5.5%) προγράμματα προέκυψαν αποδοτικά, ενώ περίπου το 93% των προγραμμάτων έχουν δείκτη αποδοτικότητας μικρότερο ή ίσο του 0.5.

Έπειτα, εφαρμόστηκε πάλι η μέθοδος DEA μόνο με τα μη-αποδοτικά απ' όπου προέκυψε το Tier 2. Η ίδια διαδικασία επαναλήφθηκε με επαναληπτικούς βρόχους στο περιβάλλον προγραμματισμού της R μέχρι ο αριθμός των υπολειπόμενων παραγωγικών μονάδων να είναι τουλάχιστον τρεις φορές μεγαλύτερος (3x5=15) από το άθροισμα του αριθμού των διαφορετικών μεταβλητών εισροών και εκροών (4+1=5), όπως προτείνεται από τον Banker et al. (1984). Μέσα από την ανάλυση του επιπέδου αποδοτικότητας, τα 2232 προγράμματα ομαδοποιήθηκαν σε 14 ομάδες.

Στη συνέχεια, αναπτύχθηκε ένα Δένδρο Ταξινόμησης, με μάθηση με επίβλεψη, εισάγοντας τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας ως δεδομένα εισόδου. Το Δένδρο Ταξινόμησης που αναπτύχθηκε αποδείχτηκε να είναι ένα πολύ αξιόπιστο εργαλείο για τη πρόβλεψη του επιπέδου αποδοτικότητας των έργων E&A, το οποίο μπορεί να αξιοποιηθεί από όλους τους οργανισμούς του τομέα κατά τη λήψη των αποφάσεων, την αξιολόγηση και την επιλογή έργων.

Μετέπειτα, πραγματοποιήθηκε η ομαδοποίηση των προγραμμάτων, χρησιμοποιώντας τον αλγόριθμο K-μέσων, με μάθηση χωρίς επίβλεψη, χωρίζοντας τα προγράμματα σε 4 διαφορετικές ομάδες, ανάλογα με τα χαρακτηριστικά του κάθε προγράμματος.

Τέλος, κάθε μη αποδοτικό πρόγραμμα αντιστοιχίστηκε, με βάση τη μικρότερη Ευκλείδεια απόσταση, με ένα πρόγραμμα που βρίσκεται στο αμέσως μεγαλύτερο Tier του ίδιου cluster, που μοιράζεται, δηλαδή, παρόμοια χαρακτηριστικά. Έτσι, προσδιορίστηκε η καλύτερη διαδρομή για τη σταδιακή βελτιστοποίηση των προγραμμάτων κάθε επιπέδου εκτός του επιπέδου 1.

Το βασικό συμπέρασμα που προκύπτει από τα αποτελέσματα της ανάλυσης των επιπέδων αποδοτικότητας των προγραμμάτων είναι ότι η συμμετοχή πολλών διαφορετικών χωρών αλλά και οργανισμών σε ένα έργο επηρεάζει αρνητικά την αποδοτικότητα του έργου. Αυτό οφείλεται στις δυσκολίες που πιθανότατα προκύπτουν κατά τον συντονισμό και την επικοινωνία πολλών διαφορετικών χωρών και οργανισμών.

Το εργαλείο που αναπτύχθηκε μπορεί να αξιοποιηθεί άμεσα από την Ευρωπαϊκή Ένωση στη διαχείριση των έργων Ε&Α για τη πρόβλεψη του επιπέδου αποδοτικότητας που έπεται να έχουν νέα έργα Ε&Α ανάλογα με τους πόρους που διατίθενται και για την ανάπτυξη στρατηγικών μεθόδων βελτίωσης της αποδοτικότητας έργων που είτε είναι σε εξέλιξη, είτε που δεν έχουν ξεκινήσει ακόμα.

Τα παρόν μοντέλο, λόγω των περιορισμένων δεδομένων που διατίθενται, περιλαμβάνει μόνο ένα μικρό σύνολο εισροών που έχουν αντίκτυπο στα προϊόντα ή τις υπηρεσίες που παράγονται. Άλλοι παράγοντες που θα έπρεπε να ληφθούν υπόψη είναι η πολυπλοκότητα του έργου, η ποιότητα των διαθέσιμων υλικών και λογισμικών εργαλείων, το πλήθος, η εμπειρία του προσωπικού κλπ.. Επίσης, θα ήταν ενδιαφέρουσα η προσέγγιση της μεθοδολογίας με προσανατολισμό προς τις εκροές, δηλαδή με σκοπό την ανάπτυξη στρατηγικών μεθόδων για τη μεγιστοποίηση των εκροών, διατηρώντας σταθερές τις εισροές. Τέλος, θα ήταν σκόπιμη μια μετα-ανάλυση (meta-analysis), μέσω της οποίας θα διερευνηθούν τα ακόλουθα: α) η συσχέτιση μεταξύ των εισροών των έργων Ε&Α, β) οι διαφοροποιήσεις που υπάρχουν στα επίπεδα αποδοτικότητας ανάλογα το τομέα που βρίσκονται και γ) η ανάπτυξη καινούριων μοντέλων DEA για κάθε διαφορετικό τομέα.

Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Ευρετήριο Διαγραμμάτων

1. Εισαγωγή

1.1.Γενική Ανασκόπηση

Η δραστηριότητα Έρευνας και Ανάπτυξης (Ε&Α) είναι μια καλά οργανωμένη διαδικασία δημιουργίας, παραγωγής, διάδοσης και εφαρμογής της γνώσης. Περιλαμβάνει καινοτομία στην επιστημονική τεχνολογία, στα μέτρα διαχείρισης και στα κοινωνικά και πολιτικά συστήματα κ.λπ. Ο Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ, 2003) όρισε την επένδυση στη γνώση ως το άθροισμα των δαπανών Ε&Α, των δαπανών για την τριτοβάθμια εκπαίδευση και των επενδύσεων σε λογισμικό υπολογιστών. Δεδομένου ότι οι επενδύσεις Ε&Α είναι ένα από τα πιο κρίσιμα στοιχεία για την προώθηση της επιστημονικής και τεχνολογικής προόδου, κάθε χώρα που χρησιμοποιεί τους πόρους αναποτελεσματικά, επιτυγχάνει πιο αργή πρόοδο. Επιπλέον, εάν οι πόροι Ε&Α δεν χρησιμοποιούνται αποτελεσματικά, οι πρόσθετες επενδύσεις προσφέρουν μικρή βοήθεια στην τόνωση της προόδου. Ωστόσο, η σχετική βιβλιογραφία έχει επικεντρωθεί κυρίως στις προσπάθειες για νέες επενδύσεις Ε&Α και συγκριτικά λίγη προσοχή έχει δοθεί στην αποτελεσματική χρήση των πόρων μόλις τεθούν σε εφαρμογή. Αυτή είναι μια δυνητικά σημαντική παράλειψη, καθώς η ανεπαρκής διαχείριση των δραστηριοτήτων Ε&Α ενδέχεται να ευθύνεται για την επιστημονική και οικονομική καθυστέρηση. Η κατανόηση της φύσης της αποδοτικότητας και της αναποτελεσματικότητας της Ε&Α είναι σημαντική για τον σχεδιασμό τακτικών για τη βελτίωση της κατανομής των πόρων.

Οι νέες τεχνολογίες και τα επιχειρηματικά μοντέλα, που αναπτύσσονται, έχουν μεγάλη επίδραση στον τρόπο δημιουργίας και παράδοσης προϊόντων και υπηρεσιών, ιδίως όσον αφορά την καινοτομία, τα συστήματα, τις τεχνολογίες, τη διαχείριση και τις μεθόδους παράδοσης. Η νέα οικονομία της πληροφορίας αναγκάζει τις εταιρείες υψηλής τεχνολογίας να ανταγωνίζονται μεταξύ τους μέσω της αύξησης της απόδοσης Ε&Α και της μείωσης του κόστους ταυτόχρονα. Μια αποτελεσματική λειτουργία Ε&Α είναι μια σημαντική πηγή ανταγωνιστικού πλεονεκτήματος στη σημερινή ταχέως παγκοσμιοποιημένη οικονομία.

Έτσι, η αξιολόγηση της απόδοσης της έρευνας και ανάπτυξης (Ε&Α) είναι ένα σημαντικό πρόβλημα τόσο ακαδημαϊκού ενδιαφέροντος όσο και πρακτικής ανάγκης.

1.2.Στόχοι και Μεθοδολογία

Το πρόβλημα αξιολόγησης έργων Ε&Α είναι ένα δύσκολο πρόβλημα λήψης αποφάσεων που αντιμετωπίζουν οι υπεύθυνοι που ασχολούνται με τη διαχείριση έργων Ε&Α. Η αξιολόγηση περιλαμβάνει πολλαπλά κριτήρια που μετρούν τις ανταμοιβές, τη συνάφεια με την αποστολή και τους στόχους του οργανισμού, τη δυνατότητα στρατηγικής μόχλευσης, την πιθανότητα τεχνικής και εμπορικής επιτυχίας κ.λπ.. Η εστίαση σε μελλοντικά γεγονότα και ευκαιρίες σε ένα δυναμικό περιβάλλον προκαλεί πολλές από τις απαιτούμενες πληροφορίες να είναι στην καλύτερη περίπτωση αβέβαιες και στη χειρότερη μη διαθέσιμες. Οι γνώμες και οι κρίσεις, συχνά, πρέπει να αντικαθιστούν τα δεδομένα και τα μέτρα θα μπορούσαν να εκτιμηθούν μόνο ποιοτικά. Η έλλειψη αξιόπιστων ποσοτικών μέτρων είναι ιδιαίτερα εμφανής σε μη κερδοσκοπικούς οργανισμούς, όπου τα ποιοτικά μέτρα συνήθως έχουν μεγαλύτερο μερίδιο στη συνολική αξιολόγηση. Παρά τις δυσκολίες αυτές, τα έργα πρέπει να αξιολογηθούν και να δοθούν προτεραιότητες, καθώς ανταγωνίζονται για πόρους.

Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη μιας μεθοδολογίας αξιολόγησης έργων Ε&Α που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση τα τελευταία χρόνια. Επιπλέον, προτείνονται στρατηγικές μέθοδοι για τη σταδιακή βελτιστοποίηση έργων Ε&Α που είναι ήδη σε εξέλιξη, αλλά και για τη βελτιστοποίηση μελλοντικών έργων, ανάλογα με το επίπεδο αποδοτικότητας που προβλέπεται να έχουν από το μοντέλο.

Η μεθοδολογία που αναπτύσσεται στη παρούσα διπλωματική, προσπαθεί να ανταποκριθεί στις προκλήσεις της σημερινής εποχής, συνδυάζοντας τη μέθοδο Περιβάλλουσας Ανάλυσης Δεδομένων (Data Envelopment Analysis, DEA) με μοντέλα μηχανικής μάθησης (machine learning), στο προγραμματιστικό περιβάλλον της R.

Η DEA είναι ένα πολύ-παραγοντικό μοντέλο ανάλυσης παραγωγικότητας για τη μέτρηση της σχετικής απόδοσης κάθε Μονάδας Λήψης Αποφάσεων, όταν υπάρχουν πολλαπλές είσοδοι και έξοδοι. Ένας τυπικός τρόπος μέτρησης της απόδοσης DEA είναι η εκτίμηση του μοντέλου που περιλαμβάνει όλες τις εισόδους και εξόδους. Στη παρούσα διπλωματική, ως εισροές για το μοντέλο της DEA τέθηκαν: το συνολικό κόστος του έργου, ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο, το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο και η συνολική διάρκεια του κάθε έργου (Duration). Ως εκροές για το μοντέλο της DEA τέθηκε το άθροισμα των παραδοτέων με τις δημοσιεύσεις. Η μέθοδος DEA εφαρμόστηκε με σκοπό να προσδιοριστεί το επίπεδο αποδοτικότητας στο οποίο ανήκει κάθε έργο, με εστίαση στην ελαχιστοποίηση των δεδομένων εισόδου (input-oriented).

Έπειτα, χρησιμοποιήθηκαν μοντέλα μηχανικής μάθησης με σκοπό την ταξινόμηση (classification) και την ομαδοποίηση (clustering) των δεδομένων. Αναπτύχθηκε ένα

Δένδρο Ταξινόμησης, με μάθηση με επίβλεψη, εισάγοντας τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας ως δεδομένα εισόδου, το οποίο θα έχει τη δυνατότητα να προβλέπει σε ποιο επίπεδο αποδοτικότητας βρίσκεται οποιοδήποτε υπάρχον ή νέο πρόγραμμα, με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών). Επίσης, πραγματοποιήθηκε η ομαδοποίηση των δεδομένων, ανάλογα με τα χαρακτηριστικά του κάθε προγράμματος. Τέλος, συνδυάστηκαν τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας και τα αποτελέσματα της ομαδοποίησης με σκοπό να προσδιοριστεί η καλύτερη διαδρομή για τη σταδιακή βελτιστοποίηση (improvement path) των μη αποδοτικών προγραμμάτων.

Η μεθοδολογία που αναπτύχτηκε, εφαρμόστηκε σε 2232 προγράμματα Ε&Α, που συλλέχτηκαν και επεξεργάστηκαν από το CORDIS (Community Research and Development Information Service - Κοινοτική Υπηρεσία Πληροφοριών Έρευνας και Ανάπτυξης). Το CORDIS δημιουργήθηκε το 1990 και διοικείται από την Υπηρεσία Εκδόσεων της Ευρωπαϊκής Ένωσης . Είναι το κύριο δημόσιο αποθετήριο και «πύλη» της Ευρωπαϊκής Επιτροπής για τη διάδοση πληροφοριών για όλα τα ερευνητικά έργα που χρηματοδοτούνται από την Ε. Ε. και τα αποτελέσματά τους. Αποστολή του CORDIS είναι η διάδοση και η εκμετάλλευση των ερευνητικών αποτελεσμάτων από τους επαγγελματίες του χώρου, με σκοπό τη προώθηση της ανοιχτής επιστήμης, τη δημιουργία καινοτόμων προϊόντων και υπηρεσιών και την τόνωση της ανάπτυξης σε ολόκληρη την Ευρώπη.

1.3.Δομή Διπλωματικής Εργασίας

Παρακάτω παρουσιάζεται συνοπτικά η διάταξη και το περιεχόμενο των κεφαλαίων της διπλωματικής εργασίας.

Στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στο αντικείμενο της εργασίας. Περιγράφεται συνοπτικά η σημασία των έργων Ε&Α, η σημασία της επαρκής αξιολόγησής τους. Στη συνέχεια, αναφέρονται τα προβλήματα που προκύπτουν κατά την αξιολόγηση των έργων Ε&Α και εισάγεται η έννοια της Περιβάλλουσας Ανάλυσης Δεδομένων. Τέλος, παρατίθενται τα εργαλεία που χρησιμοποιήθηκαν και η μεθοδολογία που αναπτύχθηκε.

Στο δεύτερο κεφάλαιο πραγματοποιείται η βιβλιογραφική ανασκόπηση, όπου αναλύονται, ταξινομούνται και συγκρίνονται οι διαφορετικές μέθοδοι αξιολόγησης έργων Ε&Α. Στην συνέχεια, ακολουθεί μία βιβλιογραφική ανασκόπηση για την εφαρμογή της Περιβάλλουσας Ανάλυσης Δεδομένων σε διάφορους τομείς και συγκεκριμένα στην αξιολόγηση των έργων Ε&Α και αιτιολογείται η καταλληλόλητα της.

Στο τρίτο κεφάλαιο περιγράφεται διεξοδικά η μεθοδολογία που αναπτύχθηκε. Επιπλέον, αναλύεται το θεωρητικό υπόβαθρο της Περιβάλλουσας Ανάλυσης Δεδομένων, του μοντέλου ταξινόμησης και του μοντέλου ομαδοποίησης. Τέλος παρατίθενται τα εργαλεία που χρησιμοποιήθηκαν για την εκπόνηση της εργασίας.

Στο τέταρτο κεφάλαιο αναλύεται η διαδικασία συλλογής και επεξεργασίας των στοιχείων πάνω στο οποίο εφαρμόστηκε η μεθοδολογία και δίνεται η γενική εικόνα των δεδομένων μέσα από την περιγραφική στατιστική ανάλυση των μεταβλητών.

Στο πέμπτο κεφάλαιο παρουσιάζονται και απεικονίζονται τα αποτελέσματα της μεθοδολογίας, ενώ γίνεται και σύντομη αναφορά της σημασίας της μεθοδολογίας στη διαχείριση των έργων.

Στο έκτο κεφάλαιο περιγράφεται συνοπτικά η μεθοδολογίας και τα αποτελέσματα. Στη συνέχεια, παρουσιάζονται τα συμπεράσματα που προκύπτουν από τα αποτελέσματα. Τέλος, γίνονται προτάσεις για περαιτέρω έρευνα που θα συμπλήρωναν τη παρούσα εργασία.

Στο έβδομο κεφάλαιο παρατίθεται η πλήρης βιβλιογραφία που χρησιμοποιήθηκε για την πραγματοποίηση της διπλωματικής εργασίας.

2. Βιβλιογραφική Ανασκόπηση

2.1. Μέθοδοι αξιολόγησης έργων E&A

Οι εταιρείες και οι κυβερνήσεις, που ασχολούνται με έργα έρευνας και ανάπτυξης - E&A (Research and Development, R&D), αντιμετωπίζουν προβλήματα κατά την αξιολόγηση έργων E&A. Οι αξιολογήσεις επικεντρώνονται σε εναλλακτικά έργα E&A που μπορεί να αναλάβει η εταιρεία ή η κυβέρνηση. Τα αποτελέσματα των αξιολογήσεων, επηρεάζουν τις αποφάσεις που λαμβάνονται και άρα έχουν αντίκτυπο στην επιβίωση και την ανάπτυξη ενός τεχνολογικού οργανισμού. Κατά την αξιολόγηση έργων E&A, μια σημαντική απόφαση που πρέπει να ληφθεί είναι η μέθοδος αξιολόγησης που θα χρησιμοποιηθεί. Πολύ συχνά, η ποιότητα των αποτελεσμάτων εξαρτάται τόσο από την αναλυτική εμπειρογνωμοσύνη της εταιρείας όσο και από την ίδια τη μέθοδο αξιολόγησης.

Τις τελευταίες δεκαετίες, ένα ευρύ φάσμα μεθόδων αξιολόγησης των έργων E&A έχει αναπτυχθεί και αναφερθεί στη βιβλιογραφία. Πλήρεις ανασκοπήσεις αυτών των μεθόδων αξιολόγησης E&A έχουν διεξαχθεί από τους Souder (1978), Danila (1980) και Jackson (1983).

Γενικά, οι μέθοδοι αξιολόγησης έργων E&A μπορούν να ταξινομηθούν σε δύο κύριες κατηγορίες:

- Μέθοδοι στάθμισης και κατάταξης (weighting and ranking methods)
- Μέθοδοι συνεισφοράς παροχών (benefit-contribution methods)

2.1.1. Μέθοδοι στάθμισης και κατάταξης

Οι μέθοδοι στάθμισης και κατάταξης (weighting and ranking methods) υπολογίζουν τα σχετικά βάρη και ταξινομούν ένα σύνολο προτεινόμενων έργων κατά σειρά προτίμησης. Οι πιο συνηθισμένοι τύποι μεθόδων στάθμισης και κατάταξης είναι η συγκριτική μέθοδος (Comparative method) (Pessemier and Baker, 1971, Easton, 1973, Ormala, 1986), μέθοδος βαθμολόγησης (Scoring Method) (Bradbury et al., 1973, Bedell, 1983, Krawiec, 1984, Pinto and Slevin, 1989, Balachandra and Brockhoff, 1996), και η διαδικασία αναλυτικής ιεραρχίας (Analytical Hierarchy Process, AHP) (Saaty, 1980, Lockett et al., 1986, Liberatore, 1987).

Στη συγκριτική μέθοδο, ένα έργο συγκρίνεται με ένα άλλο έργο ή ένα σύνολο εναλλακτικών έργων. Χρησιμοποιούνται μαθηματικά μοντέλα για να υπολογίσουν ρητά τη συνολική αξία κάθε έργου που οδηγεί στον προσδιορισμό του καλύτερου έργου. Η συγκριτική μέθοδος είναι εύχρηστη και κατανοητή, αλλά βασίζεται υποκειμενικές

κρίσεις και έχει ορισμένα μειονεκτήματα. Πρώτον, οι αξιολογήσεις μπορεί να είναι ασταθείς, καθώς ενδέχεται να αλλάξουν με το χρόνο και οι αξιολογήσεις που πραγματοποιούνται από διαφορετικά άτομα δεν είναι άμεσα συγκρίσιμες. Δεύτερον, οι αλλαγές στο σύνολο των εναλλακτικών έργων, ενδέχεται να επηρεάσουν την κατάταξη όλων των έργων. Τέλος, ορισμένες συγκριτικές μέθοδοι δεν λαμβάνουν υπόψη πολλαπλούς στόχους. Αφήνουν όλες τις δυσκολίες που σχετίζονται με τη συγκέντρωση των πολλαπλών στόχων στους υπεύθυνους λήψης αποφάσεων (Ormala, 1986).

Η μέθοδος βαθμολόγησης αξιολογεί τα έργα δίνοντας σε κάθε έργο μια βαθμολογία που αντικατοπτρίζει πόσο καλά ανταποκρίνεται στους καθορισμένους στόχους σε ορισμένες κλίμακες. Ένα μοντέλο βαθμολόγησης είναι ένας μαθηματικός τύπος ή αλγεβρική έκφραση που παράγει μια βαθμολογία για κάθε υπό εξέταση έργο. Η έκφραση ενσωματώνει αυτούς τους παράγοντες που θεωρούνται σημαντικοί. Κάθε παράγοντας σταθμίζεται για να αντικατοπτρίζει τη σημασία του σε σχέση με άλλους παράγοντες. Στη συνέχεια, τα έργα ταξινομούνται κατά σειρά των βαθμολογιών τους. Η μέθοδος βαθμολόγησης μπορεί να ενσωματώσει όλα τα είδη δεδομένων που μπορεί να είναι διαθέσιμα σχετικά με τα υποψήφια έργα. Ο Krawiec (1984) ανέφερε ότι η μέθοδος βαθμολόγησης είναι ένα αποτελεσματικό αναλυτικό εργαλείο όταν οι απαιτήσεις για δεδομένα και η πολυπλοκότητα πιο εξελιγμένων προσεγγίσεων δεν δικαιολογούνται. Το κύριο μειονέκτημα της μεθόδου βαθμολόγησης είναι ότι η δομή της μεθόδου συνήθως δεν είναι καλά καθορισμένη, καθιστώντας δύσκολη την αιτιολόγηση της χρήσης της (Jackson, 1983).

Η διαδικασία αναλυτικής ιεραρχίας (Analytical Hierarchy Process, AHP), η οποία αναπτύχθηκε από την Saaty (1980), είναι μια μέθοδος για τη σύγκριση ενός συνόλου εναλλακτικών λύσεων που βοηθούν στη διαδικασία λήψης αποφάσεων σε ένα περίπλοκο περιβάλλον. Είναι μια εύκολη μέθοδος για τη διατύπωση και την ανάλυση σύνθετων αποφάσεων. Το AHP διαμορφώνει ένα πολύπλοκο πρόβλημα σε μια ιεραρχία. Τα κριτήρια και οι σχετικοί παράγοντες αποσυντίθενται ιεραρχικά ανάλογα με την κατάσταση. Τα ιεραρχικά επίπεδα συνήθως αποτελούνται από τον συνολικό στόχο στην κορυφή της ιεραρχίας, ακολουθούμενο από τα κριτήρια που συμβάλλουν στον στόχο, τα υπό-κριτήρια (εάν υπάρχουν) και τέλος τις εναλλακτικές στο χαμηλότερο επίπεδο. Έπειτα, εκτελείται μια σειρά συγκρίσεων κατά ζεύγη που παράγουν τοπικά βάρη σε κάθε επίπεδο της ιεραρχίας. Στη συνέχεια, αυτά τα τοπικά βάρη συνδυάζονται χρησιμοποιώντας ένα μοντέλο προστιθέμενης αξίας για να παράγουν ένα σύνολο καθολικών βαρών ή προτεραιοτήτων για τις εναλλακτικές λύσεις. Οι εναλλακτικές μπορούν να ταξινομηθούν με βάση τα καθολικά τους βάρη που υπολογίζονται από το AHP.

2.1.2. Μέθοδοι συνεισφοράς παροχών

Οι μέθοδοι συνεισφοράς παροχών (benefit-contribution methods) χρησιμοποιούνται για την εξέταση έργων, με σκοπό τον προσδιορισμό του βαθμού ικανοποίησης των βασικών στόχων E&A ενός οργανισμού. Οι πιο συνηθισμένοι τύποι μεθόδων συνεισφοράς παροχών είναι η οικονομική ανάλυση (economic analysis) (Freeman, 1982, Ellis, 1984, Irvine, 1988, Fahrni and Spatig, 1990, Graves and Ringuest, 1991), η ανάλυση κόστους/οφέλους (cost/benefit analysis) (Augood, 1975, Kuwahara and Takeda, 1990, Link, 1993), και η ανάλυση δένδρων αποφάσεων (decision tree analysis) (Savage, 1954, Gear, 1974, Thomas, 1985, Morris et al., 1991, Faulkner, 1996).

Η οικονομική ανάλυση βασίζεται σε τεχνικές προϋπολογισμού κεφαλαίου. Τα πιο συχνά οικονομικά κριτήρια είναι η καθαρή παρούσα αξία, η περίοδος αποπληρωμής και το ποσοστό απόδοσης της επένδυσης. Η χρήση της οικονομικής ανάλυσης είναι θεωρητικά καλά αιτιολογημένη εάν μπορούν να ικανοποιηθούν οι αυστηροί όροι για τους οποίους ισχύουν τα μοντέλα (Ormala, 1986). Ωστόσο, στην πράξη, οι συνεισφορές των έργων E&A είναι δύσκολο να μετρηθούν και να διαχωριστούν από αυτές άλλων δραστηριοτήτων. Επιπλέον, είναι συχνά απαραίτητο να εκτιμηθούν τα άμεσα οικονομικά οφέλη ή οι ταμειακές ροές των έργων σε έναν μακρύ ορίζοντα προγραμματισμού. Όμως, τα ακριβή δεδομένα εισόδου που απαιτούνται από τις μεθόδους, σε νομισματικούς όρους, είναι δύσκολο να εκτιμηθούν. Ένα άλλο πρόβλημα της οικονομική ανάλυση είναι ότι οι αποφάσεις E&A βασίζονται σε πολλαπλά κριτήρια, αλλά οι οικονομικές μέθοδοι λαμβάνουν υπόψη μόνο ένα μόνο κριτήριο, δηλαδή την οικονομική απόδοση. Έχει επίσης παρατηρηθεί ότι η μεγάλη εξάρτηση από τα μέτρα οικονομικής επιστροφής μπορεί να οδηγήσει σε ένα μη ισορροπημένο χαρτοφυλάκιο προσπαθειών βελτίωσης προϊόντων και διαδικασιών (Liberatore, 1987).

Η ανάλυση κόστους/οφέλους είναι μια άλλη μέθοδος που χρησιμοποιείται συχνά για την αξιολόγηση έργων και τον προϋπολογισμό κεφαλαίου (Porter et al., 1980). Σε αυτήν τη μέθοδο, η αποτελεσματικότητα ενός έργου περιγράφεται σε μέτρα κόστους και οφέλους. Η κύρια αξία αυτού του τύπου ανάλυσης είναι ότι ενθαρρύνει τη ρητή εξέταση του κόστους και των οφελών των έργων και επιτρέπει τον εντοπισμό κρίσιμων παραγόντων στην αξιολόγηση (Ormala, 1986). Το κύριο μειονέκτημά του είναι ότι απαιτεί διαφορετικά οφέλη ή κόστη που μετρούνται στην ίδια μονάδα. Επίσης, δεν επιτρέπει την άμεση σύγκριση των έργων ή με κάποιο γενικό μέτρο αποδοχής. Τέλος, αυτή η μέθοδος δεν υποστηρίζει τη συγκέντρωση κόστους και οφέλους σε ένα μόνο μέτρο.

Η ανάλυση δέντρων αποφάσεων χρησιμοποιείται σε καταστάσεις στις οποίες οι υπεύθυνοι λήψης αποφάσεων αντιμετωπίζουν μια ακολουθία αποφάσεων, και μεταξύ κάθε δύο διαδοχικών αποφάσεων, παρεμβαίνει ένα αποτέλεσμα της προηγούμενης

απόφασης (Martino, 1995). Η ανάλυση δένδρων αποφάσεων έχει τις ρίζες της στη θεωρία της κανονιστικής απόφασης (Savage, 1954) και είναι μια καθιερωμένη μεθοδολογία για την ανάλυση αποφάσεων (Howard and Metheson, 1993, Clemen, 1996). Περιλαμβάνει τη διάρθρωση του προβλήματος, απαριθμώντας όλα τα πιθανά παρεμβατικά και τελικά επακόλουθα αποτελέσματα, και εφαρμόζει την αρχή της μέγιστης αναμενόμενης χρησιμότητας, για τον προσδιορισμό της καλύτερης εναλλακτικής λύσης του έργου.

Το κύριο πλεονέκτημα της εφαρμογής της ανάλυσης δέντρων αποφάσεων είναι ότι είναι δυνατή η αναπαράσταση και η ανάλυση μιας σειράς διαδοχικών αποφάσεων που πρέπει να λαμβάνονται με την πάροδο του χρόνου, και αυτό είναι ένα πολύ τυπικό χαρακτηριστικό της αξιολόγησης των έργων E&A (Jackson, 1983). Ωστόσο, η κατασκευή του δέντρου αποφάσεων δεν είναι μόνο χρονοβόρα, αλλά μπορεί επίσης να είναι εξαιρετικά ακατάστατη όταν το πρόβλημα είναι μεγάλο και περίπλοκο (Raiffa, 1968). Τα διαγράμματα επιρροής έχουν αναπτυχθεί ως εναλλακτική λύση για το δέντρο αποφάσεων στην ανάλυση αποφάσεων (Howard and Metheson, 1981). Τα διαγράμματα επιρροής είναι γραφικές αναπαραστάσεις των προβλημάτων απόφασης όπου οι κόμβοι στα κατευθυνόμενα άκυκλα γραφήματα υποδηλώνουν τις μεταβλητές που προκαλούν ανησυχία και τα κατευθυνόμενα τόξα αναπαριστούν τις σχέσεις μεταξύ των μεταβλητών. Επομένως, ένα διάγραμμα επιρροής αυξάνεται γραμμικά σε μέγεθος ανάλογα με τον αριθμό των μεταβλητών ενώ ένα δέντρο αποφάσεων μεγαλώνει εκθετικά. Ένα άλλο πλεονέκτημα του διαγράμματος επιρροής έναντι του δέντρου αποφάσεων είναι ότι οι πιθανότητες εξάρτησης μεταξύ βασικών μεταβλητών αντιπροσωπεύονται ρητά είτε από την παρουσία είτε από την απουσία τόξων μεταξύ των κόμβων. Σε ένα δέντρο αποφάσεων, όλες οι σχέσεις ανεξαρτησίας μπορούν να αποκαλυφθούν μόνο μέσω αριθμητικών υπολογισμών και συγκρίσεων των πιθανοτήτων.

Ένα άλλο πρόβλημα με τη χρήση των δέντρων αποφάσεων στην αξιολόγηση E&A είναι ότι όπως όλες οι θεωρητικές μέθοδοι αποφάσεων, ο υπεύθυνος λήψης αποφάσεων οφείλει να εκχωρήσει πιθανότητες σε αβέβαιες μεταβλητές και προτιμήσεις σε επακόλουθα αποτελέσματα. Οι ψυχολόγοι έχουν δείξει ότι οι άνθρωποι είναι «κακοί» εκτιμητές πιθανοτήτων (Tversky and Kahneman, 1974). Για να ξεπεραστεί το πρόβλημα έχουν αναπτυχθεί τυποποιημένα πρωτόκολλα για την πρόβλεψη πιθανοτήτων από εμπειρογνώμονες (Morganand Henrion, 1990).

2.1.3. Σύγκριση Μεθοδολογιών

Οι K.L. Poh, B.W. Ang και F. Bai (2002) έκαναν μια συγκριτική μελέτη των τεχνικών αξιολόγησης έργων E&A. Η μελέτη τους έδειξε ότι η μέθοδος βαθμολόγησης είναι η πιο συμφέρουσα μέθοδος για την αξιολόγηση έργων E&A. Πιο συγκεκριμένα, τα αποτελέσματα της μελέτης τους φανερώνουν πως οι μέθοδοι βαθμολόγησης έχουν τη

δυνατότητα να αντιμετωπίζουν προβλήματα έργων E&A πολλαπλών διαστάσεων, ενώ παράλληλα είναι εύκολοι στη χρήση και στην κατανόηση τους.

Στο Πίνακα 2.1.1 συνοψίζονται οι μεθοδολογίες που περιγράφηκαν.

Πίνακας 2.1.1: Σύγκριση Μεθοδολογιών

Σύγκριση Μεθοδολογιών	Μέθοδοι Αξιολόγησης Έργων E&A	Σύντομη Περιγραφή μεθόδου	Συγγραφείς
Μέθοδοι Στάθμισης και Κατάταξης (Weighting and Ranking Methods)	**Συγκριτική Μέθοδος (Comparative Method)**	Ένα έργο συγκρίνεται με ένα άλλο έργο ή ένα σύνολο εναλλακτικών έργων με μαθηματικά μοντέλα	Pessemier and Baker, 1971
			Easton, 1973
			Ormala, 1986
	Μέθοδος Βαθμολόγησης (Scoring Method)	Σε κάθε έργο δίνεται ένας βαθμός ανάλογα με την ανταπόκρισή του στους καθορισμένους στόχους	Bradbury et al., 1973
			Bedell, 1983
			Krawiec, 1984
			Pinto and Slevin, 1989
			Balachandra and Brockhoff, 1996
	Διαδικασία Αναλυτικής Ιεραρχίας (Analytical Hierarchy Process, AHP)	Ένα πολύπλοκο πρόβλημα διαμορφώνεται σε μια ιεραρχία	Saaty, 1980
			Lockett et al., 1986
			Liberatore, 1987
Μέθοδοι Συνεισφοράς Παροχών (Benefit-Contribution Methods)	**Οικονομική Ανάλυση (Economic Analysis)**	Βασίζεται σε τεχνικές προϋπολογισμού κεφαλαίου (όπως η καθαρή παρούσα αξία, η περίοδος αποπληρωμής και το ποσοστό απόδοσης της επένδυσης)	Freeman, 1982
			Ellis, 1984
			Irvine, 1988
			Fahrni and Spatig, 1990
			Graves and Ringuest, 1991
	Ανάλυση Κόστους/Οφέλους (Cost/Benefit Analysis)	Περιγράφεται η αποτελεσματικότητα ενός έργου σε μέτρα κόστους και οφέλους	Augood, 1975
			Kuwahara and Takeda, 1990
			Link, 1993
	Ανάλυση Δένδρων Αποφάσεων (Decision Tree Analysis)	Αναπαρηστάται και αναλύεται μια σειρά διαδοχικών αποφάσεων που πρέπει να λαμβάνονται με την πάροδο του χρόνου	Gear, 1974
			Thomas, 1985
			Morris et al., 1991
			Faulkner, 1996

2.2.Μέθοδος Περιβάλλουσας Ανάλυσης Δεδομένων (DEA)

Η μέθοδος Περιβάλλουσας Ανάλυσης Δεδομένων (Data Envelopment Analysis, DEA) είναι μια εναλλακτική προσέγγιση της μεθόδου βαθμολόγησης, αφού και οι δύο μέθοδοι αποδίδουν μια βαθμολογία σε ένα σύνολο ομότιμων οντοτήτων που παράγουν έργο, ανάλογα με την αποτελεσματικότητα τους. Πιο συγκεκριμένα, η DEA είναι μια τεχνική που διερευνήθηκε από τους Charnes et al. (1978) και που βασίζεται σε γραμμικό προγραμματισμό για τη μέτρηση της σχετικής απόδοσης των οργανισμών ή των μονάδων λήψης αποφάσεων (DMUs) όπου η παρουσία πολλαπλών εισόδων και εξόδων καθιστά

τις συγκρίσεις δύσκολες. Η σχετική απόδοση ενός DMU, μετριέται με την εκτίμηση του συνόλου των σταθμισμένων εξόδων με τις σταθμισμένες εισόδους και τη σύγκριση με άλλα DMUs. Η DEA επιλέγει για κάθε DMU το βάρος των εισόδων και εξόδων που μεγιστοποιούν την αποτελεσματικότητά του. Τα DMU που επιτυγχάνουν απόδοση 100% θεωρούνται αποδοτικά, ενώ τα άλλα DMU με βαθμολογία απόδοσης κάτω από 100% είναι αναποτελεσματικά.

Η DEA αναπτύχθηκε σε επιχειρησιακές έρευνες και οικονομικές μελέτες ως μέθοδος για την αξιολόγηση της αποτελεσματικότητας των μονάδων δραστηριότητας, κάνοντας τις ελάχιστες δυνατές παραδοχές σχετικά με τη λειτουργική μορφή της υποκείμενης λειτουργίας παραγωγής. Η DEA έχει χρησιμοποιηθεί εκτενώς για την αξιολόγηση της σχετικής αποτελεσματικότητας των μονάδων δραστηριότητας μη κερδοσκοπικού χαρακτήρα (π.χ. σχολεία, νοσοκομεία) και κερδοσκοπικούς οργανισμούς (π.χ. τράπεζες, αεροπορικές εταιρείες).

Δίνονται μερικά ακόμα παραδείγματα εφαρμογών της DEA σε διάφορους τομείς.

- Δημόσιες συγκοινωνίες (Karlaftis, M.G., Tsamboulas, D., 2012)
- Σχολεία ή πανεπιστήμια (Beasley, 1990, Ahn, 1987)
- Δικαστήρια (Lewin et al., 1982)
- Τράπεζες (Thompson et al., 1996, 1997, Athanassopoulos, 1997, Brockett et al., 1997 Drake and Howcroft, 1994, Oral et al., 1992, Schaffnit et al., 1997, Sherman and Ladino, 1995)
- Αεροπορικές εταιρείες (Schefczyk, 1993)
- Προγράμματα λογισμικού (Mahmood et al., 1996)
- Συντήρηση εξοπλισμού (Clark , 1992, Hjalmmarsson and Odeck, 1996),
- Ιατροφαρμακευτική περίθαλψη ή νοσοκομεία (Pina and Torres, 1992, Rutledge et al., 1995)
- Μονάδες γεωργικής παραγωγής (Haag et al., 1992),
- Αστυνομικές δυνάμεις (Thanassoulis, 1995)
- Σιδηρόδρομοι (Adolphson et al., 1989)
- Περιπολίες αυτοκινητόδρομων (Clark, 1992, Cook et al., 1990)
- Βιομηχανία ζυθοποιίας (Day et al., 1995)

Υπάρχουν διάφοροι λόγοι για τους οποίους η DEA χρησιμοποιείται για πολλές εφαρμογές. Πρώτον, δεν απαιτεί υποκείμενες παραδοχές για τα δεδομένα εισόδου και εξόδου. Δεύτερον, επιτρέπει στους διαχειριστές να εξετάζουν ταυτόχρονα πολλαπλά δεδομένα εισόδου και εξόδου ενός DMU. Τρίτον, παρέχει στους διαχειριστές μια μέθοδο για τη διάκριση μεταξύ αποδοτικών και μη αποδοτικών DMU. Τέταρτον, επισημαίνει τις πηγές και το περιθώριο βελτίωσης για όλα τα αναποτελεσματικά DMU. Τέλος, μπορεί να χρησιμοποιηθεί για την ανίχνευση της αναποτελεσματικότητας των DMUs που μπορεί

να μην είναι ανιχνεύσιμες μέσω άλλων τεχνικών όπως η γραμμική παλινδρόμηση ή αναλύσεις αναλογίας.

Παρά τις εκτεταμένες εφαρμογές της, η DEA παρουσιάζει ορισμένα μειονεκτήματα. Πρώτον, αν και η DEA είναι αποτελεσματική στην εκτίμηση της «σχετικής» απόδοσης ενός DMU συγκριτικά με ένα προκαθορισμένο σύνολο ομότιμων DMUs, δεν είναι αποτελεσματική στην εκτίμηση της «σχετικής» απόδοσης ενός DMU συγκριτικά με ένα «θεωρητικό μέγιστο». Έτσι, για να μετρηθεί η αποδοτικότητα ενός νέου DMU, πρέπει να αναπτυχθεί μια νέα DEA με τα δεδομένα DMUs που χρησιμοποιήθηκαν προηγουμένως. Επίσης, δεν μπορεί να προβλέψει το επίπεδο απόδοσης του νέου DMU χωρίς άλλη ανάλυση DEA. Τέλος, δεν παρέχει μια διαδρομή σταδιακής πορείας τη βελτίωση της αποτελεσματικότητας κάθε μη αποδοτικού DMU λαμβάνοντας υπόψη τις διαφορές στους δείκτες αποδοτικότητας.

Ορισμένοι ερευνητές πρότειναν τη DEA ως εργαλείο για την αξιολόγηση έργων E&A (Khouja M., 1995, Baker RC, Talluri S., 1997). Κατηγοριοποίησαν τα σχετικά μέτρα αξιολόγησης είτε ως εισροές είτε ως εκροές μοντέλου DEA και ταξινόμησαν τα έργα με βάση τους δείκτες απόδοσής τους. Οι Linton et al. (2002) χρησιμοποίησαν τη DEA για να διαχωρίσουν ένα χαρτοφυλάκιο έργων σε ομάδες «αποδοχής», «περαιτέρω σκέψης» και «απόρριψης», σαν πρώτο βήμα στην ανάλυση χαρτοφυλακίου και στη συνέχεια χρησιμοποίησαν μια προσέγγιση γραφικής ανάλυσης για να ολοκληρώσουν την αξιολόγηση. Οι Oralet et al. (1991) χρησιμοποίησαν τη DEA για να αξιολογήσουν τη διασταυρούμενη αποτελεσματικότητα σε συλλογικές ρυθμίσεις λήψης αποφάσεων.

Οι H. Eilat et al. (2008) παρουσίασαν μια προσέγγιση πολλαπλών κριτηρίων για την αξιολόγηση των έργων E&A, με βάση την ενσωμάτωση δύο διαφορετικών καινοτόμων μεθοδολογιών διαχείρισης. Συνδύασαν τις έννοιες που λαμβάνονται από την DEA και τον ισορροπημένο πίνακα επιδόσεων (Balanced Scorecard, BSC), οι οποίες έχουν αποδειχθεί χρήσιμες μετρήσεις και ανάλυση εργαλεία σε πολλές πρακτικές εφαρμογές. Αυτές οι έννοιες ενσωματώθηκαν σε ένα μόνο μοντέλο DEA-BSC. Οι τιμές που λαμβάνονται μέσω αυτού του μοντέλου αντιπροσωπεύουν τα «οφέλη» (δεδομένα εξόδου), το «κόστος» (δεδομένα εισόδου) και τις προτιμήσεις του οργανισμού. Το μοντέλο διακρίνει τα έργα σύμφωνα με τα επιθυμητά χαρακτηριστικά και τα κατατάσσει σύμφωνα με την επιδιωκόμενη έμφαση του οργανισμού.

Οι H. Lee et al. (2009) σύγκριναν την απόδοση έξι εθνικών προγραμμάτων E&A με ετερογενείς στόχους χρησιμοποιώντας τη DEA, αποδεικνύοντας ότι η DEA αποτελεί αποτελεσματικό εργαλείο για τη σύγκριση των επιδόσεων μεταξύ προγραμμάτων E&A με ετερογενείς στόχους.

Οι H.K. Hong et al. (1999) ανέπτυξαν μια μεθοδολογία υβριδικής ανάλυσης που χρησιμοποιεί τη DEA μαζί με τη μηχανική μάθηση, με σκοπό την αξιολόγηση της αποτελεσματικότητας των έργων ενοποίησης συστημάτων (System Integration, SI).

Αξίζει να σημειωθεί ότι, το πρώτο μοντέλο DEA που προτάθηκε από τους Charnes et al. (1978) είναι το μοντέλο CCR που υποθέτει ότι η παραγωγή παρουσιάζει σταθερή απόδοση κλίμακας. Οι Banker et al. (1984) το επέκτειναν το μοντέλο, στο μοντέλο BCC για μεταβαλλόμενη απόδοση κλίμακας. Όσον αφορά την κλίμακα απόδοσης των E&A, τα ευρήματα από προηγούμενες μελέτες είναι κάπως ανάμεικτα (Graves and Langowitz, 1996). Βρέθηκε ότι η δραστηριότητα E&A μπορεί να εμφανίζει αύξηση ή μείωση των αποδόσεων στην κλίμακα καθώς και σταθερή κλίμακα επιστροφής (Bound et al., 1984, Scherer, 1983). Για το λόγο αυτό, στη συγκεκριμένη διπλωματική επιλέγεται το μοντέλο BCC.

2.3. Συμπεράσματα Βιβλιογραφικής Ανασκόπησης

Η αξιολόγηση έργων E&A είναι ένα σημαντικό βήμα στη λήψη αποφάσεων E&A των οργανισμών. Αν και έχουν αναπτυχθεί πολλές τεχνικές αξιολόγησης E&A, η καθεμία έχει τα δικά της πλεονεκτήματα και μειονεκτήματα. Η εφαρμογή διαφορετικών τεχνικών μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα αξιολόγησης και συνεπώς σε διαφορετικές αποφάσεις E&A. Λόγω του μεγάλου αριθμού διαθέσιμων τεχνικών, η επιλογή των κατάλληλων μεθόδων αξιολόγησης θεωρείται κρίσιμη για τους υπεύθυνους λήψης αποφάσεων E&A.

Η DEA έχει προταθεί από ερευνητές ως εργαλείο για την αξιολόγηση έργων E&A. Τα σχετικά μέτρα αξιολόγησης κατηγοριοποιήθηκαν είτε ως εισροές είτε ως εκροές του μοντέλου DEA και ταξινομήθηκαν τα έργα με βάση τους δείκτες απόδοσής τους. Η DEA είναι αποτελεσματική στην εκτίμηση της «σχετικής» απόδοσης ενός DMU συγκριτικά με ένα προκαθορισμένο σύνολο ομότιμων DMUs, αλλά κρίνεται αναποτελεσματική στην εκτίμηση της «σχετικής» απόδοσης ενός DMU συγκριτικά με ένα «θεωρητικό μέγιστο». Έτσι, για να μετρηθεί η αποτελεσματικότητα ενός νέου DMU, πρέπει να αναπτυχθεί μια νέα DEA με τα δεδομένα DMUs που χρησιμοποιήθηκαν προηγουμένως. Επιπλέον, η DEA προσδιορίζει τα DMUs αναφοράς και τους στόχους των αναποτελεσματικών DMUs. Δεν παρέχει, όμως, μια διαδρομή σταδιακής πορείας για τη βελτίωση της αποτελεσματικότητας κάθε αναποτελεσματικού DMU λαμβάνοντας υπόψη τις διαφορές στους δείκτες αποδοτικότητας. Τέλος, δε μπορεί να προβλέψει το αποτελεσματικό επίπεδο ενός νέου DMU.

Απαιτείται, λοιπόν, περαιτέρω έρευνα για τη συνδυαστική χρήση εργαλείων, όπως η DEA και τα μοντέλα μηχανικής μάθησης, για την ανάπτυξη στρατηγικών μεθόδων

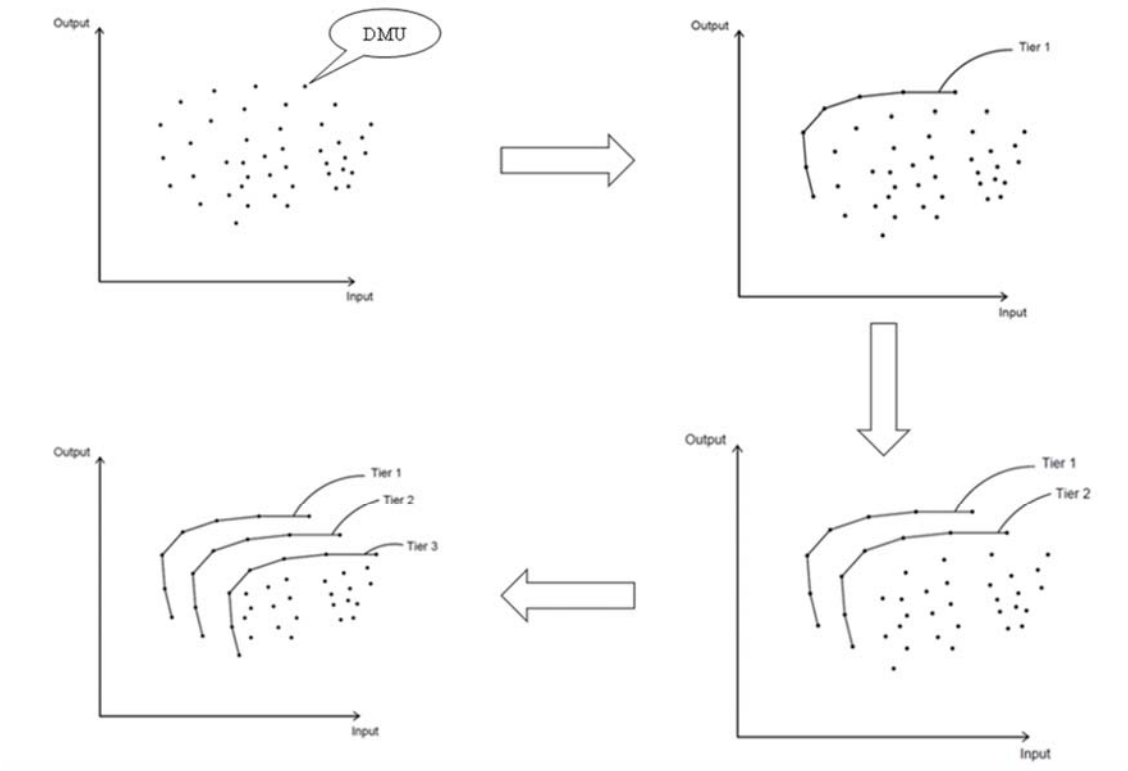
βελτίωσης της αποτελεσματικότητας των αναποτελεσματικών έργων Ε&Α και τη πρόβλεψη της απόδοσης νέων έργων Ε&Α.

3. Μεθοδολογία

3.1. Προσέγγιση

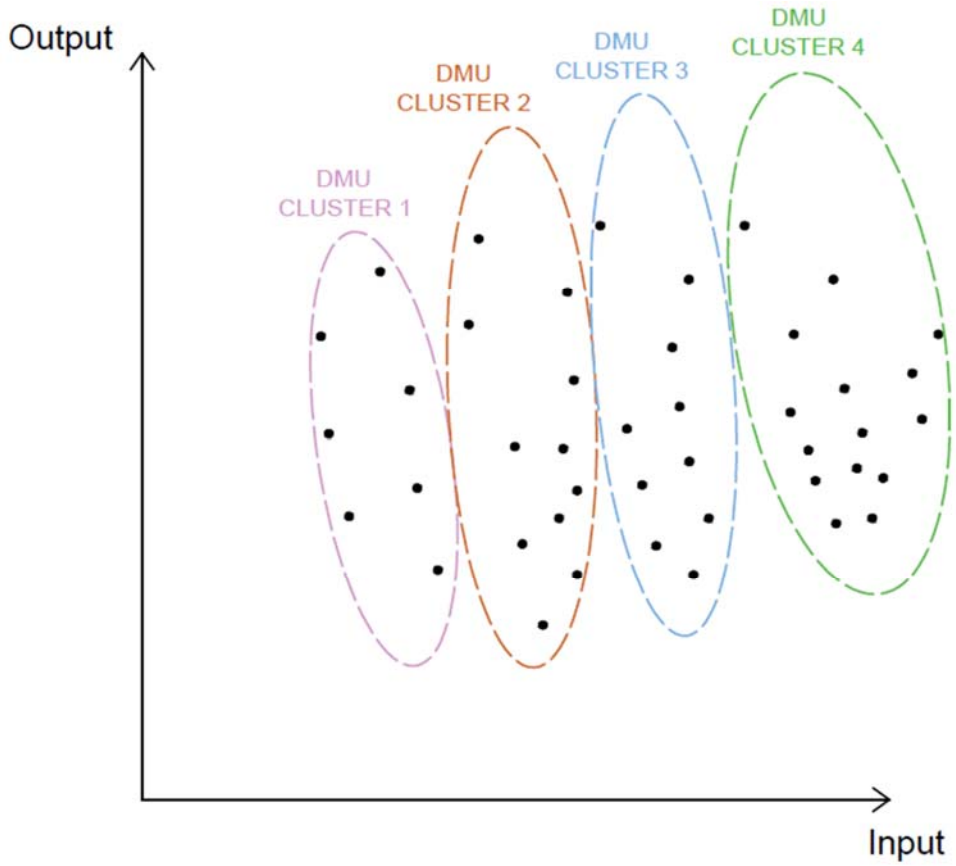
Για την επίτευξη των προαναφερθέντων στόχων, ακολουθείται η εξής πορεία, με τη βοήθεια του προγραμματιστικού περιβάλλοντος R.

Πρώτα, συλλέγονται και επεξεργάζονται όλα τα δεδομένα που είναι απαραίτητα για την αξιολόγηση των προγραμμάτων έρευνας και ανάπτυξης και καθορίζονται οι μεταβλητές εισροών (inputs) και εκροών (outputs). Έπειτα, εφαρμόζεται η μέθοδος Data Envelopment Analysis (DEA), γνωστή στα ελληνικά ως Περιβάλλουσα Ανάλυση Δεδομένων, στα προγράμματα έρευνας και ανάπτυξης, με προσανατολισμό στις εισροές. Μέσα από αυτή τη διαδικασία προκύπτει ένα σύνολο προγραμμάτων που έχουν δείκτη αποδοτικότητας ίσο με 1 και άρα είναι αποδοτικά. Αυτό το σύνολο αποκαλείται Tier 1 (Επίπεδο 1). Στο επόμενο βήμα, εφαρμόζεται πάλι η μέθοδος DEA μόνο με τα μη-αποδοτικά προγράμματα, εκείνα δηλαδή που δεν βρίσκονται στο Tier 1. Με αυτό τον τρόπο προκύπτει ένα σύνολο προγραμμάτων που έχουν δείκτη αποδοτικότητας ίσο με 1 και άρα είναι αποδοτικά. Αυτό το σύνολο αποκαλείται Tier 2. Η ίδια διαδικασία επαναλαμβάνεται όσο ο αριθμός των υπολειπόμενων παραγωγικών μονάδων είναι τουλάχιστον τρεις φορές μεγαλύτερος (3x5=15) από το άθροισμα του αριθμού των διαφορετικών μεταβλητών εισροών και εκροών (4+1=5), όπως προτείνεται από τον Banker et al. (1984).



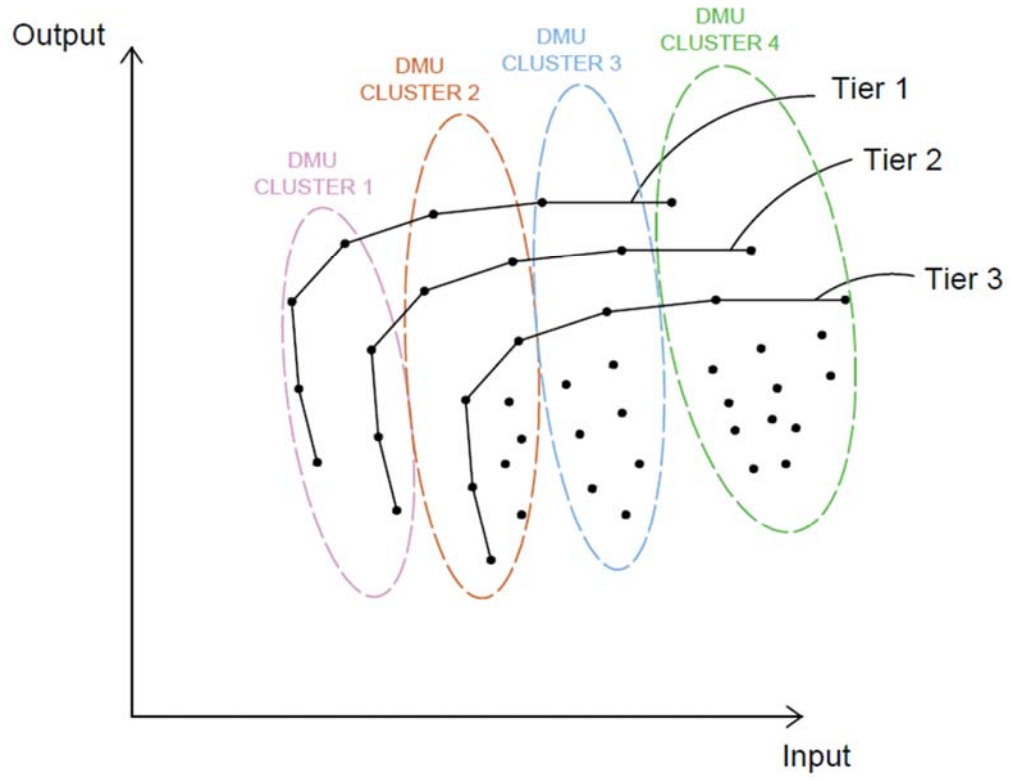
Εικόνα 3.1.1: Η διαδικασία ανάλυσης επιπέδων αποδοτικότητας (Tier Analysis)

Στη συνέχεια, χρησιμοποιήθηκαν μοντέλα μηχανικής μάθησης με σκοπό την ταξινόμηση (classification) και την ομαδοποίηση (clustering) των δεδομένων. Αρχικά, αναπτύχθηκε ένα Δένδρο Ταξινόμησης, με μάθηση με επίβλεψη, εισάγοντας τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας ως δεδομένα εισόδου, το οποίο θα έχει τη δυνατότητα να προβλέπει σε ποιο επίπεδο αποδοτικότητας βρίσκεται οποιοδήποτε υπάρχον ή νέο πρόγραμμα, με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών). Για το σκοπό αυτό, χρησιμοποιήθηκε ο αλγόριθμος C4.5 (J48). Μετέπειτα, πραγματοποιήθηκε η ομαδοποίηση των δεδομένων, χρησιμοποιώντας τον αλγόριθμο K-μέσων (k-means), με μάθηση χωρίς επίβλεψη, χωρίζοντας τα προγράμματα (DMUs) σε διαφορετικές ομάδες (clusters), ανάλογα με τα χαρακτηριστικά του κάθε προγράμματος.

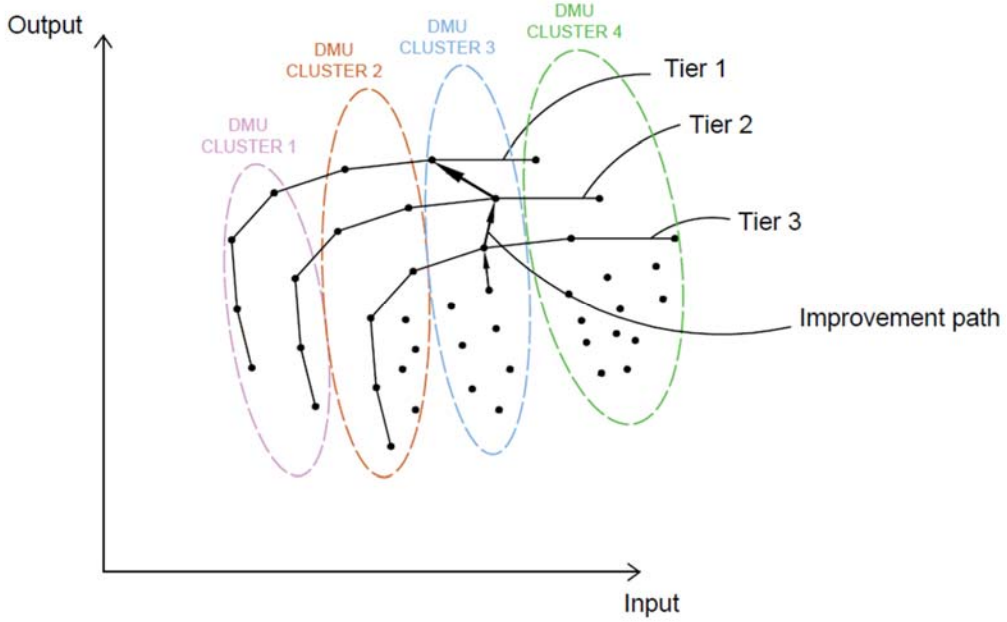


Εικόνα 3.1.2: Ομαδοποίηση δεδομένων (clustering)

Έπειτα, συνδυάζονται τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας και τα αποτελέσματα της ομαδοποίησης των DMUs, με σκοπό να προσδιοριστεί το σύνολο αναφοράς των μη αποδοτικών DMUs. Τα αποδοτικά DMUs στο ανώτερο Tier ενός cluster, αποτελούν το σύνολο αναφοράς για τα μη αποδοτικά DMUs των χαμηλότερων Tiers, που βρίσκονται στο ίδιο cluster. Τέλος, προσδιορίζεται η καλύτερη διαδρομή για τη σταδιακή βελτιστοποίηση (improvement path) των μη αποδοτικών DMUs, με βάση την μικρότερη Ευκλείδεια απόσταση ενός μη αποδοτικού DMU με ένα DMU που βρίσκεται στο αμέσως μεγαλύτερο Tier, ενώ παράλληλα βρίσκεται και στο ίδιο cluster.

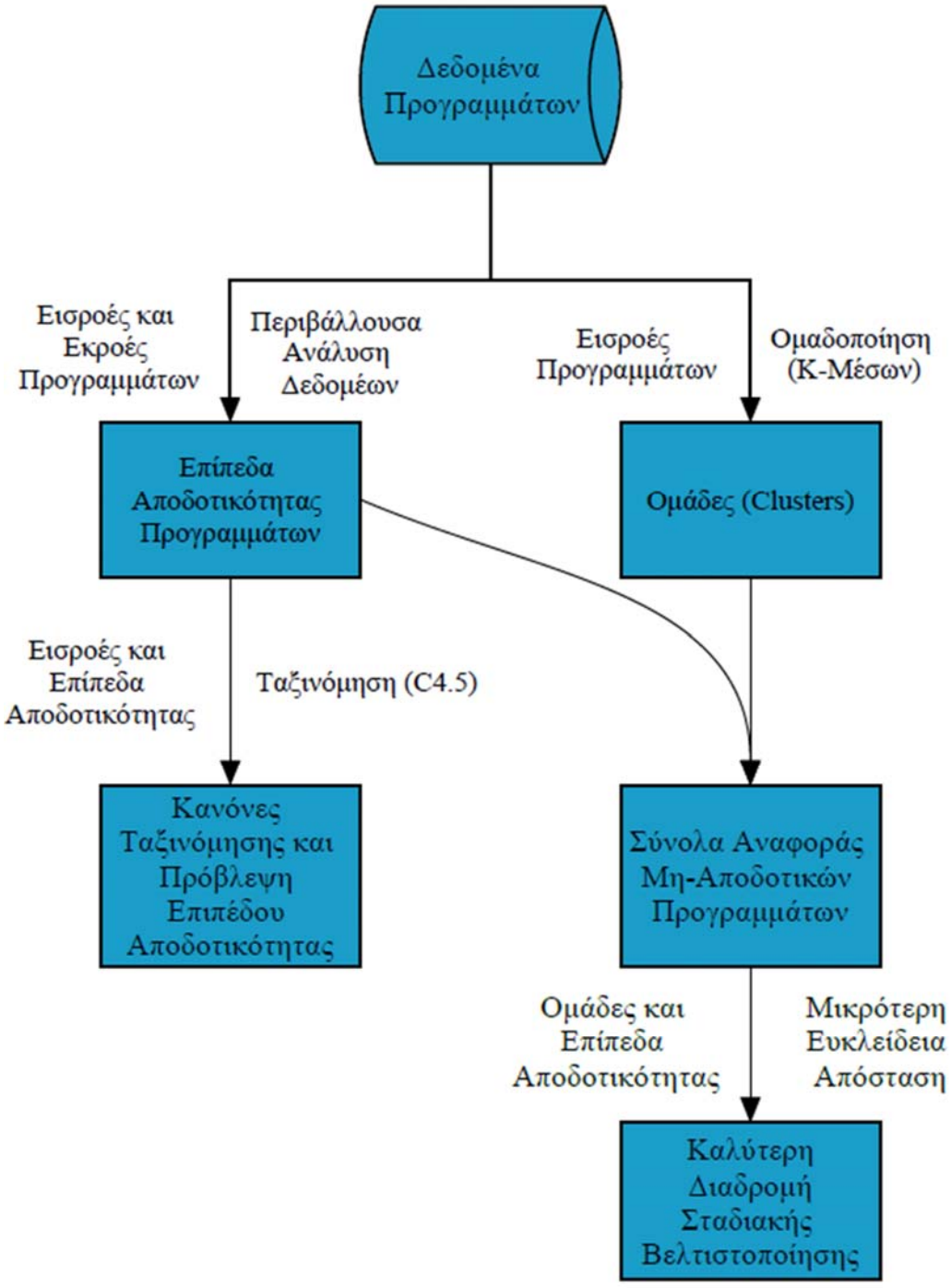


Εικόνα 3.1.3: Σύνολο αναφοράς των μη αποδοτικών DMUs



Εικόνα 3.1.4: Διαδρομή σταδιακής βελτιστοποίησης ενός μη-αποδοτικού DMU

Στο παρακάτω διάγραμμα (Διάγραμμα 3.1.1) φαίνεται η ροή των εργασιών.



Διάγραμμα 3.1.1: Ροή Εργασιών

3.2. Θεωρητικό Υπόβαθρο

3.2.1. Data Envelopment Analysis (DEA)

Η μέθοδος Data Envelopment Analysis (DEA), ή αλλιώς Περιβάλλουσα Ανάλυσης Δεδομένων (ΠΑΔ), αποτελεί μια μη παραμετρική μέθοδο γραμμικού προγραμματισμού, που αξιολογεί την αποδοτικότητα ενός συνόλου Μονάδων Λήψης Αποφάσεων (Decision-Making Units, DMUs).

Για να εφαρμοστεί σωστά η μέθοδος DEA, πρέπει πρώτα να καθοριστούν τα εξής:

- Οι Μονάδες Λήψης Αποφάσεων (DMUs):

Μονάδες Λήψης Αποφάσεων (Decision-Making Units, DMUs), είναι ένα σύνολο συγκρίσιμων και ομοιογενών ομάδων, οι οποίες μετατρέπουν πολλαπλές εισροές (inputs) σε πολλαπλές εκροές (outputs). Η διαδικασία μετατροπής εισροών σε εκροές περιγράφετε από την τεχνολογία παραγωγής της παραγωγικής μονάδας. Η τεχνολογία παραγωγής μετατρέπει εισροές σε εκροές, μέσω των οποίων και εκφράζεται. Μια επιτυχής εφαρμογή της χαρακτηρίζεται ως αποδοτική.

$$Απόδοση\ DMU = \frac{Εκροές}{Εισροές}, (Charnes, Cooper, Rhodes - 1978)$$

Εικόνα 3.2.1: Απεικόνιση της λειτουργίας των Μονάδων Λήψης Αποφάσεων

- Οι εισροές (inputs) και οι εκροές (outputs) των DMUs:

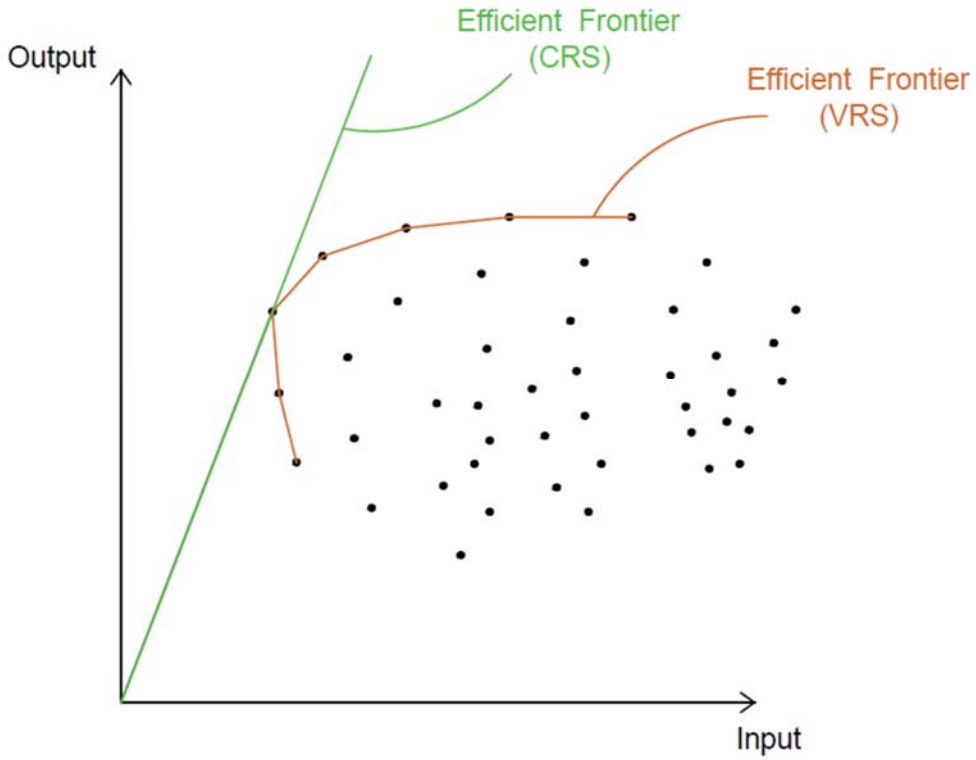
Εισροές θεωρούνται οι πόροι που χρησιμοποιούνται για την παραγωγή των εκροών, ενώ εκροές θεωρούνται τα προϊόντα ή οι υπηρεσίες που παράγονται από τις DMUs. Ο προσδιορισμός των εισροών και των εκροών στην αξιολόγηση μιας DMU είναι πολύ κρίσιμος για την επιτυχημένη εφαρμογή της DEA. Οι εισροές πρέπει να περιέχουν όλους τους πόρους που έχουν αντίκτυπο στις εξόδους και οι έξοδοι θα πρέπει να αντανακλούν όλα τα χρήσιμα αποτελέσματα με βάση τα οποία θα αξιολογηθεί η DMU.

- Ο προσδιορισμός του προσανατολισμού μοντέλου είτε προς τις εισροές, είτε προς τις εκροές (Input-oriented Model/ Output-oriented Model):

Τα μοντέλα DEA μπορούν να διαχωριστούν με βάση το αντικείμενο του μοντέλου, δηλαδή μπορεί να εστιάζουν στην ελαχιστοποίηση των δεδομένων εισόδου με προσανατολισμό στα δεδομένα εισόδου (Input-oriented) ή στη μεγιστοποίηση των δεδομένων εξόδου με προσανατολισμό στα δεδομένα εξόδου (Output-oriented).

- Ο προσδιορισμός των αποδόσεων κλίμακας (Constant Returns to Scale, CRS/ Variable Returns to Scale, VRS):

Τα μοντέλα DEA προσδιορίζουν ποια DMUs είναι αυτά που ορίζουν την Περιβάλλουσα επιφάνεια ή το σύνορο αποδοτικότητας. Η καταλληλόλητα ενός συγκεκριμένου συνόρου υπαγορεύεται από προϋποθέσεις, κυρίως οικονομικές, που αφορούν στο σύνολο των δεδομένων το οποίο υπεισέρχεται στην ανάλυση. Το μοντέλο CCR χρησιμοποιεί μόνο σταθερές οικονομίες κλίμακος ενώ το μοντέλο BCC μεταβλητές οικονομίες κλίμακος οδηγώντας με αυτό τον τρόπο σε διαφορετικό σύνορο αποδοτικότητας. Το σύνορο αυτό αποτελείται από την κυρτή θήκη των DMUs σε αντίθεση με το σύνορο της CCR που είναι ένα ευθύγραμμο τμήμα με ακμές τις αποδοτικές μονάδες. Οπότε τα DEA μοντέλα μπορούν να διαχωριστούν με βάση την απόδοση κλίμακας. Πιο συγκεκριμένα, διαχωρίζονται στα μοντέλα σταθερής απόδοσης κλίμακας (Constant Returns to Scale, CRS), όπου τα δεδομένα εξόδου αυξάνονται αναλογικά με τα δεδομένα εισόδου και στα μοντέλα με μεταβαλλόμενη απόδοση κλίμακας (Variable Returns to Scale, VRS), όπου τα δεδομένα εξόδου αυξάνονται με διαφορετικό ρυθμό σε σχέση με τα δεδομένα εισόδου.



Εικόνα 3.2.2: Αποδοτοτικά Σύνορα - Σταθερή και Μεταβαλλόμενη Απόδοση Κλίμακας (CRS, VRS)

Μετά τον καθορισμό των παραπάνω και σύμφωνα με το μοντέλο των Charnes, Cooper, Rhodes (CCR - 1978), αξιολογούνται τα n DMUs. Κάθε Μονάδα Αποφάσεων, DMU_j (j= 1, 2, …, n) χρησιμοποιεί m εισροές x_{ij} (i= 1, 2, …, m) και s εκροές y_{rj} (r= 1, 2, …, s). Έστω ότι $x_{ij} \geq 0$ και $y_{rj} \geq 0$ και ότι κάθε DMU έχει τουλάχιστον μια θετική εισροή και μια θετική εκροή. Οι εισροές και εκροές χρησιμοποιούνται για να μετριέται η σχετική τεχνική αποδοτικότητα των DMU_j, η οποία αξιολογείται συγκριτικά με τα δεδομένα όλων των DMUj (όπου j= 1, 2, …, n). Η σχετική τεχνική αποδοτικότητα μίας παραγωγικής μονάδας υπολογίζεται σχηματίζοντας τον λόγο του σταθμισμένου αθροίσματος των εκροών προς το σταθμισμένο άθροισμα των εισροών, όπου οι συντελεστές βαρύτητας επιλέγονται κατά τέτοιο τρόπο, ώστε να υπολογίζεται η κατά Pareto αποδοτικότητα της υπό εξέταση παραγωγικής μονάδας. Ένας συνηθισμένος τρόπος μέτρησης της σχετικές αποδοτικότητας είναι ο εξής:

$$Σχετική\ Αποδοτικότητα\ DMU_{jo} = \frac{\text{Σταθμισμένο Αθροισμα Εκροών}}{\text{Σταθμισμένο Αθροισμα Εισροών}}$$

Η DEA επιλύει το ακόλουθο μη γραμμικό κλασματικό πρόβλημα μαθηματικού προγραμματισμού για κάθε DMU.

$$E_{jo} = max\ \theta\ (u, v) = \frac{\sum_{r=1}^{s} y_{r0} * u_r}{\sum_{i=1}^{m} x_{i0} * v_i} \qquad \textit{(Μοντέλο CCR)}$$

Με περιορισμούς:

$$\frac{\sum_{r=1}^{s} y_{rj} * u_r}{\sum_{i=1}^{m} x_{ij} * v_i} \leq 1 \ \textit{για κάθε DMU}j \ \text{όπου}\ j = 1, \dots, n$$

$$\forall u_r, v_i \geq 0$$

$o = \eta$ *μονάδα προς αξιολόγηση*

$j = 1, \dots, n$ *πλήθος παραγωγικών μονάδων*

$r = 1, \dots, s$ *αριθμός εκροών*

$i = 1, \dots, m$ *αριθμός εισροών*

y_{rj} *= εκροή r της μονάδας j*

x_{ij} *= εισροή i της μονάδας j*

vi, ur συντελεστές για την εισροή i και την εκροή j που μεγιστοποιούν την αντικειμενική συνάρτηση της μονάδας προς αξιολόγηση

Η σχετική αποδοτικότητα κάθε μονάδας επιτυγχάνεται θεωρώντας στο πρόβλημα CCR τους συντελεστές v_i, u_r σαν μεταβλητές και μεγιστοποιώντας την αποδοτικότητα του DMU_o κάτω από τον περιορισμό ότι κανένα DMU με το ίδιο σύνολο συντελεστών βαρύτητας δεν θα έχει αποδοτικότητα μεγαλύτερη από το την μονάδα.

Το παραπάνω μοντέλο αποτελεί ένα μοντέλο κλασματικού προγραμματισμού. Για την επίλυση του μοντέλου DEA, η σχέση μετατρέπεται σε γραμμική, για να μπορούν να εφαρμοστούν οι μέθοδοι γραμμικού προγραμματισμού. Θέτεται ο παρονομαστής του κλάσματος ίσος με 1 ($\sum_{i=1}^{m} x_{ij} * v_i = 1$) και μεγιστοποιείται ο αριθμητής. Η τελική γραμμική εξίσωση έχει την παρακάτω μορφή.

$$Efficiency = max \sum_{r=1}^{s} y_{r0} * u_r$$

Με περιορισμούς:

$$\sum_{r=1}^{s} y_{rj} * u_r - \sum_{i=1}^{m} x_{ij} * v_i \ \leq 0 \ \text{όπου}\ j = 1, \dots, n \qquad (1)$$

$$\sum_{i=1}^{m} x_{ij0} * v_i = 1 \qquad (2)$$

$$\forall u_r, v_i \geq 0$$

Για το προηγούμενο μοντέλο, υπάρχει και αντίστοιχο δυικό. Για τη μετατροπή του παραπάνω μοντέλου σε δυικό, στον περιορισμό (1) αντιστοιχεί η μεταβλητή θ_o, χωρίς περιορισμό στο πρόσημο και στον περιορισμό (2) αντιστοιχεί η μεταβλητή $\lambda_j \geq 0$.

Στον παρακάτω πίνακα (Πίνακας 3.2.1), δίνονται συνοπτικά τα μαθηματικά μοντέλα που χρησιμοποιούνται ανάλογα με το αν ο προσανατολισμός του μοντέλου είναι προς τις εισροές ή προς τις εκροές και ανάλογα η απόδοση κλίμακας είναι σταθερή ή μεταβαλλόμενη.

Πίνακας 3.2.1: Μαθηματικά Μοντέλα

Μαθηματικά Μοντέλα	Προσανατολισμό στα δεδομένα εισόδου (Input-oriented)	Προσανατολισμό στα δεδομένα εξόδου (Output-oriented)
CCR - Μοντέλο σταθερής απόδοσης κλίμακας (Constant Returns to Scale, CRS)	Αντικειμενική Συνάρτηση: **minθ** Περιορισμοί: $\sum_{j=1}^{n} x_{ij} * \lambda_j \leq \theta * x_{io}, i = 1, \dots, m$ $\sum_{j=1}^{n} y_{rj} * \lambda_j \leq y_{ro}, i = 1, \dots, s$ $\lambda_j \geq 0, j=1, \dots, n$	Αντικειμενική Συνάρτηση: **maxθ** Περιορισμοί: $\sum_{j=1}^{n} x_{ij} * \lambda_j \leq x_{io}, i = 1, \dots, m$ $\sum_{j=1}^{n} y_{rj} * \lambda_j \leq \theta * y_{ro}, i = 1, \dots, s$ $\lambda_j \geq 0, j=1, \dots, n$
BBC - Μοντέλο μεταβαλλόμενης απόδοσης κλίμακας (Variable Returns to Scale, VRS)	Αντικειμενική Συνάρτηση: **minθ** Περιορισμοί: $\sum_{j=1}^{n} x_{ij} * \lambda_j \leq \theta * x_{io}, i = 1, \dots, m$ $\sum_{j=1}^{n} y_{rj} * \lambda_j \leq y_{ro}, i = 1, \dots, s$ $\lambda_j \geq 0, j=1, \dots, n$ $\sum_{j=1}^{n} \lambda_j = 1$	Αντικειμενική Συνάρτηση: **maxθ** Περιορισμοί: $\sum_{j=1}^{n} x_{ij} * \lambda_j \leq x_{io}, i = 1, \dots, m$ $\sum_{j=1}^{n} y_{rj} * \lambda_j \leq \theta * y_{ro}, i = 1, \dots, s$ $\lambda_j \geq 0, j=1, \dots, n$ $\sum_{j=1}^{n} \lambda_j = 1$

Εάν θ=1, τότε η DMU η οποία βρίσκεται υπό αξιολόγηση είναι στο αποδοτικό σύνορο, δηλ. δεν υπάρχει κάποια άλλη DMU η οποία λειτουργεί περισσότερο αποτελεσματικά από αυτή. Διαφορετικά, η DMU που βρίσκεται υπό αξιολόγηση δεν

είναι αποδοτική. Η συγκεκριμένη DMU μπορεί είτε να αυξήσει το επίπεδο εκροών είτε να μειώσει το επίπεδο εισροών.

3.2.2. Ταξινόμηση (Classification) με Δένδρα Απόφασης (C4.5)

Η ταξινόμηση (classification) είναι μια από τις βασικότερες τεχνικές της Εξόρυξης Δεδομένων. Η ταξινόμηση είναι εργασία επιβλεπόμενης μάθησης. Στόχος της επιβλεπόμενης μάθησης είναι η ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο και σε ένα σύνολο άλλων γνωρισμάτων. Το γνώρισμα στόχος αναφέρεται και ως εξαρτημένη μεταβλητή, ενώ τα υπόλοιπα γνωρίσματα αναφέρονται και ως ανεξάρτητες μεταβλητές. Με την επιβλεπόμενη μάθηση επιτυγχάνεται η δημιουργία ενός μηχανισμού λήψης αποφάσεων ή υπολογισμών, ο οποίος είναι ικανός να προβλέπει τις τιμές της εξαρτημένης μεταβλητής χρησιμοποιώντας τις ανεξάρτητες μεταβλητές.

Η διαδικασία της ταξινόμησης αποτελείται από δύο βασικά βήματα: την εκμάθηση και την ταξινόμηση.

Κατά την εκμάθηση (learning) δημιουργείται το μοντέλο με βάση ένα σύνολο προ κατηγοριοποιημένων παραδειγμάτων, που ονομάζεται δεδομένα εκπαίδευσης (training data). Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο ταξινόμησης, προκειμένου να σχηματιστεί το μοντέλο. Λόγω του ότι τα δεδομένα εκπαίδευσης ανήκουν σε μία προκαθορισμένη κατηγορία, η οποία είναι γνωστή, η κατηγοριοποίηση αποτελεί μέθοδος εποπτευομένης μάθησης (supervised learning). Το μοντέλο αναπαρίσταται με τη μορφή κανόνων ταξινόμησης (classification rules), δέντρων απόφασης (decision trees) ή μαθηματικών τύπων.

Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα (test data) για να υπολογίσουν την ακρίβεια του μοντέλου. Το μοντέλο ταξινομεί τα δοκιμαστικά δεδομένα. Έπειτα, η κατηγορία που σχηματίστηκε με βάση τα δοκιμαστικά δεδομένα συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται από το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπό εκπαίδευση μοντέλο. Στην περίπτωση που το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη.

Μια από τις βασικότερες και πιο δημοφιλείς μεθόδους ταξινόμησης είναι τα Δένδρα Αποφάσεων. Βασική λογική της κατασκευής τους είναι η διαδοχική διάσπαση του συνόλου των παρατηρήσεων σε υποσύνολα. Κριτήριο για τη διάσπαση είναι οι τιμές των μεταβλητών. Η διαδικασία των διαδοχικών διασπάσεων αναπαρίσταται με μια ανεστραμμένη δενδρική δομή. Στην κορυφή βρίσκεται ο κόμβος-ρίζα του δένδρου. Σε

κατώτερα επίπεδα βρίσκονται επιπλέον κόμβοι, οι οποίοι συνδέονται με ακμές με άλλα στοιχεία του δένδρου. Στο κατώτερο επίπεδο κάθε κλάδου βρίσκονται τα φύλλα του δένδρου. Ο κόμβος - ρίζα έχει μόνο εξερχόμενες ακμές που τον συνδέουν με στοιχεία του κατώτερου επιπέδου. Οι υπόλοιποι κόμβοι έχουν εισερχόμενες ακμές που τους συνδέουν με τους κόμβους του ανώτερου επιπέδου και εξερχόμενες ακμές που τους συνδέουν με στοιχεία του κατώτερου επιπέδου. Τέλος, τα φύλα έχουν μόνο εισερχόμενες ακμές, οι οποίες τα συνδέουν με τους κόμβους του ανώτερου επιπέδου. Κάθε κόμβος αντιπροσωπεύει έναν έλεγχο στα δεδομένα και αντίστοιχη διάσπαση τους σε δύο ή περισσότερα υποσύνολα, ανάλογα με το αποτέλεσμα του ελέγχου. Η συνηθέστερη εκδοχή είναι ο έλεγχος να περιλαμβάνει μία μόνο μεταβλητή, έχουν προταθεί ωστόσο αλγόριθμοι όπου σε έναν κόμβο ελέγχονται περισσότερες μεταβλητές. Κάθε ακμή αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και το αντίστοιχο υποσύνολο των δεδομένων. Τέλος, κάθε φύλο αντιπροσωπεύει μια απόφαση κατηγοριοποίησης.

Έχουν προταθεί διάφοροι αλγόριθμοι για τη δημιουργία Δένδρων Αποφάσεων. Ένας από τους πιο διαδεδομένους είναι ο ID3, καθώς και οι μετεξελίξεις του, ο C4.5 και η εμπορική του εκδοχή C5.0.

Στον ID3 το κριτήριο που χρησιμοποιείται ονομάζεται Κέρδος Πληροφορίας (ΚΠ). Για κάθε διαθέσιμο γνώρισμα υπολογίζεται το Κέρδος Πληροφορίας και επιλέγεται το γνώρισμα με τη μεγαλύτερη τιμή ΚΠ.

Το Κέρδος Πληροφορίας(S,A) (Information Gain (S,A)) εκφράζει τη μείωση της εντροπίας που θα προκύψει, εάν ένα σύνολο παρατηρήσεων S διαχωριστεί σε υποσύνολα με βάση τις τιμές του γνωρίσματος A. Η εντροπία μετρά την ανομοιογένεια του συνόλου S, ανάλογα με τη διασπορά των παρατηρήσεων ως προς την κλάση στην οποία ανήκουν. Ας θεωρήσουμε ένα σύνολο S το οποίο περιέχει s παρατηρήσεις. Εάν η κλάση είναι δυαδική, εάν δηλαδή υπάρχουν δύο δυνατές τιμές για το γνώρισμα της κλάσης, τότε οι παρατηρήσεις της μίας τιμής κλάσης μπορούν να χαρακτηριστούν θετικές, ενώ οι υπόλοιπες μπορούν να χαρακτηριστούν αρνητικές. Το πλήθος των θετικών παρατηρήσεων είναι s_p και το πλήθος των αρνητικών παρατηρήσεων είναι s_n. Η εντροπία του συνόλου S ορίζεται ως:

$$E(S) = -p_p * log_2(p_p) - p_n * log_2(p_n)$$

όπου p_p είναι το ποσοστό των θετικών παρατηρήσεων ($p_p = s_p / s$) και p_n είναι το ποσοστό των αρνητικών παρατηρήσεων ($p_n = s_n / s$).

Εάν το γνώρισμα της κλάσης μπορεί να πάρει c διαφορετικές τιμές και το πλήθος των παρατηρήσεων με τιμή κλάσης i είναι s_i, τότε η εντροπία του S ορίζεται ως:

$$E(S) = -\sum_{i=1}^{c} p_i * log_2(p_i)$$

όπου p_i είναι το ποσοστό των παρατηρήσεων που ανήκουν στην κλάση i ($p_i=s_i/s$)

Έστω ότι το γνώρισμα A μπορεί να πάρει u δυνατές διακριτές τιμές ($a_1,a_2,\ldots,a_u$). Το σύνολο S μπορεί να χωριστεί στα υποσύνολα ($S_1,S_2,\ldots,S_u$). Το S_1 αποτελείται από τις παρατηρήσεις οι οποίες έχουν τιμή a_1 στο γνώρισμα A. Αντιστοίχως, τα υπόλοιπα υποσύνολα S_j απαρτίζονται από τις παρατηρήσεις που έχουν την εκάστοτε τιμή a_j στο γνώρισμα A. Εάν επιλεχθεί ως μεταβλητή διαχωρισμού το γνώρισμα A, τότε η εντροπία του διαχωρισμού του συνόλου S σε υποσύνολα ανάλογα με τις τιμές του A, ορίζεται ως:

$$E(S,A) = \sum_{j=1}^{u} \frac{S_j}{S} * E(S_j)$$

όπου u το πλήθος των δυνατών τιμών του γνωρίσματος A, S_j το υποσύνολο των παρατηρήσεων οι οποίες έχουν την τιμή a_j στο γνώρισμα A, s_j το πλήθος των μελών του S_j, s είναι το πλήθος των μελών του S και $E(S_j)$ είναι η εντροπία του S_j.

Ουσιαστικά η εντροπία που προκύπτει από τον διαχωρισμό του S ισούται με το άθροισμα των εντροπιών των S_j πολλαπλασιασμένες με έναν συντελεστή βαρύτητας, ο οποίος σχετίζεται με το πλήθος των μελών τους. Όσο μικρότερη είναι η εντροπία τόσο αυξάνει ο βαθμός ομοιογένειας των υποσυνόλων. Το Κέρδος Πληροφορίας είναι η μείωση της εντροπίας, η οποία προκύπτει από τον διαχωρισμό και ορίζεται ως:

$$IG(S,A) = E(S) - E(S,A)$$

ID3 υπολογίζει για κάθε γνώρισμα το Κέρδος Πληροφορίας. Το γνώρισμα με το μεγαλύτερο Κέρδος Πληροφορίας επιλέγεται και ο διαχωρισμός των παρατηρήσεων γίνεται με βάση τις τιμές αυτού του γνωρίσματος. Με τον τρόπο αυτόν μεταβαίνουμε σε υποσύνολα μεγαλύτερης ομοιογένειας.

Ο αλγόριθμος C4.5, ο οποίος χρησιμοποιείται στη συγκεκριμένη διπλωματική, αποτελεί επέκταση του ID3 και προτάθηκε από τον ίδιο ερευνητή (Quinlan, 1993). Μια από τις βασικές βελτιώσεις αφορά το κριτήριο διαχωρισμού. Σύμφωνα με τον Quinlan το Κέρδος Πληροφορίας τείνει να ευνοεί γνωρίσματα με μεγάλο πλήθος τιμών. Τα γνωρίσματα αυτά οδηγούν σε μεγάλο αριθμό μικρών και πολύ ομοιογενών υποσυνόλων.

Σε πολλές περιπτώσεις όμως, τα γνωρίσματα αυτά δεν περιέχουν ουσιαστική πληροφορία. Αν για παράδειγμα τα δεδομένα περιέχουν πεδίο για κάποιον κωδικό, όπως ο αριθμός ταυτότητας, τότε το πεδίο αυτό θα έχει μεγάλο κέρδος πληροφορίας και θα επιλεγεί. Ωστόσο, δεν περιέχει πληροφορία χρήσιμη για την κατηγοριοποίηση. Για την αντιμετώπιση αυτού του προβλήματος, στον C4.5 χρησιμοποιείται το κριτήριο Λόγος Κέρδους (Gain Ratio), το οποίο ορίζεται ως:

$$Gain\ Ratio(S, A) = \frac{Information\ Gain\ (S, A)}{Entropy\ (S, A)}$$

Ο Λόγος Κέρδους κανονικοποιεί το κέρδος πληροφορίας ως προς την εντροπία. Μελέτες έχουν δείξει ότι ο Λόγος Κέρδους βελτιώνει την ακρίβεια και μειώνει την πολυπλοκότητα των δένδρων.

3.2.3. Ομαδοποίηση (K-μέσων) – Clustering (k-means)

Στην επιβλεπόμενη μάθηση (Supervised Learning) μας δίνεται ένα σύνολο δεδομένων με τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής. Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο να μπορεί να κατηγοριοποιήσει νέα δεδομένα σε κάποια από τις προϋπάρχουσες κλάσεις. Αντίθετα, στη μη επιβλεπομένη μάθηση (Unsupervised Learning) μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις-ετικέτες κάθε εγγραφής και στόχος είναι η χρήση κάποιου αλγορίθμου, ώστε αυτόματα να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέρουσα δομή των δεδομένων. Για παράδειγμα, η ομαδοποίηση είναι μια από τις τεχνικές μη επιβλεπόμενης μάθησης. Δοθέντων κάποιων δεδομένων χωρίς κλάσεις, οι αλγόριθμοι ομαδοποίησης, ομαδοποιούν τα δεδομένα σε ομάδες, έτσι ώστε εγγραφές, οι οποίες ανήκουν στην ίδια ομάδα, να έχουν όμοια ή παραπλήσια χαρακτηριστικά.

Ο αλγόριθμος κ-μέσων (k-means) είναι ένας αλγόριθμος ομαδοποίησης. Ο αλγόριθμος k-means ξεκινάει με k τυχαία σημεία, τα οποία ονομάζονται κεντροειδή της ομάδας και δηλώνουν το κέντρο βάρους της ομάδας. Το k υποδηλώνει τον αριθμό των ομάδων που θα δημιουργήσει ο αλγόριθμος. Ο αλγόριθμος εκτελεί επαναληπτικά δύο βήματα. Το πρώτο βήμα αφορά την ανάθεση σε κάποια ομάδα, ενώ το δεύτερο βήμα αφορά τον επαναπροσδιορισμό και τη μετατόπιση του κεντροειδούς κάθε ομάδας. Πιο συγκεκριμένα, ο αλγόριθμος εξετάζει κάθε δείγμα σε σχέση με τα κεντροειδή των ομάδων. Με χρήση του μέτρου απόστασης που έχει επιλεχτεί, αναθέτει το εξεταζόμενο δείγμα στην ομάδα, της οποίας το κεντροειδές είναι το πλησιέστερο ως προς το συγκεκριμένο δείγμα. Στο δεύτερο βήμα, παίρνοντας τον μέσο όρο των δειγμάτων κάθε ομάδας, υπολογίζονται ξανά τα κεντροειδή της κάθε ομάδας, ώστε το κεντροειδές να είναι πιο αντιπροσωπευτικό στην πρόσφατα διαμορφωμένη ομάδας. Ο αλγόριθμος

εκτελεί επαναληπτικά αυτά τα δύο βήματα, μέχρις ότου τα κεντροειδή των ομάδων να μετατοπίζονται ελάχιστα και σε απόσταση μικρότερη από κάποια δοθείσα τιμή κατωφλίου. Ως εναλλακτικό κριτήριο τερματισμού του αλγορίθμου μπορεί να χρησιμοποιηθεί και ο αριθμός επαναλήψεων του αλγορίθμου.

Η επιλογή του μέτρου απόστασης είναι κρίσιμο στην ομαδοποίηση, αφού καθορίζει τον τρόπο υπολογισμού της ομοιότητας μεταξύ των στοιχείων (x,y). Στη συγκεκριμένη εργασία χρησιμοποιείται η Ευκλείδεια Απόσταση με τύπο:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Επιπλέον, πολύ σημαντική είναι και η επιλογή του αριθμού (k) των ομάδων που θα δημιουργήσει ο αλγόριθμος. Για να επιλεχτεί ο βέλτιστος αριθμό k, αξιολογούνται οι ομαδοποιήσεις ανάλογα με το Dunn Index (DI) και η μέθοδος του Average Silhouette.

- Το DI ισούται με την ελάχιστη απόσταση μεταξύ ομάδων διαιρούμενη με το μέγιστο μέγεθος ομάδας. Ένα υψηλότερο DI συνεπάγεται καλύτερη ομαδοποίηση, που σημαίνει ότι οι ομάδες είναι συμπαγείς και καλά διαχωρισμένες από τις άλλες ομάδες.
- Η μέθοδος της Average Silhouette, μετρά την ποιότητα μιας ομαδοποίησης. Καθορίζει πόσο καλά έχει ομαδοποιηθεί κάθε στοιχείο. Ένα υψηλό πλάτος Average Silhouette υποδεικνύει καλή ομαδοποίηση. Η μέθοδος Average Silhouette υπολογίζει τη μέση σιλουέτα των παρατηρήσεων για διαφορετικές τιμές του k. Ο βέλτιστος αριθμός ομάδων k είναι αυτός που μεγιστοποιεί τη μέση σιλουέτα σε ένα εύρος πιθανών τιμών για το k.

3.3. Υπολογιστικά εργαλεία

Για την εκπόνηση της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού R και το πρόγραμμα RStudio. Η R είναι μια γλώσσα προγραμματισμού και ένα περιβάλλον λογισμικού για στατιστικούς υπολογισμούς και γραφικές απεικονίσεις. Το πρόγραμμα RStudio είναι ένα ολοκληρωμένο περιβάλλον ανάπτυξης της γλώσσας R.

Τα πακέτα που χρησιμοποιήθηκαν κατά την εκπόνηση της εργασίας είναι τα εξής:

- Cluster: Ομαδοποίηση Δεδομένων
- clValid: Επικύρωση αποτελεσμάτων ομαδοποίησης δεδομένων
- deaR: Όλες οι απαραίτητες συναρτήσεις για την εφαρμογή της DEA
- factoextra: Αλγόριθμοι ομαδοποίησης και οπτικοποίηση των ομαδοποιήσεων

- ggplot2: Οπτικοποίηση των δεδομένων
- readxl: Διαβάζει αρχεία excel (.xls και .xlsx) στη R.
- RWeka: Αλγόριθμοι μηχανικής μάθησης και εργαλεία ταξινόμησης, παλινδρόμησης, ομαδοποίησης και οπτικοποίησης
- tidyverse: Χειρισμός δεδομένων

4. Συλλογή και Επεξεργασία Στοιχείων

4.1.Συλλογή στοιχείων

Τα απαραίτητα στοιχεία για την εφαρμογή της μεθοδολογίας, που παρουσιάστηκε στο προηγούμενο κεφάλαιο, συλλέχτηκαν από τις βάσεις δεδομένων που παρέχει το CORDIS (Community Research and Development Information Service - Κοινοτική Υπηρεσία Πληροφοριών Έρευνας και Ανάπτυξης).

Το CORDIS δημιουργήθηκε το 1990 και διοικείται από την Υπηρεσία Εκδόσεων της Ευρωπαϊκής Ένωσης για λογαριασμό των Γενικών Διευθύνσεων, των Εκτελεστικών Οργανισμών και των Κοινών Επιχειρήσεων Έρευνας και Καινοτομίας της Ευρωπαϊκής Επιτροπής. ΤΟ CORDIS υποστηρίζεται από εξειδικευμένους εργολάβους για συντακτικές και τεχνικές υπηρεσίες. Η νομική βάση και η χρηματοδότηση του CORDIS προέρχονται από τα προγράμματα εργασίας του προγράμματος πλαισίου «Ορίζοντας 2020» για την έρευνα και την τεχνολογική ανάπτυξη.

Αποστολή του CORDIS είναι η διάδοση και η εκμετάλλευση των ερευνητικών αποτελεσμάτων από τους επαγγελματίες του χώρου, με σκοπό τη προώθηση της ανοιχτής επιστήμης, τη δημιουργία καινοτόμων προϊόντων και υπηρεσιών και την τόνωση της ανάπτυξης σε ολόκληρη την Ευρώπη.

Το CORDIS είναι το κύριο δημόσιο αποθετήριο και «πύλη» της Ευρωπαϊκής Επιτροπής για τη διάδοση πληροφοριών για όλα τα ερευνητικά έργα που χρηματοδοτούνται από την ΕΕ και τα αποτελέσματά τους. Ο ιστότοπος και το αποθετήριο περιλαμβάνουν όλες τις δημόσιες πληροφορίες που διαθέτει η Επιτροπή (ενημερωτικά δελτία έργου, δημοσιεύσεις και παραδοτέα) συντακτικό περιεχόμενο για την υποστήριξη της επικοινωνίας και της εκμετάλλευσης (ειδήσεις, εκδηλώσεις, ιστορίες επιτυχίας, περιοδικά, πολυγλωσσικά «αποτελέσματα εν συντομία» για το ευρύτερο κοινό) και περιεκτικούς συνδέσμους προς εξωτερικές πηγές, όπως δημοσιεύσεις ανοιχτής πρόσβασης και ιστότοπους.

Οι βάσεις δεδομένων των έργων που παρέχονται από το CORDIS και που χρησιμοποιήθηκαν για τη συγκεκριμένη εργασία, λήφθηκαν από την Πύλη Ανοιχτών Δεδομένων της Ευρωπαϊκής Ένωσης σε μορφή XLS. Αυτό το σύνολο δεδομένων περιλαμβάνει έργα και συναφείς οργανισμούς που χρηματοδοτούνται από την Ευρωπαϊκή Ένωση στο πλαίσιο του προγράμματος πλαισίου «Ορίζοντας 2020» για την έρευνα και την καινοτομία από το 2014 έως το 2020.

Πιο συγκεκριμένα, έγινε λήψη των εξής αρχείων:

- ✓ «H2020 Projects»: Περιέχει τις πληροφορίες δημόσιας επιχορήγησης για κάθε έργο, συμπεριλαμβανομένων των ακόλουθων πληροφοριών: Αριθμός ελέγχου εγγραφής (RCN), αναγνωριστικό έργου (Project ID), αρκτικόλεξο έργου, κατάσταση έργου, πρόγραμμα χρηματοδότησης, θέμα, τίτλος έργου, έργο ημερομηνία έναρξης, ημερομηνία λήξης έργου, στόχος έργου, συνολικό κόστος έργου, μέγιστη συνεισφορά ΕΚ (δέσμευση), αναγνωριστικό κλήσης , σχέδιο χρηματοδότησης (τύπος δράσης), συντονιστής, χώρα συντονιστή, συμμετέχοντες (ταξινομούνται σε λίστα διαχωρισμένων με ερωτηματικά), συμμετέχων χώρες (ταξινομούνται σε μια λίστα διαχωρισμένων με άνω τελεία).

- ✓ «H2020 Organisations»: Λίστα των συμμετεχόντων οργανισμών, συμπεριλαμβανομένων του αριθμού ελέγχου εγγραφής έργου (RCN), του αναγνωριστικού έργου (Project ID), του αρκτικόλεξου του έργου, του ρόλου του οργανισμού, του αναγνωριστικού οργανισμού, του ονόματος του οργανισμού, του μικρού ονόματος του οργανισμού, του τύπου του οργανισμού, της συμμετοχής που έληξε (αληθής / λάθος), της συνεισφοράς της Ευρωπαϊκής Επιτροπής και της χώρας του οργανισμού. Οι οργανισμοί αυτοί μπορεί να είναι οργανισμοί Ανώτατης ή δευτεροβάθμιας εκπαίδευσης, Ερευνητικοί οργανισμοί, Ιδιωτικές κερδοσκοπικές οντότητες (εκτός των ιδρυμάτων τριτοβάθμιας ή δευτεροβάθμιας εκπαίδευσης), Δημόσιοι φορείς (εκτός ερευνητικών οργανισμών και ιδρυμάτων δευτεροβάθμιας ή τριτοβάθμιας εκπαίδευσης) κ.λπ..

- ✓ «H2020 Report summaries»: Οι περιλήψεις περιοδικής αναφοράς (ή δημοσιεύσιμες περιλήψεις) από έργα H2020

- ✓ «H2020 Project deliverables»: Λίστα παραδοτέων από έργα H2020, συμπεριλαμβανομένων Αριθμός εγγραφής (RCN), τίτλος, αναγνωριστικό έργου (Project ID), αρκτικόλεξο έργου, πρόγραμμα χρηματοδότησης, θέματα, περιγραφή, τύπος παραδοτέου (πρωτότυπα, έγγραφα, εκθέσεις, ερευνητικά στοιχεία, ιστοσελίδες, συμπληρώματα διπλωμάτων ευρεσιτεχνίας, βίντεο κ.λπ.) και η διεύθυνση URL για το παραδοτέο έγγραφο.

- ✓ «H2020 Project publications»: Κατάλογος δημοσιεύσεων που συνδέονται με έργα H2020, συμπεριλαμβανομένου του αριθμού εγγραφής (RCN), του τίτλου, του αναγνωριστικού έργου (Project ID), του αρκτικόλεξου του έργου, του προγράμματος χρηματοδότησης, του θέματος, των συγγραφέων, του τίτλου του περιοδικού, του αριθμού περιοδικού, του έτους, του αριθμού

σελίδας, της έκδοσης, του ψηφιακού αναγνωριστικού αντικειμένου και της κατηγορίας δημοσίευσης (άρθρο αξιολόγησης από ομότιμους , άρθρο από μη ομότιμους κριτές, βιβλίο, μονογραφικό βιβλίο, διαδικασία συνεδρίου, διατριβή κ.λπ.).

Παράλληλα, διατίθενται και αρχεία με τα στοιχεία αναφοράς (θέματα προγραμμάτων, προγράμματα χρηματοδότησης (είδη δράσης), τύποι οργανισμών και χώρες), από την Πύλη Ανοιχτών Δεδομένων της Ευρωπαϊκής, που εμπεριέχουν τις διαφορετικές λίστες αναφοράς στις οποίες συνδέονται τα δεδομένα CORDIS, γεγονός που εξυπηρετεί στη καλύτερη κατανόηση των παραπάνω αρχείων.

4.2.Προετοιμασία

Είναι απαραίτητη η επεξεργασία των παραπάνω αρχείων, με σκοπό να προκύψει μια βάση δεδομένων, η οποία θα έχει όλα τα απαραίτητα στοιχεία που είναι αναγκαία για την εφαρμογή της DEA.

Αρχικά, δημιουργήθηκε στο αρχείο «H2020 Projects» μια στήλη με όνομα «Number of participant Countries» στην οποία, ύστερα από κατάλληλους υπολογισμούς, αποδόθηκε ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε πρόγραμμα. Επίσης, δημιουργήθηκε μια στήλη με όνομα «totalDuration», η οποία με βάση την ημερομηνία έναρξης (startDate) και την ημερομηνία λήξης (endDate) υπολογίζει τη συνολική διάρκεια του κάθε προγράμματος.

Στη συνέχεια, αθροίστηκαν οι διαφορετικοί οργανισμοί που συμμετείχαν στα διαφορετικά προγράμματα, από το αρχείο «H2020 Organisations», με τη δημιουργία ενός Συγκεντρωτικού Πίνακα (ένα ισχυρό εργαλείο για τον υπολογισμό, τη σύνοψη και την ανάλυση δεδομένων του προγράμματος Excel), με όνομα «Total Partners». Με τον ίδιο τρόπο, αθροίστηκαν όλα τα παραδοτέα του κάθε προγράμματος (με όνομα Συγκεντρωτικού Πίνακα «Publications»), από το αρχείο «H2020 Project deliverables», και όλες οι δημοσιεύσεις του κάθε προγράμματος (με όνομα Συγκεντρωτικού Πίνακα «Deliverables»), από το αρχείο «H2020 Project publications». Οι τρεις παραπάνω Συγκεντρωτικοί Πίνακες, ενσωματώθηκαν στο αρχείο «H2020 Projects» με βάση τον αναγνωριστικό αριθμό του έργου (Project ID).

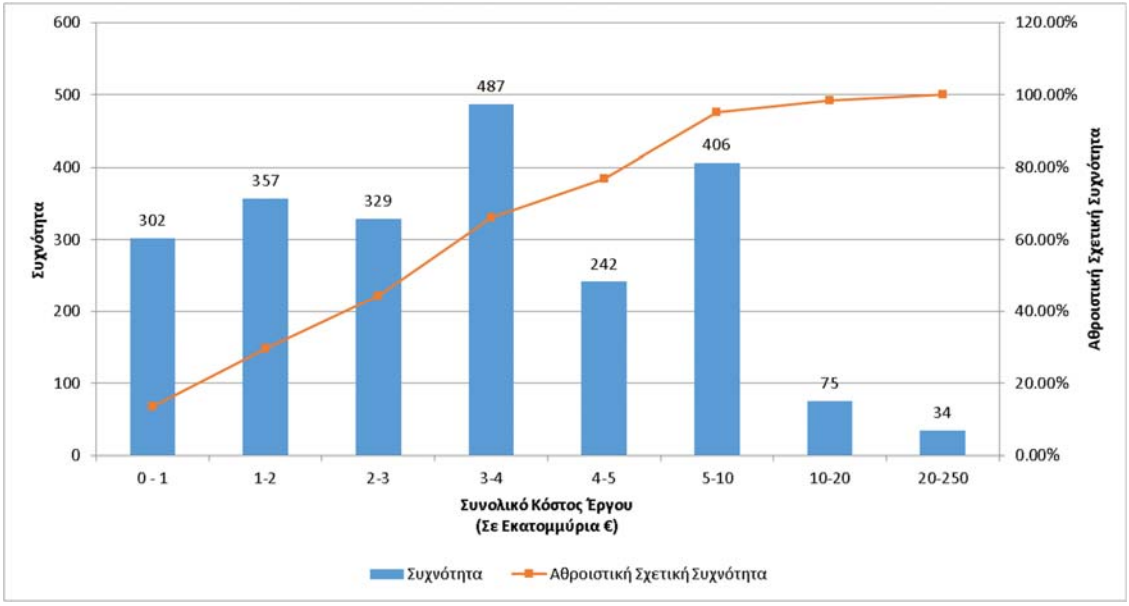
Έπειτα, πραγματοποιήθηκε «καθαρισμός» των δεδομένων. Πρώτον, διαγράφτηκαν όλα τα προγράμματα που δεν είχαν ολοκληρωθεί μέχρι τις 31/12/2019 από τη βάση δεδομένων, αφού δεν θα ήταν συγκρίσιμα με τα υπόλοιπα. Δεύτερον, διαγράφτηκαν όλα εκείνα τα προγράμματα που παρουσίαζαν κενά στα δεδομένα τους, σε σχέση με τα παραδοτέα και τις δημοσιεύσεις του κάθε προγράμματος. Τέλος, δημιουργήθηκε ένας νέος Συγκεντρωτικός Πίνακας, που περιέχει έξι στήλες με τον αναγνωριστικό αριθμό του έργου (ID), το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών

χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο (Total Partners), τη συνολική διάρκεια του κάθε έργου (Duration) και το άθροισμα των παραδοτέων με τις δημοσιεύσεις (Publications and Deliverables). Οι έξι αυτές στήλες, εμπεριέχουν τα στοιχεία που θα χρησιμοποιηθούν για την εφαρμογή της DEA.

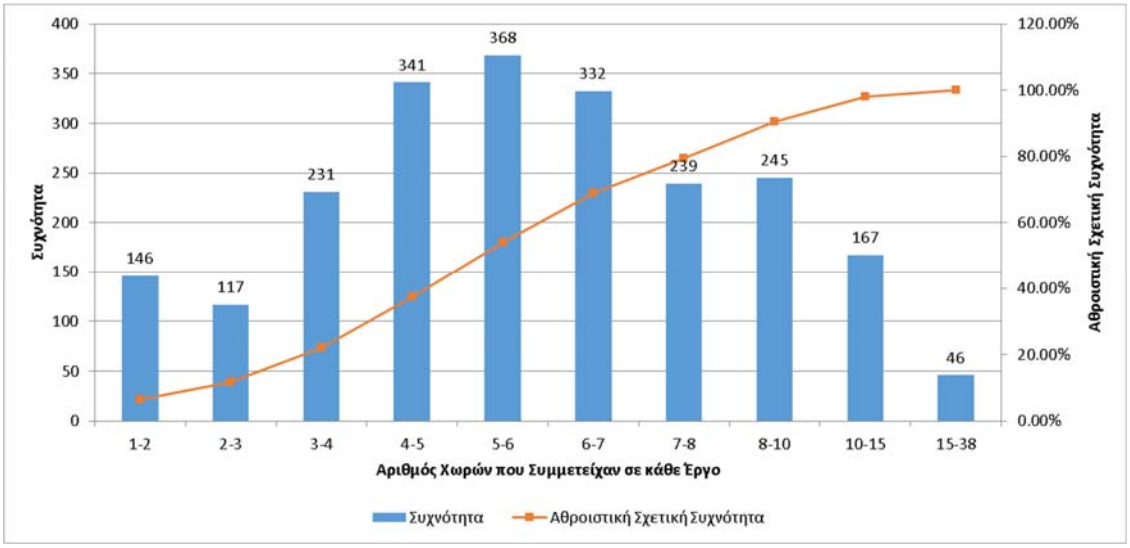
4.3. Περιγραφική Στατιστική

Τα δεδομένα που θα χρησιμοποιηθούν για την εφαρμογή της μεθοδολογίας DEA, μετά το «καθαρισμό» δεδομένων που έγινε στο προηγούμενο βήμα, αφορούν 2232 προγράμματα που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση.

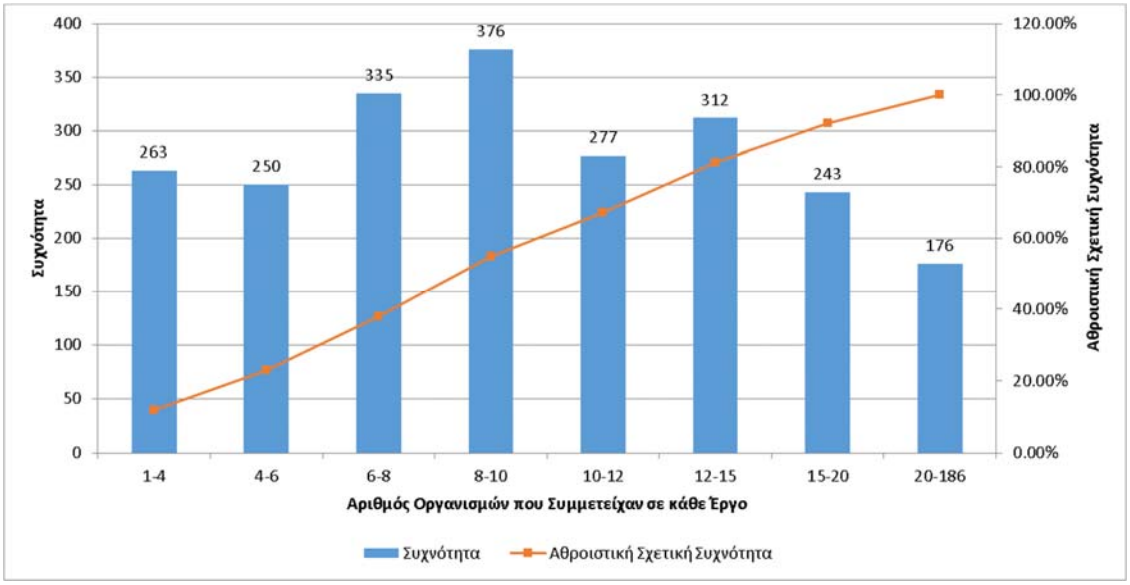
Σημαντικό σε αυτό το σημείο είναι να δοθεί η γενική εικόνα των δεδομένων που θα χρησιμοποιηθούν κατά την εκπόνηση της διπλωματικής εργασίας. Παρακάτω παρουσιάζονται γραφικά οι συχνότητες των τιμών της κάθε μεταβλητής καθώς και η αθροιστική σχετική συχνότητα τους (Διαγράμματα 4.3:1 έως 4.3.5).



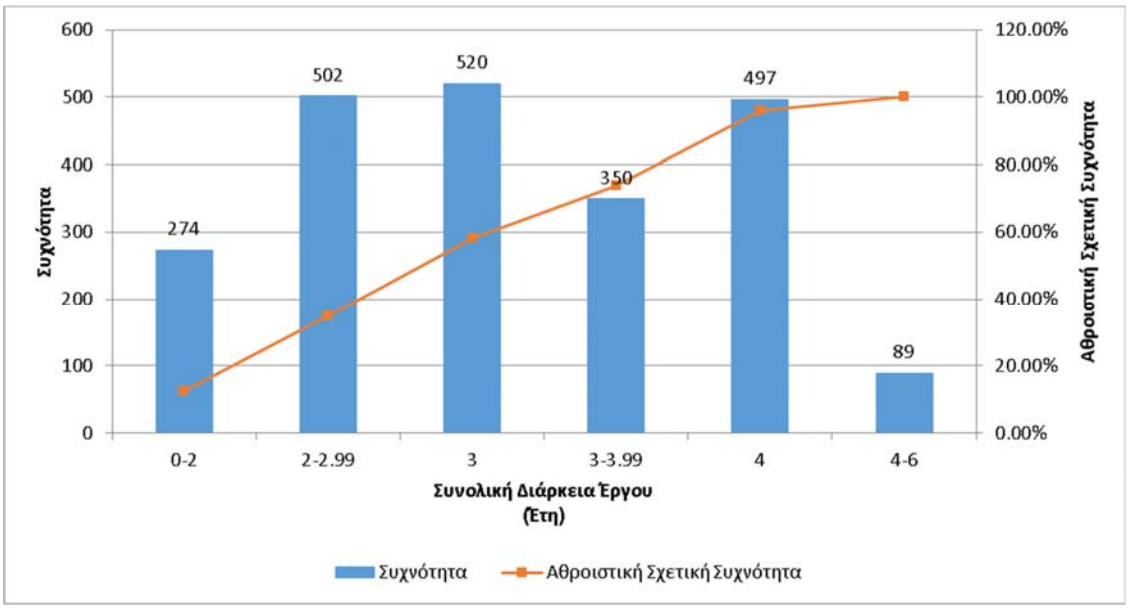
Διάγραμμα 4.3.1: Συχνότητες τιμών Συνολικού Κόστους



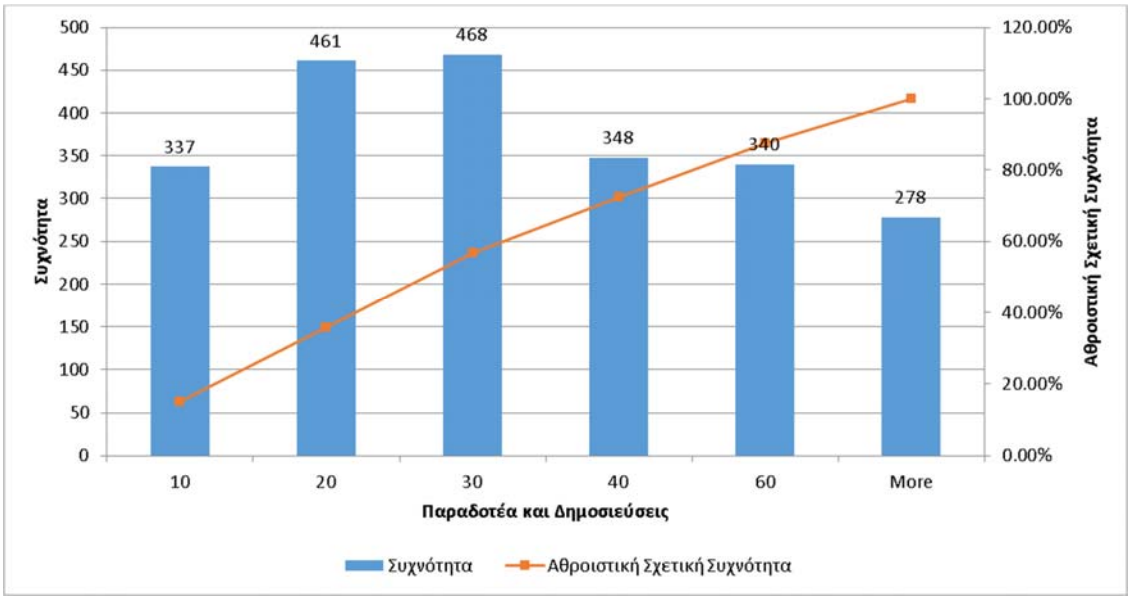
Διάγραμμα 4.3.2: Συχνότητες τιμών των Χωρών που Συμμετείχαν σε κάθε Έργο



Διάγραμμα 4.3.3: Συχνότητες τιμών των Οργανισμών που Συμμετείχαν σε κάθε Έργο



Διάγραμμα 4.3.4: Συχνότητες τιμών Συνολικής Διάρκειας Έργου



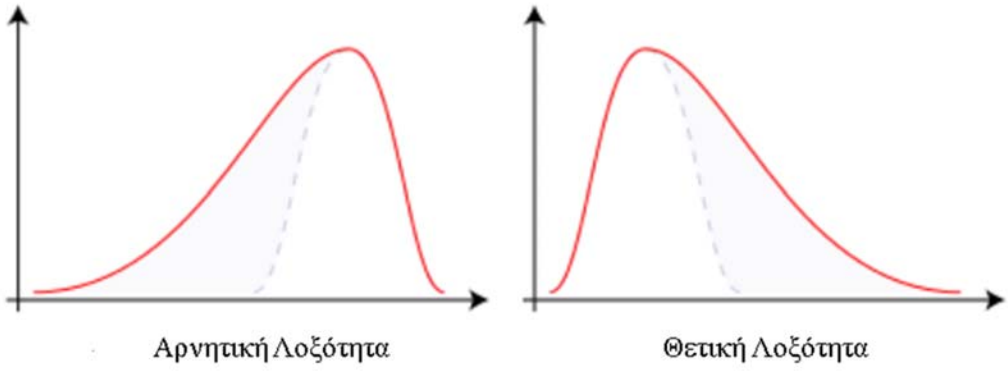
Διάγραμμα 4.3.5: Συχνότητες τιμών Παραδοτέων και Δημοσιεύσεων

Παρακάτω δίνεται μια σύντομη περιγραφική στατιστική ανάλυση των μεταβλητών που θα χρησιμοποιηθούν.

Πίνακας 4.3:1: Πίνακας Περιγραφικών Στατιστικών των Μεταβλητών

Περιγραφικά Στατιστικά Μεταβλητών	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables
Μέση Τιμή	4,419,874.98	6.68	11.34	3.14	34.91
Τυπικό Σφάλμα	170,386.56	0.07	0.18	0.02	0.81
Διάμεσος	3,354,918.59	6.00	10.00	3.00	27.00
Επικρατούσα Τιμή	1,000,000.00	6.00	9.00	3.00	24.00
Τυπική Απόκλιση	8,049,750.62	3.43	8.43	0.74	38.47
Διασπορά	64,798,485,106,347.50	11.77	71.09	0.54	1,480.29
Κυρτότητα (Kurtosis)	390.45	10.31	97.14	0.02	203.28
Ασυμμετρία (Skewness)	16.17	2.03	6.17	-0.34	9.81
Εύρος	241,255,303.98	37.00	185.00	5.48	1,013.00
Ελάχιστο	50,000.00	1.00	1.00	0.50	2.00
Μέγιστο	241,305,303.98	38.00	186.00	5.98	1,015.00

Μία κατανομή συχνοτήτων ονομάζεται συμμετρική όταν είναι φανερό πως υπάρχει ένας κατακόρυφος άξονας ο οποίος λειτουργεί ως καθρέπτης της μισής κατανομής στην άλλη μισή. Η συμμετρία των δεδομένων κάθε μεταβλητής, γίνεται φανερή από το συντελεστή ασυμμετρίας Skewness. Παρατηρείται πως όλες οι μεταβλητές έχουν μεγάλη λοξότητα προς τα αριστερά, αφού έχουν θετικό πρόσημο, εκτός από τη μεταβλητή Duration, η οποία είναι σχεδόν συμμετρική με μια λοξότητα προς τα δεξιά.



Εικόνα 4.3.1: Απεικόνιση θετικής και αρνητικής λοξότητας (Skewness)

Επιπλέον, οι μεταβλητές εξετάζονται ως προς την κυρτότητα τους (Kurtosis). Η κανονική κατανομή είναι το μέτρο και για την κυρτότητα μίας κατανομής. Όταν μια κατανομή έχει την ίδια κυρτότητα με την κανονική κατανομή και άρα συντελεστή Kurtosis ίσο με μηδέν, ονομάζεται Μεσόκυρτη. Αν κάποια κατανομή έχει περισσότερο "οξεία" κορυφή από αυτή της κανονικής κατανομής και έχει θετικό συντελεστή Kurtosis, τότε ονομάζεται Λεπτόκυρτη και παρουσιάζει μεγαλύτερη πιθανότητα να έχει ακραίες τιμές (outliers) σε σχέση με τη κανονική κατανομή. Αντίθετα, όταν μια κατανομή έχει περισσότερο "πλατιά" κορυφή και έχει αρνητικό συντελεστή Kurtosis τότε ονομάζεται Πλατύκυρτη και παρουσιάζει μικρότερη πιθανότητα να έχει ακραίες τιμές σε σχέση με τη κανονική κατανομή. Από τις μεταβλητές των δεδομένων, όλες οι μεταβλητές έχουν θετικό συντελεστή Kurtosis και άρα παρουσιάζουν μεγαλύτερη πιθανότητα να έχουν ακραίες τιμές σε σχέση με τη κανονική κατανομή, εκτός από τη μεταβλητή Duration που έχει συντελεστή οριακά ίσο με το μηδέν και άρα παρουσιάζει σχεδόν ίση πιθανότητα να έχει ακραίες τιμές σε σχέση με τη κανονική κατανομή.

Εικόνα 4.3.2: Απεικόνιση Ειδών Κυρτότητας (Kurtosis)

Η παραπάνω στατιστική ανάλυση των δεδομένων, φανερώνει ότι είναι πιθανό να παρουσιαστούν δυσκολίες κατά τη διαδικασία ομαδοποίησης των δεδομένων, λόγω των αυξημένων ακραίων τιμών.

5. Εφαρμογή Μεθοδολογίας – Αποτελέσματα

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα που προέκυψαν από την εφαρμογή της μεθοδολογίας που εξηγήθηκε αναλυτικά στο 3ο κεφάλαιο.

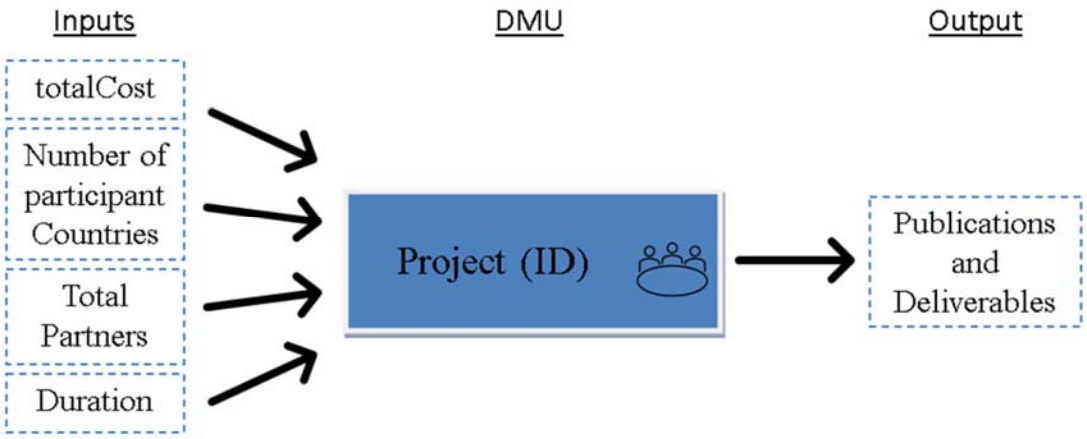
5.1. Αξιολόγηση της αποδοτικότητας των προγραμμάτων της Ευρωπαϊκής Ένωσης με Μοντέλα DEA

5.1.1. Καθορισμός Παραμέτρων

Για τη σωστή εφαρμογή της μεθοδολογίας DEA, είναι απαραίτητος ο καθορισμός των παραμέτρων που θα χρησιμοποιηθούν.

Αρχικά, πρέπει να καθοριστούν οι Μονάδες Λήψης Αποφάσεων (DMUs), που θα αξιολογηθούν. Το σύνολο των συγκρίσιμων και ομοιογενών ομάδων που αξιολογούνται στη παρούσα εργασία, είναι τα 2232 προγράμματα που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση τα τελευταία έξι χρόνια (Ιανουάριος του 2014 μέχρι Δεκέμβριος του 2019) και που προέκυψαν κατά τον «καθαρισμό» δεδομένων. Το κάθε πρόγραμμα αντιπροσωπεύεται στο μοντέλο της DEAαπό τον αναγνωριστικό αριθμό του έργου (ID).

Στη συνέχεια, το μοντέλο απαιτεί τον καθορισμό των εισροών και εκροών, οι οποίες αποτελούν τη βάση για τη μέτρηση της αποδοτικότητας. Στη συγκεκριμένη εργασία, ως εισροές τίθενται το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο (Total Partners), τη συνολική διάρκεια του κάθε έργου (Duration), ενώ ως εκροές τίθεται το άθροισμα των παραδοτέων με τις δημοσιεύσεις (Publications and Deliverables). Επισημαίνεται ότι η επιλογή των εισροών και των εκροών έγινε με βάση τα διαθέσιμα δεδομένα και με τέτοιο τρόπο ώστε να φανερώνουν τη δυναμική του κάθε DMU.



Εικόνα 5.1.1: Απεικόνιση παραμέτρων

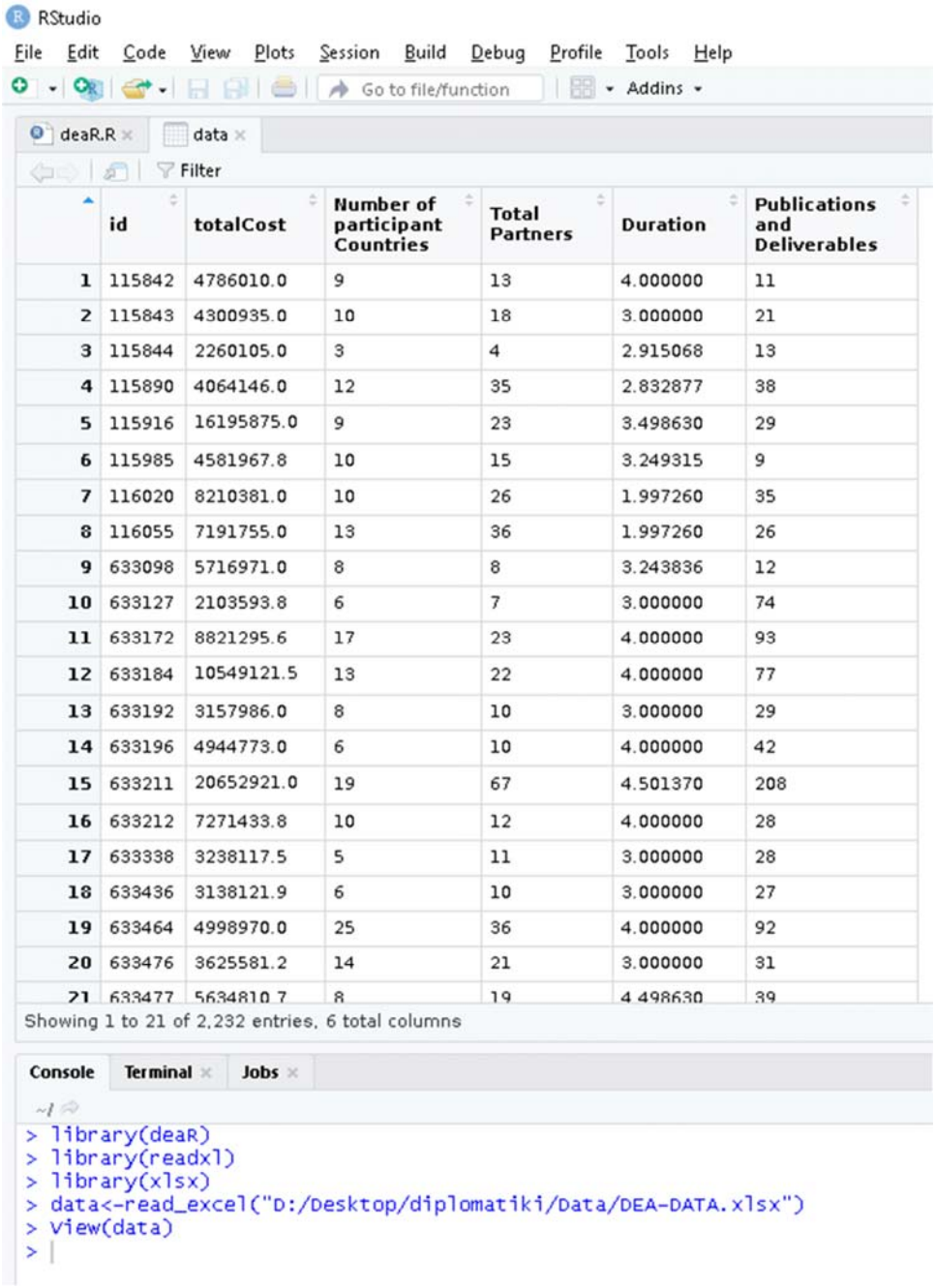
Έπειτα, επιλέχτηκε για το συγκεκριμένο μοντέλο προσανατολισμός προς τα δεδομένα εισόδου (input-oriented). Στη παρούσα εργασία, διερευνάται το περιθώριο που έχει η Ευρωπαϊκή Ένωση να αξιοποιήσει καλύτερα τους πόρους της. Όποτε, το μοντέλο που αναπτύχθηκε, εστιάζει στην ελαχιστοποίηση των δεδομένων εισόδου, διατηρώντας τα δεδομένα εξόδου σταθερά.

Τέλος, επιλέχτηκε μεταβαλλόμενη απόδοση κλίμακας (Variable Returns to Scale, VRS) για το μοντέλο, από τη στιγμή που τα προγράμματα είναι διαφορετικής κλίμακας και άρα τα δεδομένα εξόδου αυξάνονται με διαφορετικό ρυθμό σε σχέση με τα δεδομένα εισόδου.

5.1.2. Εφαρμογή μεθοδολογίας DEA

Όπως αναφέρθηκε και στο 3ο κεφάλαιο, για την υλοποίηση της μεθοδολογίας DEA, χρησιμοποιήθηκε η γλώσσα προγραμματισμού R και το περιβάλλον RStudio.

Πρώτο βήμα, ήταν η φόρτωση των πακέτων «deaR», «readxl» και «xlsx» και η εισαγωγή της βάσης δεδομένων, που δημιουργήθηκε στο 4ο κεφάλαιο, στην R. Η βάση δεδομένων έχει τη μορφή που φαίνεται στο παρακάτω στιγμιότυπο (Εικόνα 5.1.2).



	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables
1	115842	4786010.0	9	13	4.000000	11
2	115843	4300935.0	10	18	3.000000	21
3	115844	2260105.0	3	4	2.915068	13
4	115890	4064146.0	12	35	2.832877	38
5	115916	16195875.0	9	23	3.498630	29
6	115985	4581967.8	10	15	3.249315	9
7	116020	8210381.0	10	26	1.997260	35
8	116055	7191755.0	13	36	1.997260	26
9	633098	5716971.0	8	8	3.243836	12
10	633127	2103593.8	6	7	3.000000	74
11	633172	8821295.6	17	23	4.000000	93
12	633184	10549121.5	13	22	4.000000	77
13	633192	3157986.0	8	10	3.000000	29
14	633196	4944773.0	6	10	4.000000	42
15	633211	20652921.0	19	67	4.501370	208
16	633212	7271433.8	10	12	4.000000	28
17	633338	3238117.5	5	11	3.000000	28
18	633436	3138121.9	6	10	3.000000	27
19	633464	4998970.0	25	36	4.000000	92
20	633476	3625581.2	14	21	3.000000	31
21	633477	5634810.7	8	19	4.498630	39

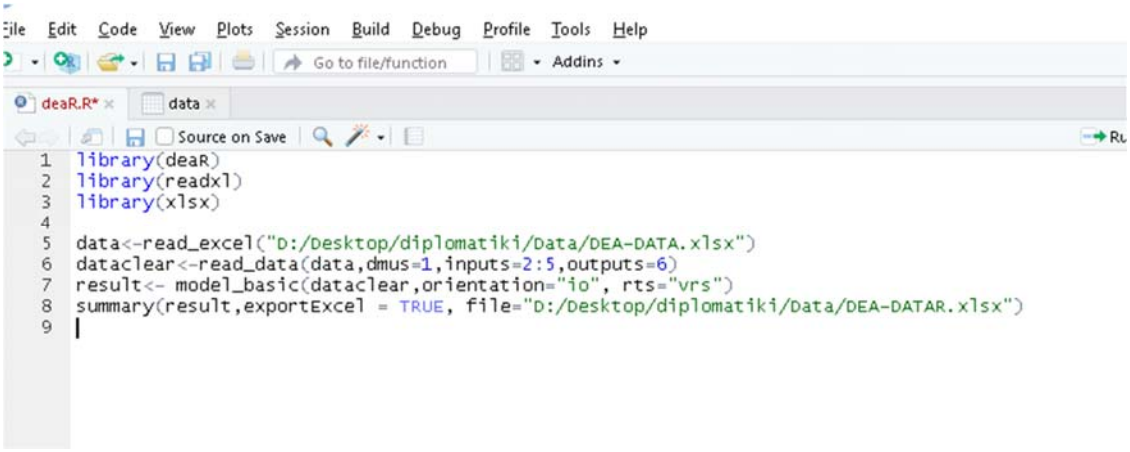
Εικόνα 5.1.2: Στιγμιότυπο βάσης δεδομένων

Όπως φαίνεται από το στιγμιότυπο, τα DMUs βρίσκονται στη 1η στήλη, οι εισροές (inputs) στις στήλες 2-5 και οι εκροές (outputs) στη 5η στήλη.

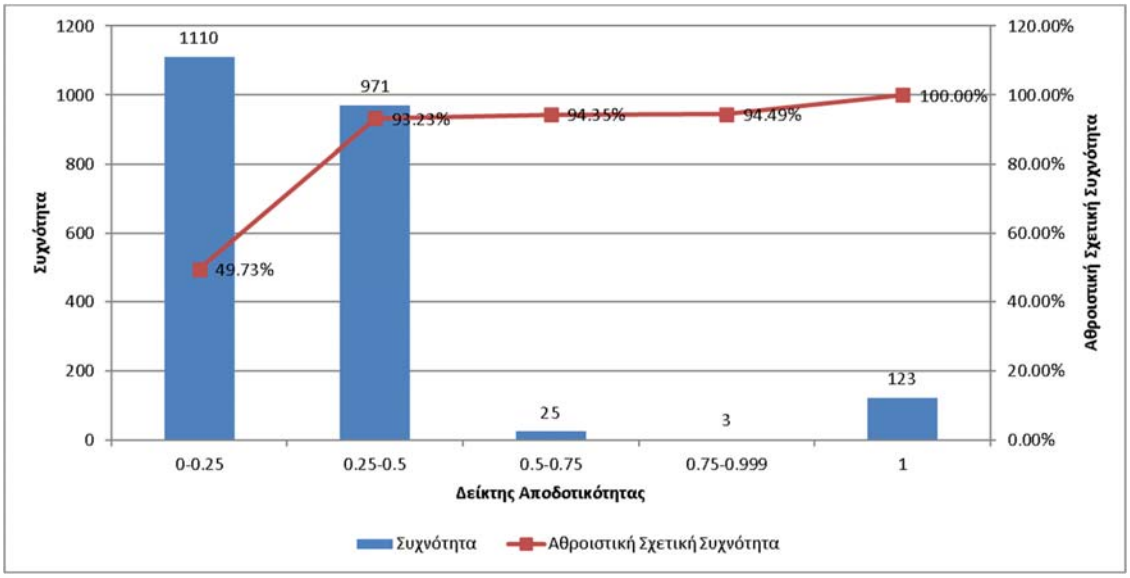
Δεύτερο βήμα, ήταν η ανάγνωση της βάσης δεδομένων από τη συνάρτηση “read_data” η οποία, υποδεικνύοντας τις στήλες των DMUs, των inputs και των outputs, δημιουργεί μια δομή πάνω στην οποία μπορεί να εφαρμοστεί η μεθοδολογία DEA. Με

βάση αυτή τη νέα δομή εφαρμόζεται η μεθοδολογία DEA με προσανατολισμό προς τα δεδομένα εισόδου (orientation="io") και μεταβαλλόμενη απόδοση κλίμακας (rts="vrs").

Τέλος, εξάγονται όλα τα αποτελέσματα της μεθόδου, τα οποία είναι: efficiencies, slacks, lambdas, targets, returns, και references, τα οποία αποθηκεύονται σε ένα αρχείο excel ώστε να είναι δυνατή η αξιοποίησή τους. Παρακάτω φαίνονται οι εντολές που χρησιμοποιήθηκαν για την εφαρμογή της μεθοδολογίας DEA και την εξαγωγή αποτελεσμάτων (Εικόνα 5.1.3) και το διάγραμμα συχνοτήτων των δεικτών αποδοτικότητας (Διάγραμμα 5.1.1).



Εικόνα 5.1.3: Εφαρμογή μεθοδολογίας DEA και εξαγωγή αποτελεσμάτων



Διάγραμμα 5.1.1: Συχνότητα των Δεικτών Αποδοτικότητας των Προγραμμάτων της Ε.Ε

Παρατηρείται, ότι μόνο τα 123 από τα 2232 (περίπου το 5.5%) προγράμματα είναι αποδοτικά, έχουν δηλαδή δείκτη αποδοτικότητας ίσο με 1, ενώ περίπου το 93% των προγραμμάτων έχουν δείκτη αποδοτικότητας μικρότερο ή ίσο του 0.5.

Παρακάτω παρουσιάζεται ένα παράδειγμα προσδιορισμού του αποδοτικού συνόρου ενός δεδομένου εισόδου, όπως παρουσιάστηκε στο κεφάλαιο 3.

Επιλέγεται τυχαία το 1ο DMU με αναγνωριστικό αριθμό έργου (ID) «115842». Με βάση τα αποτελέσματα της μεθόδου, το DMU «115842», έχει δείκτη αποδοτικότητας ίσο με 0,13 και συντελεστές λάμδα λ_{692146}, λ_{696656}, λ_{875289} με τιμές 0.014, 0.001και 0.985 αντίστοιχα. Το αποδοτικό σύνορο των δεδομένων εισόδου για το DMU «115842» υπολογίζεται ως το άθροισμα των γινομένων των λάμδα και των αντίστοιχων πραγματικών τιμών των δεδομένων εισόδου. Οπότε, προκύπτει, παραδείγματος χάρη, το αποδοτικό σύνορο του συνολικού κόστους του 1ου έργου ($totalCost_1$) ως εξής:

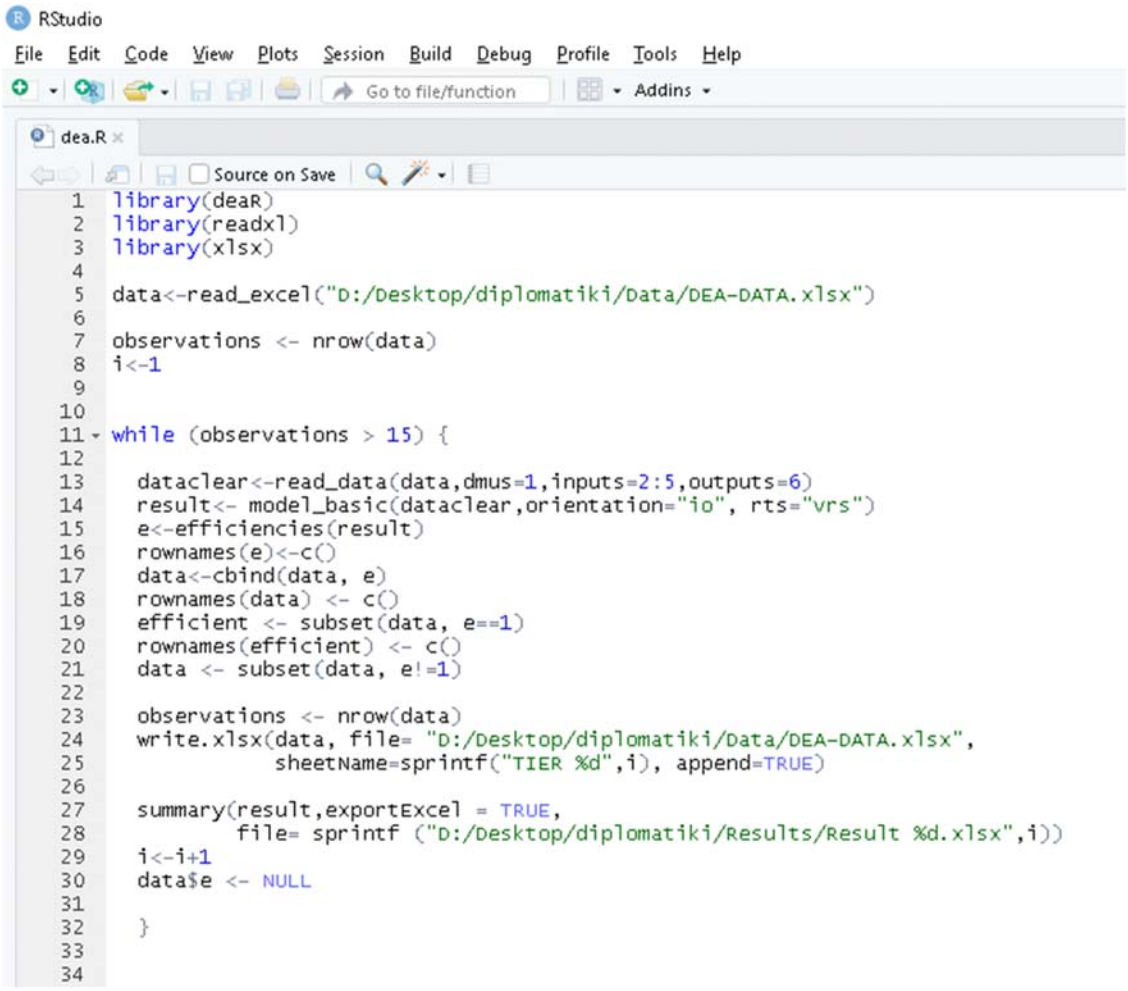
$Αποδοτικό\ Σύνορο\ του\ totalCost_1 = \lambda_{692146} * totalCost_{692146} + \lambda_{696656} * totalCost_{696656} + \lambda_{875289} * totalCost_{875289}$ = *0.014*1,000,000 + 0.001* 89,000,000.11 + 0.985* 129,156.25 = 247,905.9908 €*

Με τον ίδιο τρόπο, μπορεί να βρεθεί το αποδοτικό σύνορο όλων των δεδομένων εισόδου, όλων των DMUs.

5.2. Ανάλυση Επιπέδου Αποδοτικότητας των προγραμμάτων της Ε.Ε.

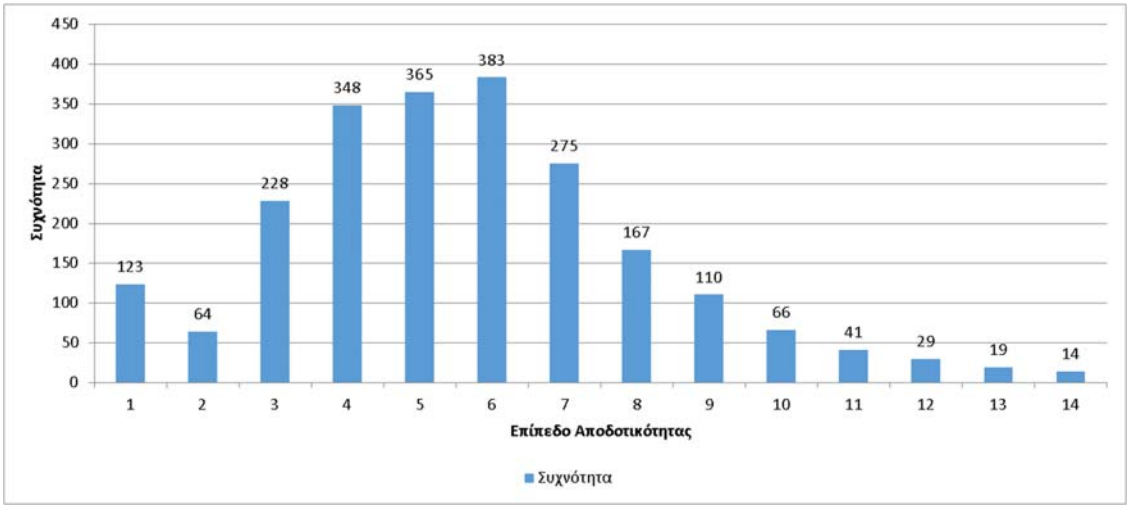
Με την εφαρμογή της μεθοδολογίας DEA προέκυψε ένα σύνολο προγραμμάτων που έχουν δείκτη αποδοτικότητας ίσο με 1 και άρα είναι αποδοτικά. Το σύνολο αυτό θα αποκαλείται Tier 1 (Επίπεδο 1). Στο επόμενο βήμα, εφαρμόζεται πάλι η μέθοδος DEA μόνο με τα μη-αποδοτικά προγράμματα, δηλαδή εκείνα που δεν βρίσκονται στο Tier 1. Με αυτό τον τρόπο προκύπτει ένα σύνολο προγραμμάτων που έχουν δείκτη αποδοτικότητας ίσο με 1 και άρα είναι αποδοτικά. Το σύνολο αυτό θα αποκαλείται Tier 2. Η ίδια διαδικασία επαναλαμβάνεται όσο ο αριθμός των υπολειπόμενων παραγωγικών μονάδων είναι τουλάχιστον τρεις φορές μεγαλύτερος (3x5=15) από το άθροισμα του αριθμού των διαφορετικών μεταβλητών εισροών και εκροών (4+1=5), όπως προτείνεται από τους Banker et al. (1984).

Η παραπάνω διαδικασία μπορεί να αυτοματοποιηθεί με επαναληπτικούς βρόχους στο περιβάλλον προγραμματισμού της R, το RStudio, όπως φαίνεται στη Εικόνα 5.2.1.



Εικόνα 5.2.1: Αυτοματοποιημένη διαδικασία Ανάλυσης Επιπέδου Αποδοτικότητας

Με αυτό τον τρόπο, τα 2232 προγράμματα ομαδοποιήθηκαν σε 14 ομάδες, ανάλογα με το επίπεδο αποδοτικότητας στο οποίο ανήκουν, μέσα από την ανάλυση του επιπέδου αποδοτικότητας. Το πλήθος των προγραμμάτων που ανήκουν σε κάθε επίπεδο αποδοτικότητας φαίνεται στο Διάγραμμα 5.2.1.



Διάγραμμα 5.2.1:Πλήθος Προγραμμάτων της Ε.Ε. ανά Επίπεδο Αποδοτικότητας

Παρατηρείται ότι το μεγαλύτερο πλήθος προγραμμάτων ανήκει στο Επίπεδο Αποδοτικότητας 6 (17% των συνολικών προγραμμάτων).

Στο παρακάτω πίνακα (Πίνακας 5.2.1) φαίνεται ο μέσος όρος κάθε μεταβλητής ανά Επίπεδο Αποδοτικότητας αλλά και ο μέσος όρος κάθε μεταβλητής για το σύνολο των δεδομένων, για να είναι εύκολη η σύγκριση τους. Πιο συγκεκριμένα, με πράσινο φόντο παρουσιάζονται οι μέσοι όροι που είναι χαμηλότεροι από το γενικό μέσο όρο όλων των δεδομένων, ενώ με κόκκινο φόντο παρουσιάζονται οι μέσοι όροι που είναι μεγαλύτεροι από τον γενικό μέσο όρο.

Πίνακας 5.2.1: Μέσοι Όροι Μεταβλητών

Average	totalCost	Number of participant Countries	Total Partners	Duration
Tier 1	2,494,495.0	1.4	3.2	2.2
Tier 2	2,333,057.5	3.0	7.1	3.1
Tier 3	1,929,461.5	3.2	5.7	3.0
Tier 4	3,179,035.9	4.1	8.1	3.2
Tier 5	3,829,809.1	5.2	9.9	3.2
Tier 6	4,544,630.6	6.1	11.9	3.2
Tier 7	5,103,131.8	7.2	13.2	3.2
Tier 8	7,177,655.6	8.4	15.7	3.1
Tier 9	6,988,870.5	9.6	18.0	3.4
Tier 10	8,154,566.1	10.8	20.2	3.3
Tier 11	5,829,122.3	12.0	20.4	3.2
Tier 12	8,151,588.6	13.1	24.1	3.4
Tier 13	7,142,970.4	14.7	29.6	3.3
Tier 14	14,586,677.2	18.1	33.9	3.8
All Data	4,419,875.0	6.0	11.3	3.1

Παρατηρείται ότι για τα Επίπεδα Αποδοτικότητας 1, 2 και 3, όλες οι μεταβλητές έχουν μέσο όρο χαμηλότερο από το γενικό μέσο όρο, γεγονός που τις καθιστά πολύ αποδοτικές.

5.3. Ταξινόμηση των προγραμμάτων με Δένδρα Απόφασης

Με τη διαδικασία Ανάλυσης των Επιπέδων Αποδοτικότητας τα 2232 προγράμματα ομαδοποιήθηκαν σε 14 ομάδες, ανάλογα με το επίπεδο αποδοτικότητας στο οποίο ανήκουν. Οπότε, κάθε πρόγραμμα περιγράφεται από ένα από τα δεκατέσσερα επίπεδα αποδοτικότητας.

Τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας, εισάγονται στον αλγόριθμο C4.5, με σκοπό την ανάπτυξη ενός Δένδρου Ταξινόμησης, με μάθηση με επίβλεψη, το οποίο θα έχει τη δυνατότητα να προβλέπει σε ποιο επίπεδο αποδοτικότητας βρίσκεται οποιοδήποτε υπάρχον ή νέο πρόγραμμα, με βάση τις τιμές των ανεξάρτητων μεταβλητών (χαρακτηριστικών) του κάθε προγράμματος.

Οι ανεξάρτητες μεταβλητές που εισάγονται στον αλγόριθμο C4.5 είναι το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που

συμμετείχαν σε κάθε έργο (Total Partners) και η συνολική διάρκεια του κάθε έργου (Duration).

Η ταξινόμηση με χρήση του αλγόριθμου C4.5, προετοιμάζει ένα σύνολο εκπαιδευτικών περιπτώσεων, καθεμία από τις οποίες περιγράφεται με βάση τα δεδομένα χαρακτηριστικά (ανεξάρτητες μεταβλητές) και μια γνωστή τάξη (Επίπεδο Αποδοτικότητας). Η επαγωγική διαδικασία του C4.5 προσπαθεί να βρει μια μέθοδο ταξινόμησης μιας υπόθεσης, εκφρασμένη ως συνάρτηση των χαρακτηριστικών, που εξηγεί τις περιπτώσεις εκπαίδευσης και που μπορεί επίσης να χρησιμοποιηθεί για την ταξινόμηση άγνωστων περιπτώσεων.

Για την εφαρμογή του αλγόριθμου C4.5, χρησιμοποιήθηκε η γλώσσα προγραμματισμού R και το περιβάλλον RStudio.

Πρώτο βήμα, ήταν η φόρτωση των πακέτων «readxl», «xlsx», «RWeka», «party» και η εισαγωγή της βάσης δεδομένων στην R. Η βάση δεδομένων που χρησιμοποιήθηκε, έχει τη μορφή που φαίνεται στο παρακάτω στιγμιότυπο (Εικόνα 5.3.1).

Filter

	totalCost	Number of participant Countries	Total Partners	Duration	Tier
1	4786010.0	8	13	4.000000	8
2	4300935.0	9	18	3.000000	9
3	2260105.0	3	4	2.915068	3
4	4064146.0	11	35	2.832877	11
5	16195875.0	8	23	3.498630	8
6	4581967.8	9	15	3.249315	9
7	8210381.0	9	26	1.997260	8
8	7191755.0	12	36	1.997260	8
9	5716971.0	7	8	3.243836	7
10	2103593.8	5	7	3.000000	5
11	8821295.6	16	23	4.000000	10
12	10549121.5	12	22	4.000000	11
13	3157986.0	7	10	3.000000	7
14	4944773.0	5	10	4.000000	5
15	20652921.0	18	67	4.501370	6
16	7271433.8	9	12	4.000000	9
17	3238117.5	4	11	3.000000	4
18	3138121.9	5	10	3.000000	5
19	4998970.0	25	36	4.000000	10
20	3625581.2	13	21	3.000000	13
21	5634810.7	7	19	4.498630	7
22	5790111.2	12	17	3.000000	12
23	21237179.5	4	6	2.671233	4
24	3108939.5	6	10	4.000000	6
25	4107405.8	12	23	4.000000	12
26	2999287.5	6	11	3.000000	6

Showing 1 to 27 of 2,232 entries, 5 total columns

Console | Terminal | Jobs

D:/Desktop/diplomatiki/

```
> library(readxl)
> library(xlsx)
> library(RWeka)
> library(party)
> t1 <- read_excel("D:/Desktop/diplomatiki/Results/DATA-DTREE.xlsx")
```

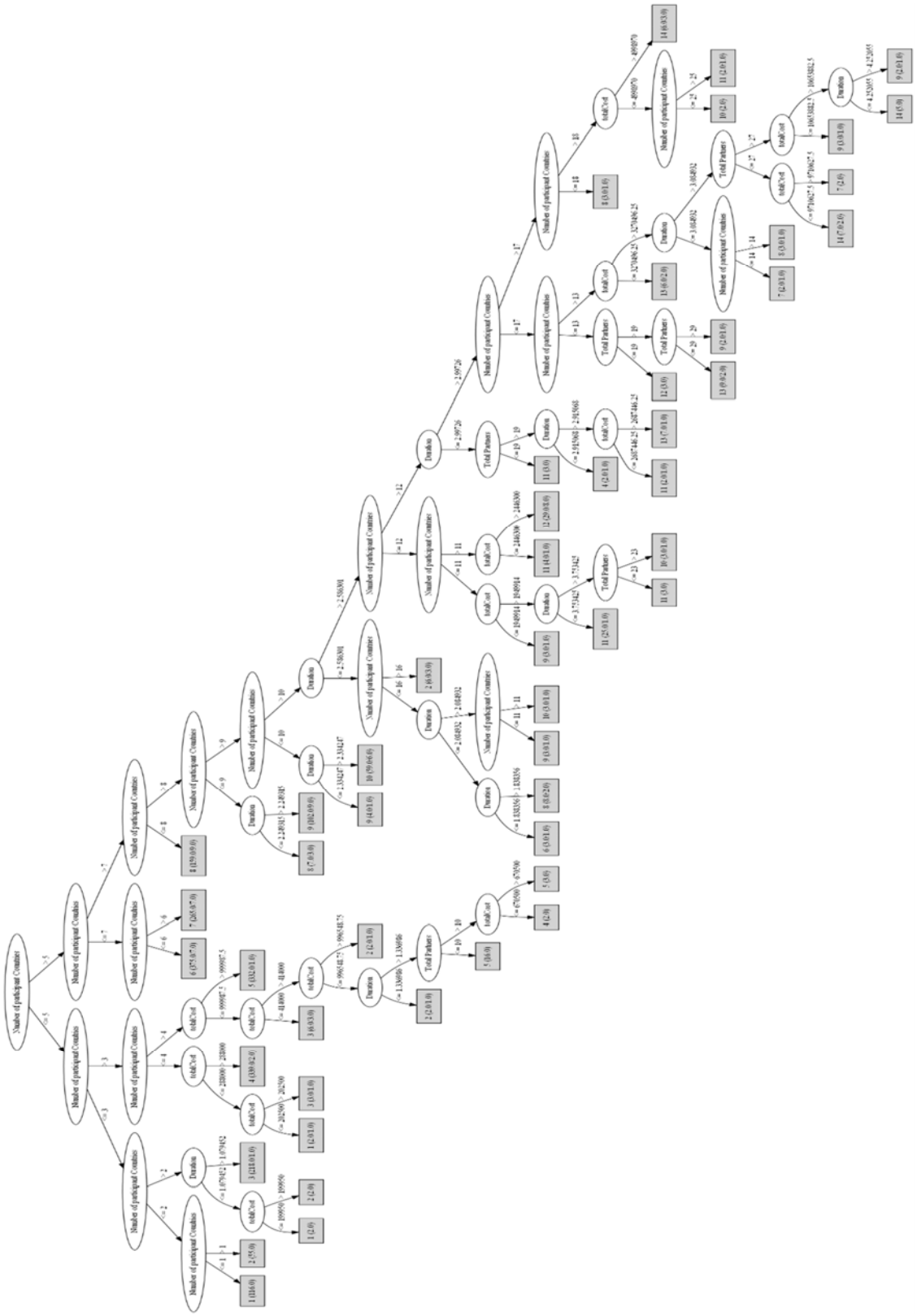
Εικόνα 5.3.1:Στιγμιότυπο βάσης δεδομένων

Όπως φαίνεται από το στιγμιότυπο, οι ανεξάρτητες μεταβλητές βρίσκονται στις στήλες 1-4 και τα Επίπεδα Αποδοτικότητας (Tier) στη στήλη 5.

Δεύτερο βήμα, ήταν η εφαρμογή του αλγορίθμου C4.5, από τη συνάρτηση «J48» (J48 είναι η εφαρμογή του αλγορίθμου C4.5 που αναπτύχθηκε από την ομάδα έργου WEKA) με εξαρτημένη μεταβλητή τα Επίπεδα Αποδοτικότητας (Tiers) και ανεξάρτητες μεταβλητές το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο (Total Partners) και η συνολική διάρκεια του κάθε έργου (Duration).

Τρίτο βήμα, ήταν η οπτικοποίηση του δέντρου απόφασης που προέκυψε από την εφαρμογή της συνάρτησης J48. Με την εντολή «write_to_dot», αναπαραστάθηκε το δέντρο απόφασης, που δημιουργήθηκε, σε γλώσσα DOT, για να είναι δυνατή η επεξεργασία του μέσω Graphviz. Το Graphviz είναι ένα λογισμικό οπτικοποίησης γραφήματος ανοιχτού κώδικα και είναι κατάλληλο για την απεικόνιση δέντρων ταξινόμησης. Αντιγράφοντας, λοιπόν, τον κώδικα, που δόθηκε από την εντολή «write_to_dot» (Εικόνα 5.3.2), δημιουργείται το δενδροδιάγραμμα της Εικόνα 5.3.3, το οποίο έχει τη δυνατότητα να προβλέπει το επίπεδο αποδοτικότητας στο οποίο βρίσκεται οποιοδήποτε υπάρχον ή νέο πρόγραμμα, με βάση τις τιμές των ανεξάρτητων μεταβλητών (χαρακτηριστικών) του κάθε προγράμματος.

Εικόνα 5.3.2: Κώδικας του δέντρου ταξινόμησης σε Περιβάλλον Graphviz



Εικόνα 5.3.3: Δένδρο Ταξινόμησης C4.5

Κάθε κλαδί καταλήγει στο επίπεδο αποδοτικότητας του προγράμματος που βρίσκεται προς εξέταση. Οι αριθμοί στα φύλλα της μορφής της μορφής (N) ή (N/E), είναι το άθροισμα των περιπτώσεων που φτάνουν στο συγκεκριμένο φύλλο και E είναι ο αριθμός των περιπτώσεων που ανήκουν σε διαφορετικό επίπεδο αποδοτικότητας από το καθορισμένο. Παράλληλα, από το δενδροδιάγραμμα (Εικόνα 5.3.3) παρατηρείται ότι ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries) είναι το χαρακτηριστικό το οποίο παρέχει τις περισσότερες πληροφορίες και για το λόγο αυτό έχει επιλεγεί ως πρώτο κριτήριο διαχωρισμού.

Μετά τη δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Προκειμένου να αξιολογηθεί το δέντρο απόφασης χρησιμοποιείται η μέθοδος k-fold cross-validation. Στη μέθοδο k-fold cross-validation το αρχικό δείγμα χωρίζεται τυχαία σε k ίσα μεγέθους υποδείγματα. Από τα υποδείγματα k, διατηρείται ένα μεμονωμένο υποσύστημα ως δεδομένα επικύρωσης για τη δοκιμή του μοντέλου και το υπόλοιπα k-1 υποδείγματα χρησιμοποιούνται ως δεδομένα εκπαίδευσης. Η διαδικασία διασταυρούμενης επικύρωσης στη συνέχεια επαναλαμβάνεται k φορές, με καθένα από τα k υποδείγματα να χρησιμοποιείται ακριβώς μία φορά ως δεδομένα επικύρωσης. Τα αποτελέσματα k μπορούν στη συνέχεια να υπολογιστούν κατά μέσο όρο για να παραχθεί μια ενιαία εκτίμηση. Το πλεονέκτημα αυτής της μεθόδου έναντι της επανειλημμένης τυχαίας υποδειγματοληψίας είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται τόσο για την εκπαίδευση όσο και για την επικύρωση και κάθε παρατήρηση χρησιμοποιείται για την επαλήθευση ακριβώς μία φορά. Για την αξιολόγηση του μοντέλου επιλέχτηκε k=10 και χρησιμοποιήθηκε η συνάρτηση «evaluate_Weka_classifier». Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στο παρακάτω στιγμιότυπο. (Εικόνα 5.3.4).

```
Console   Terminal ×   Jobs ×
D:/Desktop/diplomatiki/
> eval_m1 <- evaluate_Weka_classifier(m1, numFolds = 10, complexity = FALSE,
+                                     seed = 1, class = TRUE)
> eval_m1
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances        2066               92.5627 %
Incorrectly Classified Instances       166                7.4373 %
Kappa statistic                          0.9156
Mean absolute error                      0.0138
Root mean squared error                  0.0938
Relative absolute error                 10.9849 %
Root relative squared error             37.3955 %
Total Number of Instances             2232

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,943    0,003    0,951      0,943    0,947      0,944    0,986     0,973     1
               0,859    0,003    0,887      0,859    0,873      0,869    0,937     0,881     2
               0,952    0,003    0,969      0,952    0,960      0,956    0,975     0,955     3
               0,968    0,004    0,977      0,968    0,973      0,968    0,987     0,969     4
               0,953    0,007    0,961      0,953    0,957      0,949    0,969     0,947     5
               0,966    0,005    0,976      0,966    0,971      0,965    0,982     0,935     6
               0,938    0,006    0,959      0,938    0,949      0,941    0,980     0,931     7
               0,940    0,009    0,892      0,940    0,915      0,909    0,977     0,866     8
               0,864    0,010    0,812      0,864    0,837      0,829    0,956     0,724     9
               0,773    0,008    0,750      0,773    0,761      0,754    0,918     0,650     10
               0,634    0,004    0,765      0,634    0,693      0,691    0,910     0,655     11
               0,759    0,006    0,629      0,759    0,688      0,686    0,875     0,462     12
               0,474    0,006    0,391      0,474    0,429      0,425    0,838     0,285     13
               0,357    0,005    0,313      0,357    0,333      0,330    0,853     0,277     14
Weighted Avg.  0,926    0,006    0,928      0,926    0,927      0,921    0,970     0,899

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   <-- classified as
 116   3   1   2   0   0   0   0   0   0   0   0   1 |   a = 1
   3  55   3   0   2   0   1   0   0   0   0   0   0 |   b = 2
   1   0 217   2   5   1   0   1   1   0   0   0   0 |   c = 3
   0   0   2 337   3   3   0   1   0   0   0   0   1 |   d = 4
   0   3   0   1 348   3   2   2   3   2   0   1   0 |   e = 5
   0   0   0   0   0 370   5   4   1   2   1   0   0 |   f = 6
   0   1   0   0   0   0 258   7   4   1   1   0   2 |   g = 7
   0   0   0   0   1   1   1 157   4   1   0   1   1 |   h = 8
   0   0   0   1   1   0   0   1  95   8   0   0   2 |   i = 9
   0   0   0   0   2   0   1   1   3  51   3   2   1 |   j = 10
   0   0   0   1   0   1   0   0   1   0  26   8   3 |   k = 11
   0   0   1   1   0   0   0   0   0   1   2  22   1 |   l = 12
   0   0   0   0   0   0   0   1   4   1   1   1   9 |   m = 13
   2   0   0   0   0   0   1   1   1   1   0   0   3 |   n = 14
>
```

Εικόνα 5.3.4: Αξιολόγηση μοντέλου J48 (k-fold cross-validation, k=10)

Από τα αποτελέσματα της αξιολόγηση του μοντέλου μεγάλη σημασία έχει:

- ✓ Το ποσοστό των σωστά ταξινομημένων περιπτώσεων και των λάθος ταξινομημένων περιπτώσεων (Correctly Classified Instances, Incorrectly Classified Instances) με τιμές 95.878% και 4.122 %.

- ✓ Ο Πίνακας Σύγχυσης (Confusion Matrix), που είναι ένας πίνακας οπτικοποίησης των προβλέψεων του μοντέλου σε σχέση με τα πραγματικά δεδομένα. Κάθε σειρά του πίνακα σύγχυσης αντιπροσωπεύει τις περιπτώσεις σε μια προβλεπόμενη κλάση και κάθε στήλη αντιπροσωπεύει τις παρουσίες σε μια πραγματική κλάση. Παρατηρείται ότι οι περισσότερες προβλέψεις είναι σωστές.

- ✓ Η Ακρίβεια (Precision), που ορίζεται ως οι αληθινά θετικές περιπτώσεις (True Positive, TP) διαιρεμένες με το άθροισμα των αληθινά θετικών περιπτώσεων και των λανθασμένων θετικών περιπτώσεων (False Positive, FP), έχει τιμή 0.928.

- ✓ Το ROC Area. Το ROC (receiver operating curve) είναι μια καμπύλη που παρουσιάζει το ποσοστό των στοιχείων που προβλέπονται θετικά και είναι πραγματικά θετικά (TPR) έναντι του ποσοστού των στοιχείων που προβλέπονται θετικά ενώ είναι αρνητικά (FPR). Στόχος είναι να προκύψουν μικρές τιμές FP και μεγάλες τιμές TP, να σχηματιστεί δηλαδή ορθή γωνία. Το ROC Area υπολογίζει το εμβαδόν κάτω από την καμπύλη που σχηματίζεται από την καμπύλη ROC και είναι επιθυμητό να προκύπτει μεγαλύτερο από 0.70 για να είναι το μοντέλο αποδεκτό. Στην συγκεκριμένη περίπτωση προκύπτει τιμή 0.970, που σημαίνει ότι το μοντέλο είναι πάρα πολύ καλό.

Από τα παραπάνω προκύπτει ότι το μοντέλο που δημιουργήθηκε είναι αποδεκτό, αφού ικανοποίει όλα τα κριτήρια και έχει πολύ μικρό ποσοστό σφαλμάτων.

Παρακάτω (Εικόνα 5.3.5), φαίνεται ο κώδικας που χρησιμοποιήθηκε για τη διαδικασία της ταξινόμησης και της αξιολόγησης.

```
2  library(readxl)
3  library(xlsx)
4  library(RWeka)
5  library(party)
6
7  t1 <- read_excel("D:/Desktop/diplomatiki/Results/DATA-DTREE.xlsx")
8  m1 <- J48(Tier~., data = t1)
9
10 write_to_dot(m1)
11
12
13
14 #Evaluate Decision Tree
15
16 eval_m1 <- evaluate_Weka_classifier(m1, numFolds = 10, complexity = FALSE,
17                                      seed = 1, class = TRUE)
18 eval_m1
```

Εικόνα 5.3.5:Ταξινόμηση προγραμμάτων και αξιολόγηση του μοντέλου ταξινόμησης

5.4. Ομαδοποίηση των προγραμμάτων με τον αλγόριθμο K-μέσων

Παράλληλα, με τη ταξινόμηση των προγραμμάτων, πραγματοποιήθηκε η ομαδοποίηση των προγραμμάτων χρησιμοποιώντας τον αλγόριθμο K-μέσων (k-means), με μάθηση χωρίς επίβλεψη, χωρίζοντας τα προγράμματα (DMUs) σε διαφορετικές ομάδες (clusters), ανάλογα με τα χαρακτηριστικά του κάθε προγράμματος.

Πρώτο βήμα, ήταν η φόρτωση των πακέτων «tidyverse», «cluster», «factoextra», «ggplot2», «readxl», «xlsx», «rlang», «DMwR» και «gridExtra»και η εισαγωγή της βάσης δεδομένων στην R. Η βάση δεδομένων που χρησιμοποιήθηκε, έχει τη μορφή που φαίνεται στο παρακάτω στιγμιότυπο.

	totalCost	Number of participant Countries	Total Partners	Duration
1	4786010.0	9	13	4.000000
2	4300935.0	10	18	3.000000
3	2260105.0	3	4	2.915068
4	4064146.0	12	35	2.832877
5	16195875.0	9	23	3.498630
6	4581967.8	10	15	3.249315
7	8210381.0	10	26	1.997260
8	7191755.0	13	36	1.997260
9	5716971.0	8	8	3.243836
10	2103593.8	6	7	3.000000
11	8821295.6	17	23	4.000000
12	10549121.5	13	22	4.000000
13	3157986.0	8	10	3.000000
14	4944773.0	6	10	4.000000
15	20652921.0	19	67	4.501370
16	7271433.8	10	12	4.000000
17	3238117.5	5	11	3.000000
18	3138121.9	6	10	3.000000
19	4998970.0	25	36	4.000000
20	3625581.2	14	21	3.000000
21	5634810.7	8	19	4.498630
22	5790111.2	12	17	3.000000
23	21237179.5	5	6	2.671233
24	3108939.5	6	10	4.000000
25	4107405.8	13	23	4.000000
26	2999287.5	7	11	3.000000
27	5918766.3	7	14	4.000000
28	7259113.2	11	17	4.000000

Showing 1 to 29 of 2,232 entries, 4 total columns

Εικόνα 5.4.1: Στιγμιότυπο βάσης δεδομένων

Τα χαρακτηριστικά του κάθε προγράμματος που χρησιμοποιήθηκαν ως δεδομένα εισόδου για το αλγόριθμο k-means, είναι το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο (Total Partners) και η συνολική διάρκεια του κάθε έργου (Duration).

Παρατηρείται πως οι μεταβλητές με βάση τις οποίες θα δημιουργηθούν οι ομάδες, έχουν διαφορετικές κλίμακες. Για το λόγο αυτό, πραγματοποιείται κλιμάκωση των δεδομένων με τη βοήθεια της συνάρτησης «scale». Η κλιμάκωση δεδομένων (επίσης γνωστή ως ομαλοποίηση δεδομένων) είναι η μέθοδος που χρησιμοποιείται για την τυποποίηση του εύρους των χαρακτηριστικών των δεδομένων και βοηθά στην επιτάχυνση των υπολογισμών του αλγορίθμου. Δεδομένου ότι, το εύρος τιμών των δεδομένων μπορεί να ποικίλλει ευρέως, καθίσταται απαραίτητο βήμα στην προεπεξεργασία δεδομένων κατά τη χρήση αλγορίθμων ομαδοποίησης. Η κλιμάκωση επιτυγχάνεται υπολογίζοντας το μέσο όρο και τη τυπική απόκλιση ολόκληρου του διανύσματος, και στη συνέχεια "κλιμακώνεται" κάθε στοιχείο αφαιρώντας το μέσο όρο και διαιρώντας το αποτέλεσμα με τη τυπική απόκλιση. Μετά τη κλιμάκωση, η βάση δεδομένων έχει τη μορφή που φαίνεται στο παρακάτω στιγμιότυπο.

	totalCost	Number of participant Countries	Total Partners	Duration
1	0.0454840206	0.67632637	0.19660727	1.16533439
2	-0.0147756101	0.96785987	0.78961733	-0.19496487
3	-0.2683027187	-1.07287463	-0.87081082	-0.31049714
4	-0.0441913040	1.55092687	2.80585150	-0.42230256
5	1.4629024641	0.67632637	1.38262738	0.48332133
6	0.0201363782	0.96785987	0.43381130	0.14417823
7	0.4708849007	0.96785987	1.73843341	-1.55899098
8	0.3443435893	1.84246037	2.92445351	-1.55899098
9	0.1611349325	0.38479287	-0.39640278	0.13672454
10	-0.2877457123	-0.19827413	-0.51500479	-0.19496487
11	0.5467772592	3.00859437	1.38262738	1.16533439
12	0.7614206712	1.84246037	1.26402537	1.16533439
13	-0.1567612509	0.38479287	-0.15919876	-0.19496487
14	0.0652067434	-0.19827413	-0.15919876	1.16533439
15	2.0165899270	3.59166137	6.60111582	1.84734744
16	0.3542418773	0.96785987	0.07800526	1.16533439
17	-0.1468067189	-0.48980763	-0.04059675	-0.19496487
18	-0.1592289199	-0.19827413	-0.15919876	-0.19496487
19	0.0719394985	5.34086238	2.92445351	1.16533439
20	-0.0986730849	2.13399387	1.14542336	-0.19496487
21	0.1509283653	0.38479287	0.90821934	1.84362059
22	0.1702209593	1.55092687	0.67101532	-0.19496487
23	2.0891708711	-0.48980763	-0.63360680	-0.64218655
24	-0.1628541750	-0.19827413	-0.15919876	1.16533439
25	-0.0388172555	1.84246037	1.38262738	1.16533439
26	-0.1764759610	0.09325937	-0.04059675	-0.19496487

Showing 1 to 29 of 2,232 entries, 4 total columns

Εικόνα 5.4.2: Στιγμιότυπο βάσης δεδομένων μετά τη κλιμάκωση των δεδομένων

Το πρώτο βήμα κατά τη χρήση του αλγόριθμου k-means είναι η υπόδειξη του αριθμού των ομάδων (k) που θα δημιουργηθούν στην τελική λύση. Η ομαδοποίηση με τον αλγόριθμο k-means, μπορεί να πραγματοποιηθεί στην R με τη συνάρτηση «kmeans», εισάγοντας τη βάση δεδομένων και τον αριθμό των ομάδων (centers=k). Η συνάρτηση kmeans έχει επίσης μια επιλογή «nstart» που επιχειρεί πολλές αρχικές διαμορφώσεις και αναφέρει την καλύτερη. Για παράδειγμα, η προσθήκη nstart = 50 θα δημιουργήσει 50 αρχικές διαμορφώσεις. Συνιστάται συχνά αυτή η προσέγγιση.

Επειδή ο αριθμός των ομάδων (k) πρέπει να οριστεί πριν ξεκινήσει ο αλγόριθμος, είναι συχνά ωφέλιμο να χρησιμοποιούνται πολλές διαφορετικές τιμές του k και να εξετάζουμε τις διαφορές στα αποτελέσματα. Εφαρμόζεται, λοιπόν, ο αλγόριθμος για 2, 3, 4 και 5 ομάδες. Στη συνέχεια, αξιολογούνται οι ομαδοποιήσεις ανάλογα με το Dunn Index (DI) και τη μέθοδος του Average Silhouette, απ' όπου προκύπτει ο βέλτιστος αριθμός ομαδοποιήσεων k=4. Τα αποτελέσματα φαίνονται στο παρακάτω στιγμιότυπο (5.4.3).

```
> k4
K-means clustering with 4 clusters of sizes 313, 1385, 526, 8

Cluster means:
    totalCost Number of participant Countries Total Partners   Duration
1  0.49056886                        1.6571084     1.40346798  0.2097136
2 -0.05902987                       -0.1111298    -0.09634397  0.4216117
3 -0.32682298                       -0.7231453    -0.68140837 -1.2196353
4 12.51465157                        1.9517854     6.57146532 -1.0055542
```

Εικόνα 5.4.3: Αποτελέσματα ομαδοποίησης

Με τον αλγόριθμο kmeans, δημιουργήθηκαν 4 ομάδες με 313, 1385, 526 και 8 παρατηρήσεις η κάθε μια. Παρατηρείται ότι η τελευταία ομάδα, αποτελείται από 8 προγράμματα, που με βάση τους μέσους όρους των μεταβλητών της ομάδας, (cluster means) είναι ακραίες (outliers). Ένα ακραίο σημείο ορίζεται ως μια παρατήρηση που βρίσκεται σε ανώμαλη απόσταση από άλλες τιμές σε ένα τυχαίο δείγμα από έναν πληθυσμό. Ο ορισμός του ακραίου σημείου αφήνει στον αναλυτή να αποφασίσει τι θα θεωρηθεί μη φυσιολογικό. Θεωρήθηκαν, λοιπόν, τα 8 αυτά προγράμματα ως ακραία και αφαιρέθηκαν από τη βάση δεδομένων. Τα στοιχεία των προγραμμάτων αυτών, προτού κλιμακωθούν, φαίνονται στο Πίνακας 5.4.1.

Πίνακας 5.4.1:Στοιχεία Ακραίων προγραμμάτων

id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	TIER
662,133.00	90,254,494.23	9.00	39.00	3.17	65.00	9
692,455.00	63,381,009.78	16.00	74.00	3.08	127.00	8
692,480.00	64,237,980.13	6.00	34.00	3.00	89.00	6
696,656.00	89,000,000.11	24.00	186.00	2.00	1,015.00	1
720,270.00	89,000,000.00	19.00	117.00	2.00	267.00	2
807,085.00	241,305,303.98	9.00	30.00	2.00	4.00	8
807,089.00	67,289,236.75	9.00	34.00	2.00	40.00	8
807,090.00	136,809,569.16	7.00	20.00	2.00	15.00	7

Μετά την αφαίρεση των ακραίων τιμών, επαναλαμβάνεται η ίδια διαδικασία με τη νέα βάση δεδομένων. Εισάγεται η νέα βάση δεδομένων στην R, κλιμακώνονται τα δεδομένα και εφαρμόζεται ο αλγόριθμος για 2, 3, 4 και 5 ομάδες. Στη συνέχεια, γίνεται οπτικοποίηση και αξιολόγηση των ομαδοποιήσεων ανάλογα με το Dunn Index (DI) και τη μέθοδος του Average Silhouette. Σημειώνεται ότι η οπτικοποίηση των ομάδων επιτυγχάνεται χρησιμοποιώντας τη συνάρτηση «fviz_cluster» το οποίο παρέχει μια απεικόνιση των ομάδων. Εάν υπάρχουν περισσότερες από δύο διαστάσεις (μεταβλητές), το «fviz_cluster» θα εκτελέσει ανάλυση κύριου παράγοντα (Principal Component Analysis, PCA) και θα σχεδιάσει τα σημεία δεδομένων σύμφωνα με τα δύο πρώτα βασικά στοιχεία που εξηγούν την πλειονότητα της διακύμανσης.

Παρακάτω φαίνεται η τελική κλιμακωμένη βάση δεδομένων με τα 2224 προγράμματα (Εικόνα 5.4.4) που προέκυψαν αφαιρώντας από τα 2232 τα 8 ακραία, ο κώδικας που χρησιμοποιήθηκε (Εικόνα 5.4.5), η οπτικοποίηση των ομάδων για k= 2, 3, 4 και 5 (Εικόνα 5.4.6) καθώς και το διάγραμμα που εμφανίζει το πώς αλλάζει η τιμή του Average Silhouette (Διάγραμμα 5.4.1) ανάλογα με την τιμή του k.

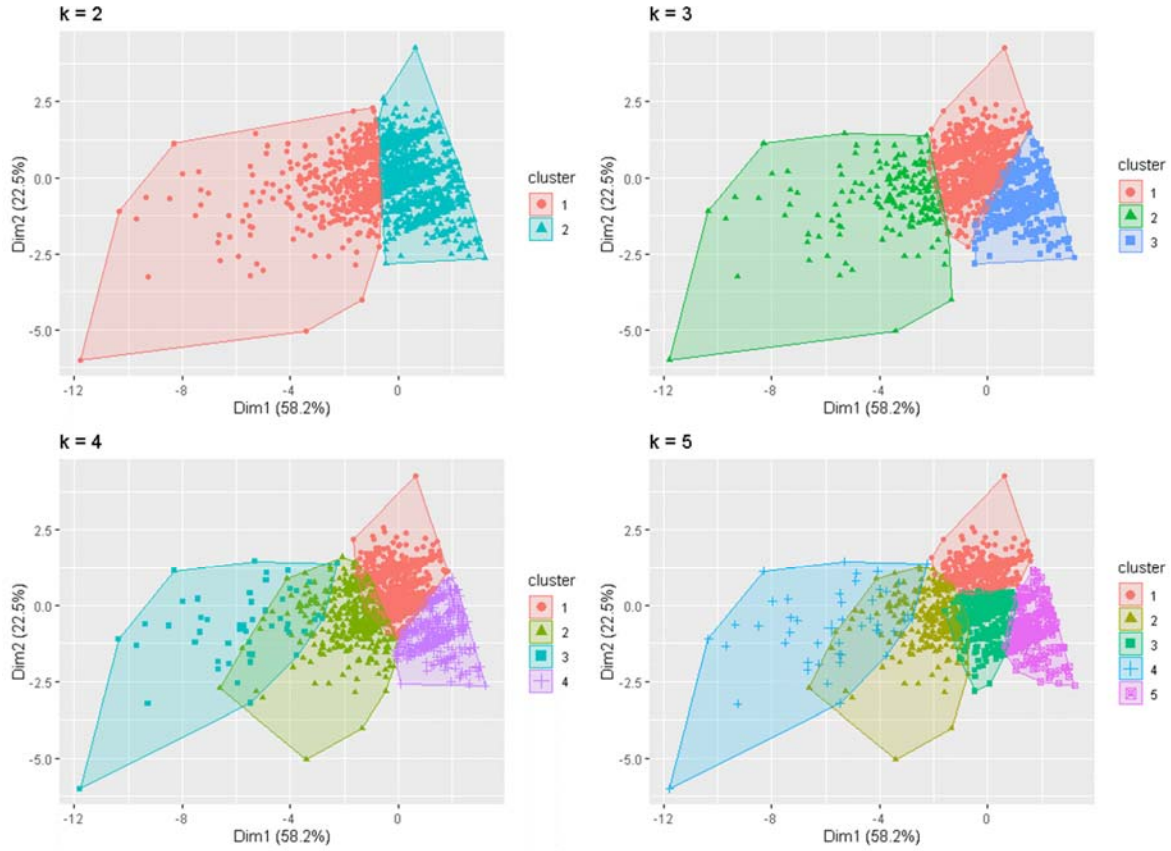
	totalCost	Number of participant Countries	Total Partners	Duration
1	0.1763203967	0.62120603	0.26314697	1.16282383
2	0.0589180364	0.92638876	0.97166861	-0.19877113
3	-0.4350226098	-0.90470762	-1.01219197	-0.31441344
4	0.0016081616	1.53675422	3.38064216	-0.42632535
5	2.9378419793	0.62120603	1.68019024	0.48016115
6	0.1269362069	0.92638876	0.54655563	0.14069501
7	1.0051184958	0.92638876	2.10530322	-1.56409648
8	0.7585811631	1.84193695	3.52234649	-1.56409648
9	0.4016402312	0.31602330	-0.44537466	0.13323422
10	-0.4729029173	-0.29434216	-0.58707899	-0.19877113
11	1.1529777161	3.06266787	1.68019024	1.16282383
12	1.5711622103	1.84193695	1.53848591	1.16282383
13	-0.2177091044	0.31602330	-0.16196601	-0.19877113
14	0.2147456937	-0.29434216	-0.16196601	1.16282383
15	4.0165776807	3.67303333	7.91518062	1.84548651
16	0.7778657548	0.92638876	0.12144265	1.16282383
17	-0.1983149339	-0.59952489	-0.02026168	-0.19877113
18	-0.2225168033	-0.29434216	-0.16196601	-0.19877113
19	0.2278629554	5.80931244	3.52234649	1.16282383
20	-0.1045373556	2.14711968	1.39678159	-0.19877113
21	0.3817550269	0.31602330	1.11337293	1.84175611
22	0.4193423144	1.84193695	0.82996428	-0.19877113
23	4.1579853536	-0.59952489	-0.72878331	-0.64641878
24	-0.2295797990	0.01084057	-0.16196601	1.16282383
25	0.0120782884	1.84193695	1.68019024	1.16282383
26	-0.2561187904	0.01084057	-0.02026168	-0.19877113

Showing 1 to 29 of 2,224 entries, 4 total columns

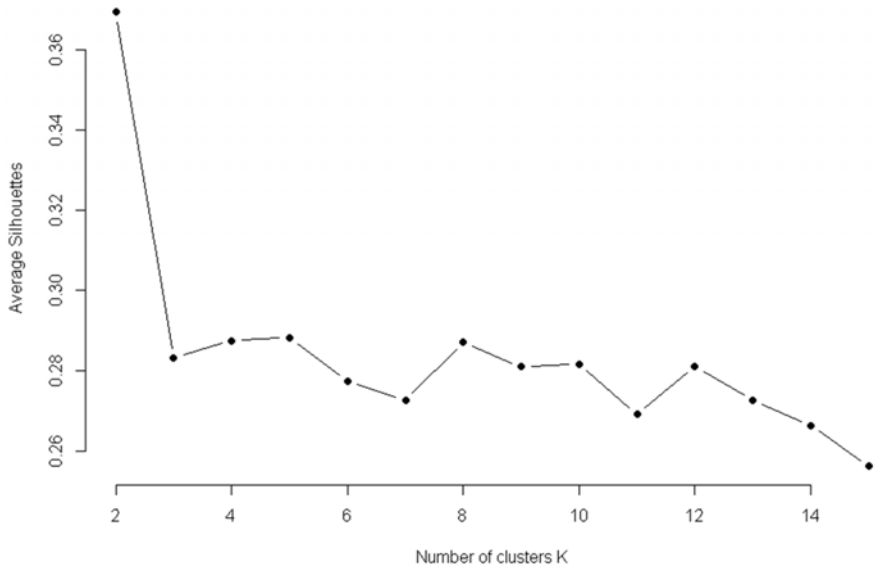
Εικόνα 5.4.4:Στιγμιότυπο τελικής κλιμακώμενης βάσης δεδομένων

```
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(ggplot2)
library(readxl)
library(xlsx)
library(rlang)
library(DMwR)
library(gridExtra)
df<- read_excel("D:/Desktop/diplomatiki/Results/DATA-NN.xlsx")
df <- scale(df)
k2 <- kmeans(df, centers =2, nstart = 50)
k3 <- kmeans(df, centers = 3, nstart = 50)
k4 <- kmeans(df, centers = 4, nstart = 50)
k5 <- kmeans(df, centers = 5, nstart = 50)
p1 <- fviz_cluster(k2, geom = "point", data = df) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = df) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = df) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = df) + ggtitle("k = 5")
grid.arrange(p1, p2, p3, p4, nrow = 2)
k.values <- 1:15
# function to compute average silhouette for k clusters
avg_sil <- function(k) {
  km.res <- kmeans(df, centers = k, nstart = 25)
  ss <- silhouette(km.res$cluster, dist(df))
  mean(ss[, 3])
}
# Compute and plot wss for k = 2 to k = 15
k.values <- 2:15
# extract avg silhouette for 2-15 clusters
avg_sil_values <- map_dbl(k.values, avg_sil)
avg_sil_values
plot(k.values, avg_sil_values,type = "b", pch = 19, frame = FALSE,      xlab = "Number of clusters K",
     ylab = "Average Silhouettes")
fviz_nbclust(df, kmeans, method = "silhouette")
# Compute k-means clustering with k = 4
set.seed(123)
final <- kmeans(df, 4, nstart = 125)
print(final)
fviz_cluster(final, data = df)
```

Εικόνα 5.4.5: Κώδικας για Ομαδοποίηση δεδομένων

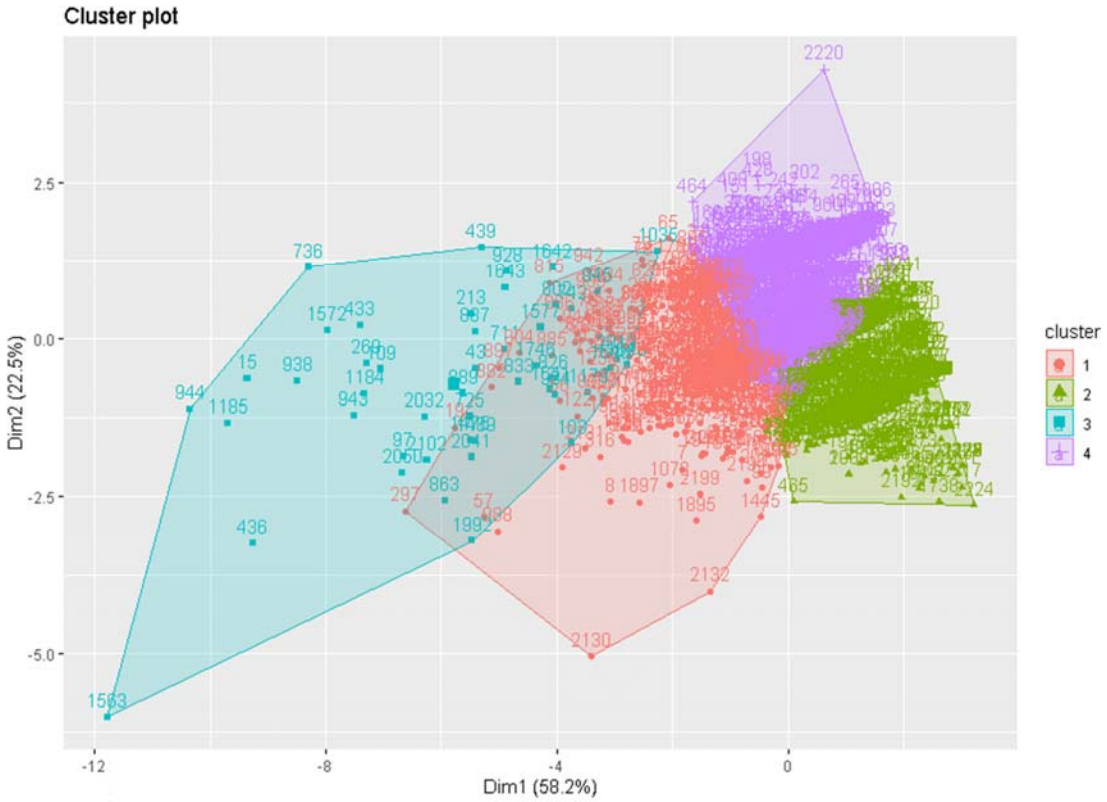


Εικόνα 5.4.6: Απεικόνιση ομάδων για k=1, 2, 3 και 4



Διάγραμμα 5.4.1: Average Silhoutte ανάλογα με τη τιμή του k

Από την οπτικοποίηση και την αξιολόγηση των ομάδων προκύπτει ο βέλτιστος αριθμός ομαδοποιήσεων k=4 με Average Silhouette = 0.288 και τιμή Dunn Index = 0.004. Παρακάτω φαίνεται η απεικόνιση των τεσσάρων ομάδων σύμφωνα με τα δύο πρώτα βασικά στοιχεία που εξηγούν την πλειονότητα της διακύμανσης (Εικόνα 5.4.7) και τα αποτελέσματα του αλγόριθμου k-means για k=4 (Εικόνα 5.4.8), τα οποία συνοψίζουν τα χαρακτηριστικά της κάθε ομάδας.



Εικόνα 5.4.7: Απεικόνιση των τεσσάρων ομάδων (cluster 1, 2, 3, 4)

```
> k4
K-means clustering with 4 clusters of sizes 1216, 418, 47, 543

Cluster means:
   totalCost Number of participant Countries Total Partners   Duration
1 -0.0792098                      -0.1894356     -0.1655785  0.4509223
2  0.4486555                       1.1906858      1.1489685  0.1577763
3  4.7357910                       2.2639982      3.3535073  0.5199256
4 -0.5779025                      -0.6883258     -0.8039414 -1.1762588
```

Εικόνα 5.4.8: Αποτελέσματα του αλγόριθμου k-means για k=4

Άρα, προκύπτουν 4 ομάδες με 1216, 418, 47 και 543 προγράμματα αντίστοιχα. Στο παρακάτω πίνακα (Πίνακας 5.4.2) φαίνεται ο μέσος όρος κάθε μεταβλητής ανά ομάδα (Cluster) αλλά και ο μέσος όρος κάθε μεταβλητής για το σύνολο των δεδομένων, για να είναι εύκολη η σύγκριση τους. Πιο συγκεκριμένα, με πράσινο φόντο παρουσιάζονται οι μέσοι όροι που είναι χαμηλότεροι από το γενικό μέσο όρο όλων των δεδομένων, ενώ με κόκκινο φόντο παρουσιάζονται οι μέσοι όροι που είναι μεγαλύτεροι από τον γενικό μέσο όρο.

Πίνακας 5.4.2:Μέσοι όροι μεταβλητών

Average	totalCost	Number of participant Countries	Total Partners	Duration	Tier
Cluster 1	23,624,517.26	13.38	34.81	3.53	9.49
Cluster 2	5,911,225.44	9.87	19.25	3.26	8.82
Cluster 3	1,669,763.68	3.71	5.47	2.28	3.62
Cluster 4	3,730,227.90	5.34	9.97	3.48	5.31
All Data	4,057,501.51	5.96	11.14	3.15	5.64

Παρατηρείται ότι στην 3η ομάδα (Cluster 3) ανήκουν τα πιο αποδοτικά προγράμματα, αφού όλες οι μεταβλητές έχουν μέσο όρο χαμηλότερο από το γενικό μέσο όρο.

Με κατάλληλη επεξεργασία των αποτελεσμάτων προκύπτει ο τελικός πίνακας των 2224 προγραμμάτων, ο οποίος συγκεντρώνει τα αποτελέσματα και έχει την εξής μορφή (παρουσιάζονται 37 από τα 2224 προγράμματα, Πίνακας 5.4.3).

Πίνακας 5.4.3:Μορφή τελικού πίνακα προγραμμάτων

	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	TIER	CLUSTERS
2	115842	4,786,010.00	8.00	13.00	4.00	11.00	8	4
3	115843	4,300,935.00	9.00	18.00	3.00	21.00	9	2
4	115844	2,260,105.00	3.00	4.00	2.92	13.00	3	3
5	115890	4,064,146.00	11.00	35.00	2.83	38.00	11	2
6	115916	16,195,875.00	8.00	23.00	3.50	29.00	8	2
7	115985	4,581,967.80	9.00	15.00	3.25	9.00	9	2
8	116020	8,210,381.00	9.00	26.00	2.00	35.00	8	2
9	116055	7,191,755.00	12.00	36.00	2.00	26.00	8	2
10	633098	5,716,971.00	7.00	8.00	3.24	12.00	7	4
11	633127	2,103,593.75	5.00	7.00	3.00	74.00	5	4
12	633172	8,821,295.56	16.00	23.00	4.00	93.00	10	2
13	633184	10,549,121.50	12.00	22.00	4.00	77.00	11	2
14	633192	3,157,986.00	7.00	10.00	3.00	29.00	7	4
15	633196	4,944,773.00	5.00	10.00	4.00	42.00	5	4
16	633211	20,652,921.00	18.00	67.00	4.50	208.00	6	1
17	633212	7,271,433.75	9.00	12.00	4.00	28.00	9	2
18	633338	3,238,117.50	4.00	11.00	3.00	28.00	4	4
19	633436	3,138,121.88	5.00	10.00	3.00	27.00	5	4
20	633464	4,998,970.00	25.00	36.00	4.00	92.00	10	2
21	633476	3,625,581.25	13.00	21.00	3.00	31.00	13	2
22	633477	5,634,810.68	7.00	19.00	4.50	39.00	7	2
23	633485	5,790,111.25	12.00	17.00	3.00	50.00	12	2
24	633545	21,237,179.50	4.00	6.00	2.67	3.00	4	4
25	633567	3,108,939.48	6.00	10.00	4.00	44.00	6	4
26	633571	4,107,405.75	12.00	23.00	4.00	57.00	12	2
27	633592	2,999,287.50	6.00	11.00	3.00	27.00	6	4
28	633595	5,918,766.27	6.00	14.00	4.00	233.00	4	4
29	633666	7,259,113.16	10.00	17.00	4.00	53.00	10	2
30	633680	5,551,125.25	13.00	31.00	4.00	78.00	11	2
31	633692	5,299,993.64	12.00	16.00	4.00	68.00	11	2
32	633724	2,993,175.00	11.00	14.00	3.42	14.00	11	2
33	633776	5,708,000.00	4.00	13.00	4.00	17.00	4	4
34	633780	5,588,101.25	8.00	11.00	4.00	12.00	8	4
35	633814	3,007,800.00	11.00	15.00	3.00	44.00	11	2
36	633838	2,991,436.25	13.00	14.00	3.00	28.00	12	2
37	633937	6,026,455.00	7.00	13.00	3.50	30.00	7	4

5.5. Σταδιακή βελτιστοποίηση των μη αποδοτικών DMUs

Με την ανάλυση επιπέδων αποδοτικότητας, τα 2224 προγράμματα χωρίστηκαν σε 14 Επίπεδα (Tier), ανάλογα με το δείκτη της απόδοσης τους. Επίσης, με τον αλγόριθμο των κ-μέσων (k-means) τα προγράμματα ομαδοποιήθηκαν σε 4 ομάδες (clusters ανάλογα) με τα χαρακτηριστικά τους. Συνδυάζοντας τα αποτελέσματα των δύο διαδικασιών, προσδιορίζεται το σύνολο αναφοράς των μη αποδοτικών προγραμμάτων, με σκοπό να βελτιώσουν την αποτελεσματικότητας μέσω εύρεσης ενός προγράμματος αναφοράς στο ανώτερο δυνατό Επίπεδο (Tier) που μοιράζεται παρόμοια χαρακτηριστικά, όπως έχει οριστεί από τον αλγόριθμο k-means. Έτσι, τα αποδοτικά προγράμματα στο ανώτερο Tier ενός cluster, αποτελούν το σύνολο αναφοράς για τα μη αποδοτικά προγράμματα των χαμηλότερων Tiers, που βρίσκονται στο ίδιο cluster.

Για να προσδιοριστεί η καλύτερη διαδρομή για τη σταδιακή βελτιστοποίηση (improvement path) των μη αποδοτικών προγραμμάτων, χρησιμοποιείται η μικρότερη Ευκλείδεια απόσταση ενός μη αποδοτικού προγράμματος με ένα πρόγραμμα που

βρίσκεται στο αμέσως μεγαλύτερο Tier, ενώ παράλληλα βρίσκεται και στο ίδιο cluster και άρα μοιράζεται παρόμοια χαρακτηριστικά.

Η διαδικασία προσδιορισμού της καλύτερης διαδρομής για τη σταδιακή βελτιστοποίηση πραγματοποιήθηκε με τα εξής βήματα.

Αρχικά, χρησιμοποιήθηκε η συνάρτηση «dist» στο πρόγραμμα R, εξάγονται 4 μητρικοί πίνακες αποστάσεων, ένας για κάθε ομάδα (cluster), με βάση την Ευκλείδεια απόσταση των δεδομένων. Σημειώνεται ότι η συνάρτηση «dist» υπολογίζει και επιστρέφει τον πίνακα απόστασης που υπολογίζεται χρησιμοποιώντας το καθορισμένο μέτρο απόστασης για τον υπολογισμό των αποστάσεων μεταξύ των σειρών μιας μήτρας δεδομένων.

Στη συνέχεια, με βάση τον μητρικό πίνακα αποστάσεων της κάθε ομάδας (cluster), βρέθηκε για κάθε μη αποδοτικό πρόγραμμα, το πρόγραμμα που βρίσκεται στην ίδια ομάδα, στο αμέσως μεγαλύτερο Tier και έχει τη μικρότερη Ευκλείδεια απόσταση από εκείνο. Το πρόγραμμα αυτό αποτελεί το πρόγραμμα αναφοράς του μη αποδοτικού προγράμματος. Στην Εικόνα 5.5.1 φαίνεται η εντολή που συντάχτηκε στο πρόγραμμα excel και επιστρέφει το πρόγραμμα αναφοράς, με βάση το μητρικό πίνακα αποστάσεων της κάθε ομάδας και το επίπεδο (tier) στο οποίο βρίσκεται το πρόγραμμα. Η εντολή αυτή έχει σκοπό την αυτοματοποίηση της παραπάνω διαδικασίας.

```
=INDEX('CLUSTER 2 MATRIX'!$B$2:$PC$2;MATCH(MIN(FILTER(INDEX('CLUSTER 2 MATRIX'!$B$3:$PC$420;MATCH(A2;
'CLUSTER 2 MATRIX'!$A$3:$A$420;0););'CLUSTER 2 MATRIX'!$B$1:$PC$1=SMALL([Tier];COUNTIF([Tier];"<"&[@Tier]))
;));INDEX('CLUSTER 2 MATRIX'!$B$3:$PC$420;MATCH([@id];'CLUSTER 2 MATRIX'!$A$3:$A$420;0););0))
```

Εικόνα 5.5.1: Εντολή εύρεσης προγράμματος αναφοράς

Έπειτα, εάν το πρόγραμμα αναφοράς δεν βρίσκεται στο ανώτερο δυνατό Επίπεδο της ομάδας (cluster), αποδίδεται σε αυτό, ένα άλλο πρόγραμμα αναφοράς το οποίο βρίσκεται στην ίδια ομάδα, στο αμέσως μεγαλύτερο Tier και έχει τη μικρότερη Ευκλείδεια απόσταση από εκείνο.

Η ίδια διαδικασία επαναλαμβάνεται μέχρις ότου να μην υπάρχει πρόγραμμα μέσα στην ομάδα (cluster) που βρίσκεται σε μεγαλύτερο Tier, από το πρόγραμμα αναφοράς.

Στο παρακάτω πίνακα φαίνεται η μορφή του πίνακα (Πίνακας 5.5.1) της 2ης ομάδας (Cluster 2), με τον αναγνωριστικό αριθμό του έργου (ID), το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο (Total Partners), τη συνολική διάρκεια του κάθε έργου (Duration), το άθροισμα των παραδοτέων με τις δημοσιεύσεις (Publications and Deliverables), το Επίπεδο Αποδοτικότητας (TIER) στο οποίο ανήκει καθώς και το πρόγραμμα αναφοράς (REF) κάθε έργου.

Πίνακας 5.5.1: Ομάδα 2 (37 από τα 418 προγράμματα της ομάδας)

	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	TIER	REF
2	115843	4,300,935.00	9.00	18.00	3.00	21.00	9	740712
3	115890	4,064,146.00	11.00	35.00	2.83	38.00	11	665220
4	115916	16,195,875.00	8.00	23.00	3.50	29.00	8	646531
5	115985	4,581,967.80	9.00	15.00	3.25	9.00	9	636494
6	116020	8,210,381.00	9.00	26.00	2.00	35.00	8	671596
7	116055	7,191,755.00	12.00	36.00	2.00	26.00	8	637107
8	633172	8,821,295.56	16.00	23.00	4.00	93.00	10	636202
9	633184	10,549,121.50	12.00	22.00	4.00	77.00	11	689450
10	633212	7,271,433.75	9.00	12.00	4.00	28.00	9	645198
11	633464	4,998,970.00	25.00	36.00	4.00	92.00	10	700416
12	633476	3,625,581.25	13.00	21.00	3.00	31.00	13	709637
13	633477	5,634,810.68	7.00	19.00	4.50	39.00	7	690772
14	633485	5,790,111.25	12.00	17.00	3.00	50.00	12	653618
15	633571	4,107,405.75	12.00	23.00	4.00	57.00	12	115890
16	633666	7,259,113.16	10.00	17.00	4.00	53.00	10	633212
17	633680	5,551,125.25	13.00	31.00	4.00	78.00	11	653811
18	633692	5,299,993.64	12.00	16.00	4.00	68.00	11	690199
19	633724	2,993,175.00	11.00	14.00	3.42	14.00	11	689687
20	633814	3,007,800.00	11.00	15.00	3.00	44.00	11	689687
21	633838	2,991,436.25	13.00	14.00	3.00	28.00	12	633724
22	633945	7,966,697.00	12.00	22.00	3.00	75.00	11	680708
23	634144	5,888,487.50	9.00	15.00	4.00	21.00	9	637232
24	634149	6,931,978.75	9.00	18.00	3.50	51.00	9	644866
25	634179	6,636,039.50	10.00	23.00	4.00	196.00	6	634495
26	634201	6,962,265.00	13.00	21.00	3.50	44.00	13	642317
27	634446	7,520,005.00	16.00	31.00	3.50	56.00	13	723986
28	634453	3,375,330.25	27.00	33.00	3.00	27.00	14	700688
29	634476	3,395,987.00	9.00	25.00	4.00	76.00	9	730349
30	634486	7,396,689.65	13.00	27.00	4.00	45.00	13	723986
31	634495	6,239,622.38	10.00	16.00	4.00	201.00	5	671650
32	634534	2,993,888.00	7.00	17.00	3.00	17.00	7	692976
33	634561	4,372,015.25	8.00	16.00	3.50	34.00	8	657466
34	634588	7,651,315.75	9.00	17.00	4.75	32.00	9	646155
35	635188	5,207,821.75	11.00	24.00	3.00	28.00	11	689592
36	635201	5,307,551.25	17.00	22.00	4.50	41.00	14	636427
37	635359	7,791,810.00	9.00	23.00	4.50	29.00	9	713794

Με βάση τα προηγούμενα αποτελέσματα που προσδιορίζουν τις αναφορές συγκριτικής αξιολόγησης κάθε προγράμματος σε κάθε επίπεδο, προσδιορίζεται η σταδιακή πορεία βελτίωσης για τα προγράμματα κάθε επιπέδου εκτός από το επίπεδο 1.

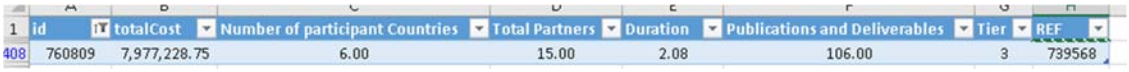
Παρουσιάζεται ένα παράδειγμα καλύτερης διαδρομής για τη σταδιακή βελτιστοποίηση (improvement path) ενός μη αποδοτικού προγράμματος. Το πρόγραμμα με ID «634495» βρίσκεται στη γραμμή 31της 2ης ομάδας (Cluster 2), όπως φαίνεται από το παραπάνω πίνακα, και βρίσκεται στο Επίπεδο 5. Το πρόγραμμα με ID «634495» έχει ως πρόγραμμα αναφοράς το πρόγραμμα με ID «671650», το οποίο έχει τα χαρακτηριστικά που φαίνονται στο Πίνακας 5.5.2.

Πίνακας 5.5.2: Χαρακτηριστικά προγράμματος με ID «671650»

	A	B	C	D	E	F	G	H
1	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	Tier	REF
223	671650	8,165,085.00	8.00	19.00	2.00	94.00	4	760809

Το πρόγραμμα με ID «671650», βρίσκεται στο Επίπεδο 4 και έχει ως πρόγραμμα αναφοράς το πρόγραμμα με ID «760809» (Πίνακας 5.5.3).

Πίνακας 5.5.3: Χαρακτηριστικά προγράμματος με ID «760809»



	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	Tier	REF
408	760809	7,977,228.75	6.00	15.00	2.08	106.00	3	739568

Το πρόγραμμα με ID «760809», βρίσκεται στο Επίπεδο 3 και έχει ως πρόγραμμα αναφοράς το πρόγραμμα με ID «739568» (Πίνακας 5.5.4).

Πίνακας 5.5.4: Χαρακτηριστικά προγράμματος με ID «739568»

	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	Tier	REF
405	739568	1,959,538.25	17.00	18.00	1.00	48.00	2	654109

Το πρόγραμμα με ID «739568», βρίσκεται στο Επίπεδο 3 και έχει ως πρόγραμμα αναφοράς το πρόγραμμα με ID «654109» (Πίνακας 5.5.5).

Πίνακας 5.5.5: Χαρακτηριστικά προγράμματος με ID «654109»

	id	totalCost	Number of participant Countries	Total Partners	Duration	Publications and Deliverables	Tier	REF
180	654109	10,126,484.54	20.00	31.00	4.00	476.00	1	#NUM!

Το πρόγραμμα με ID «654109», βρίσκεται στο Επίπεδο 1 και δεν έχει πρόγραμμα αναφοράς, αφού δεν υπάρχει πρόγραμμα μέσα στην ομάδα (cluster) που βρίσκεται σε μεγαλύτερο Tier. Για το λόγο αυτό, το πρόγραμμα excel δίνει το αποτέλεσμα «#NUM!».

Άρα η καλύτερη διαδρομή σταδιακής βελτίωσης του προγράμματος με ID «634495» είναι:

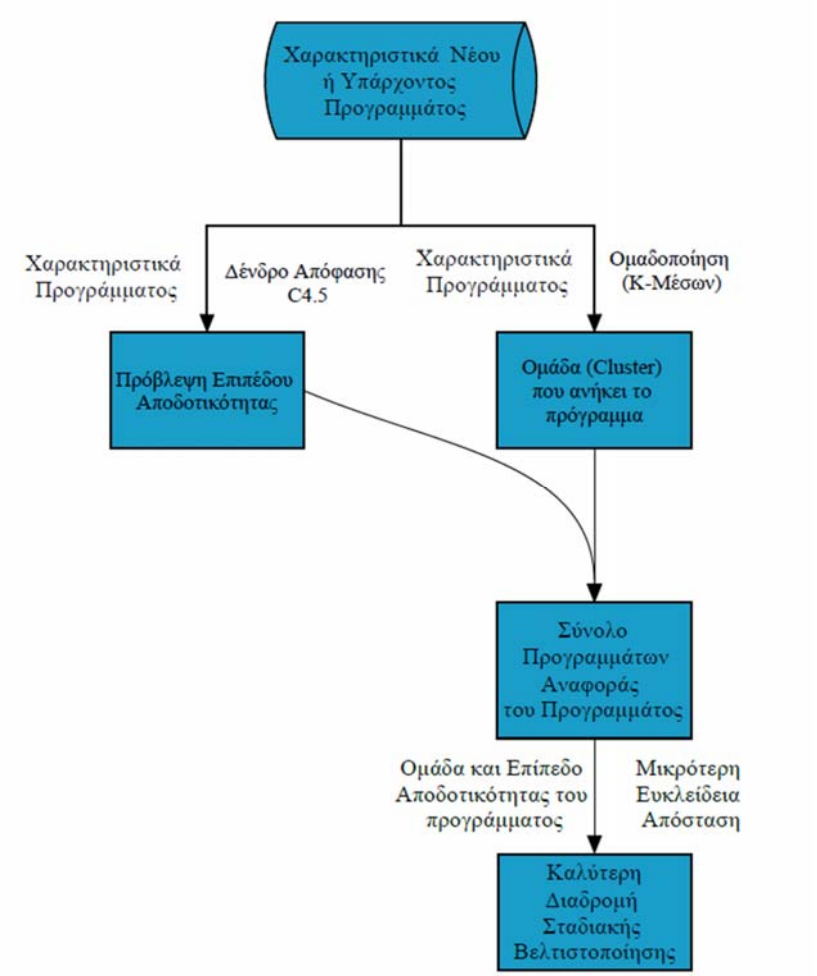
$634495_{\text{Tier 5}} \rightarrow 671650_{\text{Tier 4}} \rightarrow 760809_{\text{Tier 3}} \rightarrow 739568_{\text{Tier 2}} \rightarrow 654109_{\text{Tier 1}}$

5.6. Σημασία στη διαχείριση των έργων

Με βάση τη μεθοδολογία που αναπτύχθηκε μπορούν να αναπτυχθούν στρατηγικές μέθοδοι για τη σταδιακή βελτιστοποίηση των προγραμμάτων που είναι ήδη σε εξέλιξη, αλλά και για τη βελτιστοποίηση μελλοντικών προγραμμάτων, αξιοποιώντας κατάλληλα τους διατιθέμενους πόρους.

Για να αξιοποιηθεί η μεθοδολογία που αναπτύχθηκε, είναι απαραίτητα τα στοιχεία των πόρων που χρησιμοποιούνται για την παραγωγή των εκροών, του υπάρχοντος προγράμματος ή οι πόροι που έχουν προγραμματιστεί να χρησιμοποιηθούν για την παραγωγή των εκροών από το νέο πρόγραμμα. Οι πόροι αυτοί, αποτελούν τα χαρακτηριστικά του προγράμματος, και τα δεδομένα εισόδου της μεθοδολογίας. Στη συνέχεια, προβλέπεται το επίπεδο αποδοτικότητας του προγράμματος από το δένδρο απόφασης C4.5 που αναπτύχθηκε, με βάση τα χαρακτηριστικά του προγράμματος. Παράλληλά, προσδιορίζεται σε ποια από τις 4 ομάδες ανήκει το πρόγραμμα με βάση τον αλγόριθμο κ-μέσων που αναπτύχθηκε. Ανάλογα με το επίπεδο αποδοτικότητας και την

ομάδα στην οποία ανήκει προσδιορίζεται το σύνολο των προγραμμάτων αναφοράς του προγράμματος. Τέλος, με βάση τη μικρότερη ευκλείδεια απόσταση, προσδιορίζεται η καλύτερη διαδρομή σταδιακής βελτιστοποίησης του προγράμματος. Στο παρακάτω διάγραμμα (Διάγραμμα 5.6.1), φαίνεται η ροή των εργασιών που περιγράφηκαν.



Διάγραμμα 5.6.1: Ροή Εργασιών

6. Συμπεράσματα

6.1. Σύνοψη μεθοδολογίας και αποτελεσμάτων

Η αξιολόγηση έργων Ε&Α είναι ένα σημαντικό βήμα στη λήψη αποφάσεων Ε&Α των οργανισμών. Η επιλογή των κατάλληλων μεθόδων αξιολόγησης θεωρείται κρίσιμη για τους υπεύθυνους λήψης αποφάσεων Ε&Α, αφού η εφαρμογή διαφορετικών τεχνικών μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα αξιολόγησης και συνεπώς σε διαφορετικές αποφάσεις Ε&Α.

Στην παρούσα διπλωματική εργασία αναπτύσσεται μια μεθοδολογία αξιολόγησης των 2232 προγραμμάτων έρευνας και ανάπτυξης που έχουν χρηματοδοτηθεί από την Ευρωπαϊκή Ένωση τα τελευταία έξι χρόνια (Ιανουάριος του 2014 μέχρι Δεκέμβριος του 2019). Παράλληλα, προτείνονται στρατηγικές μέθοδοι για τη σταδιακή βελτιστοποίηση των προγραμμάτων που είναι ήδη σε εξέλιξη, αλλά και για τη βελτιστοποίηση μελλοντικών προγραμμάτων, με τη πρόβλεψη του επιπέδου αποδοτικότητας τους.

Για τη μεθοδολογία που αναπτύχθηκε, χρησιμοποιήθηκε η μέθοδος Περιβάλλουσας Ανάλυσης Δεδομένων (DEA) σε συνδυασμό με μοντέλα μηχανικής μάθησης (machine learning), με τη βοήθεια του προγραμματιστικού περιβάλλοντος R.

Αρχικά, εφαρμόστηκε η μέθοδος DEA στα 2232 προγράμματα Ε&Α, που συλλέχτηκαν και επεξεργάστηκαν από το CORDIS (Community Research and Development Information Service - Κοινοτική Υπηρεσία Πληροφοριών Έρευνας και Ανάπτυξης). Ως εισροές για το μοντέλο της DEA τέθηκαν: το συνολικό κόστος του έργου (totalCost), ο αριθμός των διαφορετικών χωρών που συμμετείχαν σε κάθε έργο (Number of participant Countries), το άθροισμα των οργανισμών που συμμετείχαν σε κάθε έργο (Total Partners), η συνολική διάρκεια του κάθε έργου (Duration). Ως εκροές για το μοντέλο της DEA τέθηκε το άθροισμα των παραδοτέων με τις δημοσιεύσεις (Publications and Deliverables). Για το συγκεκριμένο μοντέλο επιλέχτηκε προσανατολισμός προς τα δεδομένα εισόδου (input-oriented) και μεταβαλλόμενη απόδοση κλίμακας (Variable Returns to Scale, VRS). Μέσα από αυτή τη μέθοδο προέκυψε ένα σύνολο προγραμμάτων που έχουν δείκτη αποδοτικότητας ίσο με 1 και άρα είναι αποδοτικά. Αυτό το σύνολο αποκαλείται Tier 1 (Επίπεδο 1). Μόνο τα 123 από τα 2232 (περίπου το 5.5%) προγράμματα είναι αποδοτικά, έχουν δηλαδή δείκτη αποδοτικότητας ίσο με 1, ενώ περίπου το 93% των προγραμμάτων έχουν δείκτη αποδοτικότητας μικρότερο ή ίσο του 0.5.

Έπειτα, εφαρμόστηκε πάλι η μέθοδος DEA μόνο με τα μη-αποδοτικά απ' όπου προέκυψε το Tier 2. Η ίδια διαδικασία επαναλήφθηκε με επαναληπτικούς βρόχους στο περιβάλλον προγραμματισμού της R μέχρι ο αριθμός των υπολειπόμενων παραγωγικών μονάδων να είναι τουλάχιστον τρεις φορές μεγαλύτερος από το άθροισμα του αριθμού

των διαφορετικών μεταβλητών εισροών και εκροών. Μέσα από την ανάλυση του επιπέδου αποδοτικότητας, τα 2232 προγράμματα ομαδοποιήθηκαν σε 14 ομάδες, ανάλογα με το επίπεδο αποδοτικότητας στο οποίο ανήκουν. Παρατηρήθηκε ότι το μεγαλύτερο πλήθος προγραμμάτων ανήκει στο Επίπεδο Αποδοτικότητας 6 (17% των συνολικών προγραμμάτων) και ότι στα Επίπεδα Αποδοτικότητας 1, 2 και 3 ανήκουν τα πιο αποδοτικά προγράμματα.

Στη συνέχεια, αναπτύχθηκε ένα Δένδρο Ταξινόμησης, με μάθηση με επίβλεψη, εισάγοντας τα αποτελέσματα της διαδικασίας ανάλυσης επιπέδων αποδοτικότητας ως δεδομένα εισόδου. Το Δένδρο Ταξινόμησης που αναπτύχθηκε έχει τη δυνατότητα να προβλέπει το επίπεδο αποδοτικότητας στο οποίο ανήκει οποιοδήποτε υπάρχον ή νέο πρόγραμμα, με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών). Το μοντέλο αξιολογήθηκε με τη μέθοδο k-fold cross-validation (k=10), απ' όπου προέκυψαν οι τιμές του ποσοστού των σωστά ταξινομημένων περιπτώσεων = 95.878% Ακρίβειας = 0.928 και ROC Area = 0.97. Άρα το μοντέλο είναι αποδεκτό και μπορεί να χρησιμοποιηθεί για τη πρόβλεψη του επιπέδου αποδοτικότητας ενός προγράμματος E&A.

Μετέπειτα, πραγματοποιήθηκε η ομαδοποίηση των προγραμμάτων, χρησιμοποιώντας τον αλγόριθμο K-μέσων (k-means), με μάθηση χωρίς επίβλεψη, χωρίζοντας τα προγράμματα σε διαφορετικές ομάδες (clusters), ανάλογα με τα χαρακτηριστικά του κάθε προγράμματος. Κατά την ομαδοποίηση, 8 από τα 2232 προγράμματα θεωρήθηκαν ως ακραία (outliers) και αφαιρέθηκαν από τη βάση δεδομένων. Η νέα βάση δεδομένων των 2224 προγραμμάτων ομαδοποιήθηκε εξαρχής. Από την οπτικοποίηση και την αξιολόγηση των ομάδων προέκυψε ο βέλτιστος αριθμός ομαδοποιήσεων k=4 με Average Silhouette = 0.288 και τιμή Dunn Index = 0.004. Στις 4 ομάδες ανήκουν 1216, 418, 47 και 543 προγράμματα αντίστοιχα.

Τέλος, κάθε μη αποδοτικό πρόγραμμα αντιστοιχίστηκε, με βάση τη μικρότερη Ευκλείδεια απόσταση, με ένα πρόγραμμα που βρίσκεται στο αμέσως μεγαλύτερο Tier του ίδιου cluster, που μοιράζεται, δηλαδή, παρόμοια χαρακτηριστικά. Έτσι, προσδιορίστηκε η καλύτερη διαδρομή για τη σταδιακή βελτιστοποίηση των προγραμμάτων κάθε επιπέδου εκτός του επιπέδου 1.

6.2.Συνολικά συμπεράσματα

Τα αποτελέσματα της DEA υποδεικνύουν ότι υπάρχει μεγάλο περιθώριο βελτίωσης των προγραμμάτων E&A, από τη στιγμή που μόνο το 5.5% των προγραμμάτων προέκυψε αποδοτικό, ενώ το 93% είχε δείκτη αποδοτικότητας κάτω από 0.5. Τα προγράμματα E&A της Ε.Ε. μπορούν να αξιοποιήσουν καλύτερα τους πόρους τους, χρησιμοποιώντας, δηλαδή, λιγότερες εισροές, για να παράγουν τις ίδιες εκροές.

Το βασικό συμπέρασμα που προκύπτει από τα αποτελέσματα της ανάλυσης των επιπέδων αποδοτικότητας των προγραμμάτων είναι ότι η συμμετοχή πολλών διαφορετικών χωρών αλλά και οργανισμών σε ένα έργο επηρεάζει αρνητικά την αποδοτικότητα του έργου. Αυτό οφείλεται στις δυσκολίες που πιθανότατα προκύπτουν κατά τον συντονισμό και την επικοινωνία πολλών διαφορετικών χωρών και οργανισμών. Παράλληλα, φαίνεται ότι υπάρχουν αρκετά μεγάλα περιθώρια μείωσης του συνολικού κόστους των έργων, χωρίς να παρουσιαστεί μείωση των εκροών. Τέλος, η συνολική διάρκεια του έργου, δε φαίνεται να επηρεάζει τη παραγωγικότητα ενός έργου.

Το δένδρο ταξινόμησης που αναπτύχθηκε επιλέγει ως πρώτο κριτήριο διαχωρισμού τον αριθμό των διαφορετικών χωρών που συμμετέχουν σε κάθε έργο. Αυτό σημαίνει ότι οι συμμετοχές διαφορετικών χωρών είναι το χαρακτηριστικό το οποίο παρέχει τις περισσότερες πληροφορίες και άρα επηρεάζει πιο έντονα την αποδοτικότητα ενός έργου. Το παραπάνω συμπέρασμα έρχεται να επιβεβαιώσει το βασικό συμπέρασμα της ανάλυσης των επιπέδων αποδοτικότητας Προτάσεις. Επιπλέον, το δένδρο ταξινόμησης που αναπτύχθηκε, αποδείχτηκε να είναι ένα πολύ αξιόπιστο εργαλείο για τη πρόβλεψη του επιπέδου αποδοτικότητας των έργων Ε&Α, το οποίο μπορεί να αξιοποιηθεί από όλους τους οργανισμούς του τομέα κατά τη λήψη των αποφάσεων, την αξιολόγηση και την επιλογή έργων.

Συνδυάζοντας τα αποτελέσματα της ανάλυση επιπέδων αποδοτικότητας και της ομαδοποίησης των προγραμμάτων σε 4 ομάδες, προσδιορίστηκε η καλύτερη διαδρομή για τη σταδιακή βελτιστοποίηση των μη αποδοτικών προγραμμάτων. Η συγκεκριμένη μεθοδολογία δίνει τη δυνατότητα της τμηματικής βελτίωσης των προγραμμάτων, που θεωρείται πολύ πιο ρεαλιστική από την προσέγγιση της μεθόδου DEA, η οποία θέτει ως προγράμματα αναφοράς εκείνα που έχουν το βέλτιστο επίπεδο αποδοτικότητας. Η προσέγγιση της DEA, αν και δε θεωρείται λάθος, κρίνεται ιδεαλιστική, αφού προτείνει μια απότομη μείωση στους πόρους που χρησιμοποιούνται, που δε θα μπορούσε να εφαρμοστεί ρεαλιστικά στα έργα Ε&Α.

Το εργαλείο που αναπτύχθηκε μπορεί να αξιοποιηθεί άμεσα από την Ευρωπαϊκή Ένωση στη διαχείριση των έργων Ε&Α. Τα αποτελέσματα του εργαλείου δίνουν μια ευρεία οπτική για το επίπεδο αποδοτικότητας των προγραμμάτων που έχουν περατωθεί τα τελευταία έξι χρόνια, απ' όπου προκύπτουν συμπεράσματα για τους παράγοντες που επηρεάζουν την αποδοτικότητα των έργων. Επιπλέον, μπορεί να αξιοποιηθεί για τη πρόβλεψη του επιπέδου αποδοτικότητας που έπεται να έχουν νέα έργα Ε&Α ανάλογα με τους πόρους που διατίθενται. Τέλος, το εργαλείο έχει μεγάλη χρησιμότητα στην ανάπτυξη στρατηγικών μεθόδων βελτίωσης της αποδοτικότητας έργων που είτε είναι σε εξέλιξη, είτε που δεν έχουν ξεκινήσει ακόμα.

Επισημαίνεται ότι η μεθοδολογία της διπλωματικής εργασίας θα μπορούσε να προσαρμοστεί σε όλους τους τομείς που απαιτείται η αξιολόγηση της αποδοτικότητας,

έχοντας δεδομένα για τους πόρους που χρησιμοποιούνται για την παραγωγή προϊόντων ή υπηρεσιών, αλλά και τα προϊόντα ή τις υπηρεσίες που παράγονται.

6.3.Προτάσεις για περαιτέρω έρευνα

Τα παρόν μοντέλο, λόγω των περιορισμένων δεδομένων που διατίθενται, περιλαμβάνει μόνο ένα μικρό σύνολο εισροών που έχουν αντίκτυπο στα προϊόντα ή τις υπηρεσίες που παράγονται. Άλλοι παράγοντες που θα έπρεπε να ληφθούν υπόψη είναι η πολυπλοκότητα του έργου, η ποιότητα των διαθέσιμων υλικών και λογισμικών εργαλείων, το πλήθος, η εμπειρία του προσωπικού κλπ. Παράλληλα, οι εκροές θα μπορούσαν να περιέχουν πιο πολλές χρήσιμες πληροφορίες που αντανακλούν την παραγωγικότητα του έργου και είναι απαραίτητες για την αξιολόγηση της αποδοτικότητας. Το μοντέλο θα μπορούσε να περιέχει δεδομένα σχετικά με την ποιότητα των δημοσιεύσεων και των παραδοτέων των έργων, δείκτες διαχείρισης έργων, τη τήρηση των κατευθυντήριων γραμμών για την εσωτερική διαδικασία της εταιρείας και των συναντήσεων επισκόπησης σχεδιασμού κ.λπ.. Σε μελλοντικές αναλύσεις της DEA, αυτοί οι παράγοντες μπορούν να ενσωματωθούν στο μοντέλο παραγωγής ως είσοδοι ή έξοδοι, με σκοπό την εξαγωγή αποτελεσμάτων για την αποδοτικότητα των έργων που αντικατοπτρίζουν καλύτερα την αποδοτικότητα.

Επίσης, πολλά έργα Ε&Α αποσκοπούν στην αύξηση των αποτελεσμάτων παρά στην μείωση των εισροών. Ενδιαφέρουσα θα ήταν, λοιπόν, η προσέγγιση της μεθοδολογίας με προσανατολισμό προς τις εκροές, δηλαδή με σκοπό την ανάπτυξη στρατηγικών μεθόδων για τη μεγιστοποίηση των εκροών, διατηρώντας σταθερές τις εισροές. Η προσέγγιση αυτή, μπορεί να είναι ωφέλιμη για οργανισμούς, όπως η Ευρωπαϊκή Ένωση, που πραγματοποιούν προκηρύξεις διαγωνισμών και έχουν προαποφασίσει τους πόρους που θα χρησιμοποιηθούν για την εκτέλεση του έργου και άρα αποβλέπουν στη μεγιστοποίηση των εκροών.

Τέλος, θα ήταν σκόπιμη μια μετα-ανάλυση (meta-analysis), μέσω της οποίας θα διερευνηθούν τα ακόλουθα: α) η συσχέτιση μεταξύ των εισροών των έργων Ε&Α, β) οι διαφοροποιήσεις που υπάρχουν στα επίπεδα αποδοτικότητας ανάλογα το τομέα που βρίσκονται και γ) η ανάπτυξη καινούριων μοντέλων DEA για κάθε διαφορετικό τομέα. Η μετα-ανάλυση θα αποσκοπεί στην εύρεση της βέλτιστης στρατηγικής για το κάθε οργανισμό, ανάλογα με το τομέα στον οποίο ανήκει.

7. Βιβλιογραφία

Augood, D.R. (1975). A new approach to R&D evolutionlike Transaction on Engineering Management.

Baker RC, Talluri S. (1997). A closer look at the use of data envelopment analysis for technology selection. Computers and Industrial Engineering.

Banker, R.D., R.F. Charnes, & W.W. Cooper. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis.

Bound, J., Cumins, C., Griliches, Z., Hall, H.H., Jaffe, A. (1984). R&D, Patent, and Productivity. University of Chicago Press, Chicago.

Bradbury, F.R., Gallagher, W.M. and Suckling, C.W. (1973). Qualitative aspects of the evaluation and control of research and development projects.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Boca Raton, FL: CRC Press.

Charnes A., Cooper W.W. and Rhodes E. (1978). Measuring the efficiency of decision-making units.

Clemen, R.T. (1996). Making Hard Decisions: An Introduction to Decision Analysis.2ndedition, California: Duxbury Press

Danila, N. (1983). Strategies Technologiques, Methodes d' Evaluation et de Selection des Projects de Recherch.

Dixit, A.K. and Pindyck, R.S. (1994). Investment under Uncertainty. New Jersey: Princeton University Press.

Ellis, L.W. (1984). Viewing R&D projects financially.

Emmanuel Thabassoulis. (2001). Definitions of efficiency and related measures, Basic Principles,General Models Introduction to the Theory and Application of Data Envelopment Analysis. Kluwer Academic Publishers, pp 1-85.

Fahrni, P. and Spatig, M. (1990). An application-oriented guide to R&D project selection and evaluation methods.

Farrell, M.J. (1957). The Measurement of Productive Efficiency.

Faulkner, T.W. (1996). Apply `Options Thinking' to R&D evaluation.

Freeman, C. (1982). The Economics of Industrial Innovation. London: Frances Printer.

Gear, A.E. (1974). A review of some recent developments in portfolio modelling in applied research and development. IEEE Transactions on Engineering Management, 21,3,119-125.

Graves, S.B. and Ringuest, J.L. (1991). Evaluating competing R&D investments. Research Technology Management, 34,4, 32-35.

Graves, S.B., Langowitz, N.S. (1996). R&D productivity: A global multi-industry comparison.

H. Lee et al. (2009). Comparative evaluation of performance of national R&D programs with heterogeneous objectives: A DEA approach. European Journal of Operational Research.

Harel Eilat, Boaz Golany, Avraham Shtub. (2008). R&D project evaluation: An integrated DEA and balanced scorecard approach. Omega, Volume 36, Issue 5, Pages 895-912.

Howard, R.A. and Metheson, J.E. (1981). Influence diagrams In Readings on the Principles and Applications of Decision Analysis.

Irvine, J. (1988). Evaluating Applied Research: Lessons from Japan. London: Pinter Publishers

Jackson, B. (1983). Decision methods for evaluating R&D projects. Research Management, 6, 4, 16-22

Jackson, B. (1983). Decision methods for evaluating R&D projects. Research Management, 6, 4, 16-22.

K.L. Poh B.W. Ang F. Bai. (2002). A comparative analysis of R&D project evaluation methods.

Karlaftis, M.G., Tsamboulas, D. (2012). Efficiency measurement in public transport: are findings specification sensitive? Transportation Research Part A: Policy and Practice.

Khouja M. (1995) The use of data envelopment analysis for technology selection. Computers and Industrial Engineering.

Krawiec, F. (1984). Evaluating and selecting research projects by scoring. Research Management, 27, 2, 21-25

Kuwahara, Y. and Takeda, Y. (1990). Managerial approach to research and development cost-effectiveness evaluation.IEEE Transaction on Engineering Management, 37, 2, 134 - 138

Liberatore, M.J. (1987). An extension of the Analytic Hierarchy Process for industrial R&D project selectionand resource allocation. IEEE Transactions on Engineering Management, 34, 4, 12 -18.

Link, A.N. (1993). Methods for evaluating the return of R&D. In Evaluating R&D Impacts: Methods and Practice.

Linton, J.D., Walsh, S.T. and Kirchhoff, B.A. (2002). R&D Management.

Lockett, G., Hetherington, B. and Yallup, P. (1986). Modeling a research portfolio using AHP: a group decision process. R&D Management, 16, 2, 151- 160.

Martino, J.P. (1995). Research and Development Project Selection. New York: John Wiley & Sons.

Morgan, M.G. and Henrion, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press.

Morris, P.A., Teisberg, E.O. and Kolbe, A.L. (1991). When choosing R&D projects, go with long shots. Research Technology Management, 34, 1, 35-40.

Oral M, Kettani O, Lang P. (1991). A methodology for collective evaluation and selection of industrial R&D projects. Management Science.

Ormala, E. (1986). Analysis and Supporting R&D Project Evaluation. Technical Research Centre of Finland, Espoo.

Porter, A.C., et al. (1980). A Guidebook for Technology Assessment and Impact Analysis. New York: North-Holland

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning.

Quinlan, J. R. (1987). Simplifying Decision Trees. International Journal of Man-Machine Studies.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufman.

Quinlan, J. R., & Rivest, R. L. (1989). Inferring Decision Trees Using the Minimum Description Length Principle. Information and Computation.

Raiffa, H. (1968). Decision Analysis: Introductory Lectures on Choices Under Uncertainty. New York: Random House.

Ramanathan, R. (2003). An Introduction to Data Envelopment Analysis: A tool for Performance Measurement, Sage Publishing.

Roll Y, Golany B. (1993). Alternate methods of treating factor weights in DEA.

Saaty, T.L. (1980). The Analytic Hierarchy Process: Planning, Priority, Setting Resource Allocation. New York: McGraw-Hill

Savage, L.J. (1954). The Foundations of Statistics. New York: John Wiley

Scherer, F.M. (1983). The propensity to patent. International Journal of Industrial Organization.

Souder, W.E. (1978). System for using R&D project evaluation methods. Research Management.

Thanassoulis E., Dyson, R.G. and Foster, M.J. (1987). Relative efficiency assessments using data envelopment analysis: an application to data on rates departments.

Thomas, H. (1985). Decision analysis and strategic management of research and development. R&D Management, 15,1, 3 -22.

Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. Science.

Συμεωνίδης, Π., Γούναρης, Α. 2015. Ομαδοποίηση Δεδομένων.