ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΣΤΗ ΔΙΑΧΕΙΡΙΣΗ ΑΝΘΡΩΠΙΝΟΥ ΔΥΝΑΜΙΚΟΥ

Κυριάκος Ν. Κοκκινόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούλιος 2024

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΠΡΟΒΛΕΨΗΣ ΣΤΗ ΔΙΑΧΕΙΡΙΣΗ ΑΝΘΡΩΠΙΝΟΥ ΔΥΝΑΜΙΚΟΥ

Κυριάκος Ν. Κοκκινόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούλιος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

- Αναπληρωτής Καθηγητής Κωνσταντίνος Πολίτης (Επιβλέπων)
- Καθηγήτρια Γεωργία Βερροπούλου
- Καθηγητής Πλάτων Τήνιος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

STATISTICAL PREDICTIVE MODELS IN HUMAN RESOURCES MANAGEMENT

By

Kyriakos N. Kokkinopoulos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
July 2024

Στη Μαρίνα μου

Ευχαριστίες

Με την εκπόνηση της παρούσας διπλωματικής εργασίας, κλείνει ο κύκλος μου στο μεταπτυχιακό πρόγραμμα Εφαρμοσμένη Στατιστική του Τμήματος Στατιστικής και Ασφαλιστικής επιστήμης του Πανεπιστημίου Πειραιώς. Θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν στην πορεία μου σε αυτό το μεταπτυχιακό, καθηγητές και μη. Πάνω απ' όλα όμως, θα ήθελα να εκφράσω τις ευχαριστίες μου και την εκτίμησή μου στον επιβλέποντα Αναπληρωτή Καθηγητή κ. Κωνσταντίνο Πολίτη ο οποίος με την καθοδήγησή, την υποστήριξη και τις εύστοχες παρατηρήσεις του με βοήθησε στην ολοκλήρωση της παρούσας διπλωματικής εργασίας. Τέλος, θα ήθελα να ευχαριστήσω τη σύντροφό μου Μαρίνα για την υποστήριξη και την υπομονή της καθώς και τους γονείς μου για τα εφόδια που μου παρείχαν όλα αυτά τα χρόνια.

Περίληψη

Η διαχείριση του ανθρώπινου δυναμικού είναι κρίσιμη για την επιτυχία και την ανταγωνιστικότητα των εταιρειών. Μια από τις βασικότερες προκλήσεις στη διαχείριση ανθρώπινου δυναμικού είναι η διατήρηση και η ανάπτυξη των εργαζομένων του κάθε οργανισμού. Στις μέρες μας, έχει παρατηρηθεί το φαινόμενο της συχνής αλλαγής εργασίας. Εργαζόμενοι όλων των ηλικιών, δεν φοβούνται να δοκιμάσουν να εργαστούν σε νέες εταιρείες, με απώτερο σκοπό καλύτερες απολαβές και καλύτερες συνθήκες εργασίας. Αυτό το φαινόμενο, έχει πολλές επιπτώσεις στις εταιρείες. Για να μπορέσει ένα τμήμα ανθρώπινου δυναμικού να ανταπεξέλθει στο παραπάνω φαινόμενο, χρειάζεται εργαλεία. Εργαλεία που βασίζονται σε δεδομένα και μπορούν να βοηθήσουν στην κατανόηση των λόγων που οδηγούν τους εργαζόμενους σε αποχώρηση ώστε να ληφθούν μέτρα για την πρόληψή τους. Στην παρούσα διπλωματική εργασία, μέσω του διαθέσιμου συνόλου δεδομένων, το οποίο περιλαμβάνει προσωπικές και επαγγελματικές πληροφορίες για κάθε εργαζόμενο και χρησιμοποιώντας λογιστική παλινδρόμηση και αλγορίθμους μηχανικής μάθησης, θα προσπαθήσουμε να προβλέψουμε την αποχώρηση ή την παραμονή του εργαζόμενου στην εταιρεία. Επίσης, δίνεται το θεωρητικό υπόβαθρο αυτών των μεθόδων, γραφήματα και γραφικές παραστάσεις των μεταβλητών καθώς και τα περιγραφικά μέτρα. Τέλος, γίνεται σύγκριση των αποτελεσμάτων των μοντέλων με χρήση κατάλληλων μετρικών για την αξιολόγηση της απόδοσής τους ώστε να καταλήξουμε στο καταλληλότερο και αποδοτικότερο μοντέλο για την πρόβλεψη μας.

Abstract

Human Resource Management is critical to the success and competitiveness of companies. One of the key challenges in Human Resources Management is the retention and development of employees in any organization. In our days, the phenomenon of frequent job change has been observed. Employees of all ages are not afraid to work in new companies, with the ultimate goal of better remuneration and better working conditions. This phenomenon has many consequences for the companies. In order to face this phenomenon, a Human Resources department needs tools. These data-based tools can help identify the reasons that lead employees to leave, in order to take measures and prevent them. In this thesis, through the available dataset, which includes personal and professional information about each employee, using logistic regression and machine learning algorithms, we will try to predict the employee's departure or retention in the company. The theoretical background of these methods, such as graphs and plots of the variables and descriptive measures, are also given. Finally, the results of the models are compared using appropriate metrics to evaluate their performance in order to reach the most appropriate and efficient model for our prediction.

Πίνακας Περιεχομένων

Κατάλογος πινάκων 16

Κατάλογος εικόνων 17

1.Εισαγωγή 19

1.1 Το φαινόμενο της αποχώρησης των εργαζομένων 19

1.2 Οι λόγοι της αποχώρησης των εργαζομένων 20

1.3 Οι συνέπειες στις επιχειρήσεις λόγω της αποχώρησης των εργαζομένων 21

1.4 Στόχοι της διπλωματικής εργασίας 23

1.5 Βιβλιογραφική επισκόπηση 24

2.Μηχανική Μάθηση 26

2.1 Εισαγωγή 26

2.2 Κατηγορίες μηχανικής μάθησης 26

2.2.1 Εποπτευόμενη μηχανική μάθηση (Supervised Machine Learning) 27

2.2.2 Μη εποπτευόμενη μηχανική μάθηση (Unsupervised Machine Learning) 28

2.3 Αλγόριθμοι εποπτευόμενης μάθησης 29

2.3.1 Δέντρο Απόφασης (Decision Tree) 29

2.3.1.1 Αλγόριθμοι Δέντρων Απόφασης 29

2.3.2 Τυχαίο Δάσος (Random Forest) 32

2.3.3. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM) 33

3.Λογιστική Παλινδρόμηση 36

3.1 Εισαγωγή 36

3.2 Το μοντέλο logit 37

3.3 Το μοντέλο probit 39

3.4 Το μοντέλο clog-log 40

4.Παρουσίαση των δεδομένων, περιγραφικά μέτρα και διαγραμματικές απεικονίσεις 42

4.1 Παρουσίαση των δεδομένων 42

4.2 Έλεγχος για ελλιπείς τιμές και περιγραφικά μέτρα 43

4.2.1 Έλεγχος για ελλιπείς τιμές (missing values) 43

4.2.2 Περιγραφικά μέτρα 44

4.3 Διαγραμματικές απεικονίσεις 47

5.Εφαρμογή ταξινόμησης με χρήση της λογιστικής παλινδρόμησης στο σύνολο δεδομένων 54

5.1 Εισαγωγή 54

5.2 Προεπεξεργασία Δεδομένων 54

5.3. Εφαρμογή ταξινόμησης με χρήση λογιστικής παλινδρόμησης 55

5.3.1 Εφαρμογή ταξινόμησης με χρήση λογιστικής παλινδρόμησης για το πλήρες μοντέλο 56

5.3.2 Εφαρμογή ταξινόμησης με χρήση λογιστικής παλινδρόμησης για υποσύνολο του πλήρους μοντέλου 60

6.Εφαρμογή ταξινόμησης με χρήση αλγορίθμων μηχανικής μάθησης στο σύνολο δεδομένων 63

6.1 Εισαγωγή 63

6.2 Αλγόριθμος Δέντρου Απόφασης (Decision Tree Algorithm) 63

6.3 Αλγόριθμος Τυχαίου Δάσους (Random Forest Algorithm) 65

6.4 Αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines Algorithm) 67

7.Συμπεράσματα 70

Βιβλιογραφία 72

Παράρτημα: Κώδικας Python 76

Κατάλογος πινάκων

Πίνακας 1: Περιγραφή των μεταβλητών 43
Πίνακας 2: Περιγραφικά μέτρα αριθμητικών μεταβλητών 44
Πίνακας 3: Περιγραφικά μέτρα κατηγορικών μεταβλητών 44
Πίνακας 4: Μέτρηση τιμών ανά κατηγορική μεταβλητή 45
Πίνακας 5: Μορφή πίνακα σύγχυσης (Confusion Matrix) 56
Πίνακας 6: Αποτελέσματα ταξινόμησης (Classification Report) 57
Πίνακας 7: Πίνακας ANOVA reduced model 59
Πίνακας 8: Αποτελέσματα ταξινόμησης – reduced model(Classification Report) 60
Πίνακας 9: Αποτελέσματα ταξινόμησης – Decision Tree (Classification Report) 63
Πίνακας 10: Αποτελέσματα ταξινόμησης – Random Forest (Classification Report) 65
Πίνακας 11: Αποτελέσματα ταξινόμησης – SVM (Classification Report) 68
Πίνακας 12: Σύγκριση αποτελεσμάτων αλγορίθμων ταξινόμησης 71

Κατάλογος Εικόνων

Εικόνα 1: Βήματα εποπτευόμενης μάθησης (Supervised Learning) 27
Εικόνα 2: Γραφική αναπαράσταση ταξινόμησης και παλινδρόμησης 28
Εικόνα 3: Το δέντρο απόφασης 30
Εικόνα 4: Βήματα εκτέλεσης του αλγορίθμου τυχαίου δάσους 32
Εικόνα 5: Γραφική απεικόνιση λειτουργίας SVM 34
Εικόνα 6: Συνάρτηση logit 38
Εικόνα 7: Συνάρτηση probit 39
Εικόνα 8: Συνάρτηση clog-log 40
Εικόνα 9: Διάγραμμα συναρτήσεων logit, probit & C Log-log 41
Εικόνα 10: Ιστογράμματα αριθμητικών μεταβλητών 47
Εικόνα 11: Pie charts: Attrition ανά Gender 48
Εικόνα 12: Pie charts: Attrition ανά BusinessTravel 48
Εικόνα 13: Pie charts: Attrition ανά EducationField 49
Εικόνα 14: Boxplots: MonthlyIncome - Gender ανά Attrition 49
Εικόνα 15: Boxplots: Age - Gender ανά Attrition 50
Εικόνα 16: Boxplots: DistanceFromHome – Attrition 51
Εικόνα 17: Barcharts: Job Roles – Attrition 51
Εικόνα 18: Pie Charts JobSatisfaction - Attrition 52
Εικόνα 19: Barcharts: Job Involvement - Attrition 53
Εικόνα 20: Κωδικοποίηση κατηγορικών μεταβλητών 54
Εικόνα 21: Πίνακας σύγχυσης ταξινόμησης με χρήση λογιστικής παλινδρόμησης (πλήρες μοντέλο) 57
Εικόνα 22: Καμπύλη ROC πλήρους μοντέλου με χρήση λογιστικής παλινδρόμησης 58
Εικόνα 23: Πίνακας σύγχυσης ταξινόμησης με χρήση λογιστικής παλινδρόμησης (reduced model) 61
Εικόνα 24: Καμπύλη ROC reduced μοντέλου με χρήση λογιστικής παλινδρόμησης 62
Εικόνα 25: Πίνακας σύγχυσης ταξινόμησης με χρήση δέντρων απόφασης 64

Εικόνα 26: Καμπύλη ROC μοντέλου με χρήση δέντρων απόφασης 65
Εικόνα 27: Πίνακας σύγχυσης ταξινόμησης με χρήση αλγορίθμου τυχαίου δάσους 66
Εικόνα 28: Καμπύλη ROC μοντέλου με χρήση αλγόριθμου τυχαίου δάσους 67
Εικόνα 29: Πίνακας σύγχυσης ταξινόμησης με χρήση αλγορίθμου SVM 68
Εικόνα 30: Καμπύλη ROC μοντέλου με χρήση αλγόριθμου SVM 69

Κεφάλαιο 1

Εισαγωγή

1.1 Το φαινόμενο της αποχώρησης των εργαζομένων

Η αποχώρηση των εργαζομένων από έναν οργανισμό, αποτελεί πρόκληση για τις εταιρείες. Τα τελευταία χρόνια, η αντίληψη των εργαζομένων σε σχέση με την παραμονή τους σε μια εταιρεία έχει αλλάξει. Τα προηγούμενα χρόνια, ίσως και λόγω της ανασφάλειας όπου υπήρχε λόγω της οικονομικής κρίσης που περνούσε η χώρα, υπήρχε μειωμένη κινητικότητα των εργαζομένων καθώς φοβόντουσαν την ανεργία. Οι εργαζόμενοι παρέμεναν στις θέσεις εργασίας τους χωρίς να είναι ικανοποιημένοι. Πλέον όμως, όπου η τεχνολογία έχει μπει για τα καλά στις ζωές μας και μπορούμε να στείλουμε πολύ εύκολα online βιογραφικά σε εταιρείες καθώς και με το employer branding όπου οι επιχειρήσεις μέσω ενεργειών προσπαθούν να διαφημιστούν και να προσελκύσουν νέους εργαζόμενους, τα πράγματα έχουν αλλάξει. Συμβολή στην άνοδο της τάσης της αποχώρησης των εργαζομένων έχουν και οι πλατφόρμες κοινωνικής δικτύωσης. Για παράδειγμα το LinkedIn, ως μια «εργασιακού τύπου» πλατφόρμα έχει συμβάλει σε μεγάλο ποσοστό στην άμεσα επαφή των εργαζομένων με άλλους επαγγελματίες καθώς και με εταιρείες και θέσεις εργασίας. Δεν είναι τυχαίο άλλωστε πως πλέον οι υπεύθυνοι προσλήψεων των εταιρειών αναζητούν και προσεγγίζουν εργαζόμενους μέσω των προφίλ τους στη συγκεκριμένη πλατφόρμα καθώς κάθε προφίλ εργαζόμενου είναι ένα online διαθέσιμο βιογραφικό προς όλους.

Σε έρευνα που έκανε η εταιρεία EY το 2023 με τίτλο Work Reimagined Survey έδειξε πως το 34% των εργαζομένων αναφέρουν ότι είναι πρόθυμοι να αλλάξουν εργασιακή στέγη στους επόμενους 12 μήνες καθώς η καλύτερη αμοιβή είναι το κύριο μέλημα των εργαζομένων. Επίσης, το τμήμα στατιστικής του υπουργείου εργασίας των ΗΠΑ (Bureau of Labor Statistics – U.S. Department of Labor) σε έρευνα που έκανε το 2022 έδειξε πως 4,1 έτη είναι η διάμεσος παραμονής των εργαζομένων στις εταιρείες. Η εταιρεία Microsoft μαζί με την πλατφόρμα LinkedIn έδειξε πως το 46% των εργαζομένων που συμμετείχαν σκέφτεται να αλλάξει δουλειά μέσα στο 2024 ενώ η ξεχωριστή έρευνα της πλατφόρμας LinkedIn έδειξε πως τα νούμερα στην Αμερική είναι αρκετά αυξημένα με το 85% των εργαζομένων να βλέπει θετικά την αλλαγή εργασιακής στέγης. Σε έρευνα του European Data Journalism Network αναφέρεται πως στα τέλη του 2022 υπήρξε «έκρηξη» οικειοθελών παραιτήσεων κυρίως σε χώρες όπως η Γαλλία, η Ισπανία και η Ιταλία. Συγκεκριμένα, αναφέρεται πως εκτός των άλλων, στη Γαλλία και την Ιταλία μόνο το 2022 οι αριθμοί των συμβάσεων που τερματίστηκαν από την πλευρά των εργαζομένων ανέρχονται σε 2.16 και 2.2 εκατομμύρια αντίστοιχα. Στην Ολλανδία, το πρώτο τρίμηνο του 2022, 1.9 εκατομμύρια εργαζόμενοι ανέφεραν πως ξεκίνησαν να εργάζονται σε νέα δουλειά μέσα στο έτος. Τέλος,

όπως αναφέρει η έρευνα, όλα τα παραπάνω είναι σύμπτωμα της πανδημίας λόγω του Covid-19 καθώς οι Ευρωπαίοι μετά την πανδημία, δεν θέλουν να εργάζονται τόσο πολύ όσο πριν.

1.2 Οι λόγοι της αποχώρησης των εργαζομένων

Οι λόγοι αποχώρησης ενός εργαζόμενου από έναν οργανισμό μπορούν να είμαι πάρα πολλοί. Σημαντικό ρόλο σε αυτό παίζουν η προσωπικότητα και οι ανάγκες κάθε ανθρώπου. Ωστόσο υπάρχουν κάποιοι λόγοι που είναι αρκετά συνηθισμένοι και λογίζονται ως οι κυριότεροι. Μερικοί από αυτούς είναι:

- **Αμοιβή - Αποδοχές**: Είναι ο σημαντικότερος λόγος αποχώρησης ενός εργαζόμενου καθώς είναι και ο λόγος για τον οποίο εργάζεται προκειμένου να βγάλει τα προς το ζην. Οι εργαζόμενοι όταν αισθάνονται ότι δεν αμείβονται ικανοποιητικά με βάση τα όσα προσφέρουν τείνουν να αναζητούν εργασία σε διαφορετικό εργοδότη με καλύτερες αποδοχές. Σύμφωνα με έρευνα της εταιρείας Adecco το 2022 με τίτλο Exploring Workers Professional Aspirations, ο κυριότερος λόγος επιλογής εργοδότη για έναν εργαζόμενο είναι οι μηνιαίες απολαβές.
- **Παροχές**: Στις ημέρες μας, όλο και περισσότερες επιχειρήσεις εκτός από τον μηνιαίο μισθό, προσφέρουν και κάποιες επιπλέον παροχές. Κάποιες από αυτές είναι η ιδιωτική ασφάλιση, μηνιαία ή ετήσια bonus, παροχή διατακτικών σίτισης, κάλυψη μετακινήσεων (εξοδολόγιο), εκπτώσεις σε προϊόντα-καταστήματα, επιπλέον ημέρες ετήσιας κανονικής άδειας και άλλα. Οι συγκεκριμένες παροχές κάνουν τις συγκεκριμένες επιχειρήσεις πιο ελκυστικές και προσελκύουν νέους εργαζόμενους ευκολότερα.
- **Ανισορροπία μεταξύ επαγγελματικής και προσωπικής ζωής (Work-Life Balance):** Στις ημέρες μας, όπου οι εργασιακοί ρυθμοί έχουν αυξηθεί, υπάρχει η τάση για εργασία πάνω από τις 8 ώρες ημερησίως. Η συγκεκριμένη υπερωριακή εργασία, στην πληθώρα των περιπτώσεων δεν αμείβεται επιπλέον όπως θα έπρεπε με αποτέλεσμα οι εργαζόμενοι να εργάζονται πολλές περισσότερες ώρες από τις συμφωνημένες. Ο συγκεκριμένος λόγος συνδυάζεται και με την έρευνα που είδαμε στην ενότητα 1.1 πως οι εργαζόμενοι ειδικά μετά την πανδημία του Covid-19 επιθυμούν να εργάζονται λιγότερο από πριν.
- **Έλλειψη εξέλιξης και ανάπτυξης:** Οι εργαζόμενοι, ειδικότερα στις μικρότερες ηλικίες όταν βρίσκονται στην αρχή της καριέρας τους, θέλουν ευκαιρίες για να μπορέσουν να προαχθούν και να εξελιχθούν επαγγελματικά. Όταν στον οργανισμό που βρίσκονται όμως δεν υπάρχουν αυτές οι ευκαιρίες και δεν δίνονται στον εργαζόμενο επιπλέον αρμοδιότητες προκειμένου να μπορέσει να ανελιχθεί , τότε ο εργαζόμενος επιλέγει να αποχωρήσει προκειμένου να βρει μια νέα θέση με μεγαλύτερες προοπτικές.
- **Κακή/Ανεπαρκής ηγεσία:** Ο συγκεκριμένος λόγος αφορά τη σχέση και τη συμπεριφορά που έχει ένας εργαζόμενος με τον προϊστάμενο ή τον διευθυντή του. Ο ρόλος ενός προϊστάμενου είναι μέσω της καθοδήγησης να μπορέσει να εξελίξει τους υφισταμένους του, να τους κάνει να νιώθουν χρήσιμοι μέσα στην ομάδα κρατώντας

τους ικανοποιημένους. Από την άλλη, υπάρχουν προϊστάμενοι οι οποίοι έχουν τοξική συμπεριφορά, μιλούν άσχημα στον υφισταμένους τους, τους κατηγορούν για λάθη και τους πιέζουν με αποτέλεσμα να υπάρχει ένα τοξικό εργασιακό κλίμα. Τέτοιες συμπεριφορές οδηγούν σε αυξημένη αποχώρηση των εργαζομένων.

- **Έλλειψη αναγνώρισης και επιβράβευσης:** Ο συγκεκριμένος λόγος συνδυάζεται με τους προηγούμενους δύο λόγους (έλλειψη εξέλιξης και ανάπτυξης & κακή/ανεπαρκής ηγεσία). Καθώς οι εργαζόμενοι έχουν ανάγκη να νιώθουν ότι η εργασία τους αναγνωρίζεται και επιβραβεύεται. Σε αντίθετη περίπτωση, χάνουν το κίνητρό τους και αναζητούν άλλη εργασία όπου η συνεισφορά τους στην ομάδα θα εκτιμάται και θα αναγνωρίζεται περισσότερο.
- **Κακές συνθήκες εργασίας:** Το περιβάλλον και ο χώρος εργασίας παίζουν πολύ σημαντικό ρόλο στην καθημερινότητα του εργαζόμενου. Ένας μοντέρνος, πρωτότυπος και πάνω απ' όλα ασφαλής χώρος συμβάλει σημαντικά στην ψυχολογία και στη διάθεσή του εργαζόμενου. Αντιθέτως, ένας χώρος ο οποίος δεν είναι συντηρημένος, δεν έχει την κατάλληλη υλικοτεχνική δομή και δεν εγγυάται την σωματική ασφάλεια του εργαζόμενου, συμβάλει αρνητικά στην παραμονή του στην εταιρεία.
- **Ευέλικτες συνθήκες εργασίας:** Κατά τη διάρκεια της πανδημίας, δημιουργήθηκε μια νέα συνθήκη εργασίας η οποία τάραξε τα νερά στον τρόπο παροχής της εργασίας. Η νέα αυτή συνθήκη ονομάζεται τηλεργασία. Πρόκειται για εξ αποστάσεως εργασία η οποία μπορεί να παρέχεται είτε από το σπίτι, είτε σε κάποιο εξωτερικό χώρο. Πλέον, έχει αυξηθεί η ζήτηση της τηλεργασίας από τους εργαζόμενους καθώς εκτός από το ότι εργάζονται σε ένα χώρο που τους είναι οικείος, γλιτώνουν το κόστος και τον χρόνο μεταφοράς προς την εργασία παρέχοντάς τους περισσότερο προσωπικό χρόνο.
- **Αλλαγή καριέρας:** Ο συγκεκριμένος λόγος αφορά την απόφαση του εργαζόμενου να εργαστεί σε έναν διαφορετικό κλάδο. Συνήθως αυτός ο λόγος συναντάται σε εργαζόμενους οι οποίοι βρίσκονται ακόμη στην αρχή της καριέρας τους, οι οποίοι συνειδητοποιούν πως δεν τους ταιριάζει ο συγκεκριμένος κλάδος και ότι δεν θα ήθελαν να κάνουν αυτή τη δουλειά για την υπόλοιπη ζωή τους. Σε ότι αφορά τα ηλικιακά κριτήρια για την αλλαγή καριέρας, δεν παύουν να υπάρχουν και εξαιρέσεις εργαζομένων οι οποίοι σε μεγαλύτερη ηλικία, ευρισκόμενοι σε υψηλότερες θέσεις, άλλαξαν καριέρα.

Εκτός από τους παραπάνω κυριότερους λόγους όπου αναφερθήκαμε, υπάρχουν και άλλοι λόγοι οι οποίοι είναι λιγότερο διαδεδομένοι και μπορεί να διαφέρουν αναλόγως τον κλάδο, την εταιρεία και τις προσωπικές προτιμήσεις του κάθε εργαζόμενου.

1.3 Οι συνέπειες στις επιχειρήσεις λόγω της αποχώρησης των εργαζομένων

Η αποχώρηση των εργαζομένων δεν είναι κάτι απλό για μια επιχείρηση. Έχει σημαντικές επιπτώσεις στο κομμάτι της λειτουργίας μιας επιχείρησης οι οποίες αν δεν αντιμετωπιστούν εγκαίρως, μπορεί να επιφέρουν ακόμη μεγαλύτερα προβλήματα. Όπως αναφέρει το Ινστιτούτο Εργασίας της Αμερικής στην έρευνα με τίτλο 2017 Retention Report, το κόστος

της αποχώρησης ενός εργαζόμενου για μια εταιρεία ισούται με το 33,33% του ετήσιου μισθού του εργαζόμενου που αποχώρησε. Παρακάτω θα δούμε τις πιο διαδεδομένες συνέπειες που δημιουργούνται σε μια εταιρεία λόγω της αποχώρησης ενός εργαζόμενου:

- **Κόστος:** Η αντικατάσταση ενός εργαζόμενου έχει οικονομικό κόστος για μια εταιρεία. Σε πρώτη φάση, ως κόστος λογίζονται οι αποζημιώσεις που θα λάβει ο εργαζόμενος. Με μια βαθύτερη ματιά όμως, προκύπτουν και άλλα κόστη. Για την εύρεση και πρόσληψη ενός εργαζόμενου απαιτείται η δημοσίευση αγγελίας σε μια πλατφόρμα. Απαιτείται χρόνος, άρα και ωρομίσθια, από τους εργαζόμενους που βρίσκονται στον τομέα των προσλήψεων ώστε να πραγματοποιήσουν τις αξιολογήσεις των βιογραφικών και τις συνεντεύξεις. Και τέλος, όπως αναφέρουν οι O'Connell M. et al (2007), οι εργαζόμενοι δεν είναι 100% παραγωγικοί από τη στιγμή που ξεκινούν. Είναι απαραίτητο να επενδυθεί χρόνος και πόροι για την εκπαίδευση, την ένταξη και την ανάπτυξη.
- **Απώλεια Παραγωγικότητας**: Μια κενή θέση λόγω αποχώρησης ενός εργαζόμενου μειώνει την παραγωγικότητα της εταιρείας για κάποιο χρονικό διάστημα. Αυτό έχει ως αποτέλεσμα να μην τηρούνται οι προθεσμίες, είτε να εργάζονται παραπάνω οι εργαζόμενοι που έχουν παραμείνει, το οποίο μπορεί να επιφέρει επιπλέον αποχωρήσεις. Όσο και να καλύψει τη θέση σύντομα η εταιρεία, όπως αναφέραμε και παραπάνω, ένας νέος εργαζόμενος χρειάζεται χρόνο και εκπαίδευση ώστε να φτάσει τα επίπεδα παραγωγικότητας ενός παλιού.
- **Απώλεια γνώσης**: Κάθε εργαζόμενος όσο περισσότερο βρίσκεται σε μια θέση, τόσο πιο πολλές γνώσεις και δεξιότητες αποκτά και εξειδικεύεται περισσότερο. Όταν όμως ένας εργαζόμενος αποχωρεί, παίρνει μαζί του τις συγκεκριμένες γνώσεις και δεξιότητες οι οποίες χρειάζονται χρόνο προκειμένου να αποκτηθούν από τον νεοπροσληφθέντα. Οι O'Connell M. et al (2007) αναφέρουν ότι σε εταιρείες όπου υπάρχει μεγάλος βαθμός αποχωρήσεων, εργαζόμενοι χωρίς μεγάλη εμπειρία καταλήγουν να εκπαιδεύουν τους νέους εργαζόμενους. Αυτό έχει ως συνέπεια οι νέοι εργαζόμενοι να μην εκπαιδεύονται σωστά και να υποκύπτουν σε λάθη.
- **Ανασφάλεια και αβεβαιότητα:** Μια αποχώρηση πάντα τραντάζει τα νερά μιας εταιρείας, ειδικότερα της ομάδας που ανήκει ο εργαζόμενος. Έτσι, υπάρχει η πιθανότητα να επηρεαστούν αρνητικά τα μέλη της ομάδας που συνεργάζονταν με τον εργαζόμενο που αποχωρεί. Αυτό μπορεί να έχει ως συνέπεια την αρνητική επίδραση του ηθικού της ομάδας, είτε τη δημιουργία τους αισθήματος αβεβαιότητας και ανασφάλειας σχετικά με το μέλλον της ομάδας που μπορεί να οδηγήσει σε μαζικές αποχωρήσεις.
- **Επιπτώσεις στη φήμη της εταιρείας:** Η υψηλή αποχώρηση εργαζομένων από μια εταιρεία μπορεί να δημιουργήσει τη φήμη του κακού εργοδότη. Ειδικότερα στις μέρες μας όπου στα social media (πχ. LinkedIn) μπορεί κάποιος πολύ εύκολα να δει εάν αποχωρούν εργαζόμενοι από μια εταιρεία. Η συγκεκριμένη φήμη, μπορεί να έχει επιπτώσεις και στην προσέλκυση νέων ταλέντων.

- **Επιπτώσεις στις σχέσεις με τους πελάτες:** Οι εργαζόμενοι μιας εταιρείας, εάν υπάρχει μια σταθερή συνεργασία, αποκτούν σχέσεις εμπιστοσύνης με τους πελάτες. Εάν ο συγκεκριμένος εργαζόμενος που εμπιστεύεται ο πελάτης αποχωρήσει τότε μπορεί να επηρεαστεί σημαντικά η σχέση με τον πελάτη και να οδηγήσει είτε σε παροχή υπηρεσιών μειωμένης ποιότητας, είτε ακόμη και σε διακοπή της συνεργασίας. Επίσης, ένας πελάτης είναι δύσκολο να εμπιστευτεί μια εταιρεία η οποία αλλάζει συνεχώς πρόσωπα και δεν μπορεί να κρατήσει το προσωπικό της.

Από τα παραπάνω καταλήγουμε πως μια αποχώρηση έχει πάρα πολλές συνέπειες για τις επιχειρήσεις ενώ σίγουρα υπάρχουν και άλλες συνέπειες οι οποίες διαφέρουν ανάλογα με την επιχείρηση. Έτσι, δημιουργείται η ανάγκη για αντιμετώπιση αυτών των συνέπειων. Στην έρευνα 2017 Retention Report του Ινστιτούτου Εργασίας της Αμερικής αναφέρεται πως το 75% των οικειοθελών αποχωρήσεων θα μπορούσαν να αντιμετωπιστούν ενώ το υπόλοιπο 25% όχι. Συνεπώς είναι στα χέρια των εταιρειών να βρουν τρόπους οι οποίοι θα μπορέσουν να μειώσουν τις αποχωρήσεις προλαμβάνοντας τες.

1.4 Στόχοι της διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η διερεύνηση των σημαντικότερων παραγόντων που οδηγούν στην ικανοποίηση των εργαζομένων και κατά συνέπεια την διατήρησή τους στην εταιρεία ώστε να αποφευχθούν οι αποχωρήσεις. Εκτός των άλλων, η παρούσα διπλωματική εργασία αποσκοπεί να γεφυρώσει το χάσμα που υπάρχει μεταξύ της κλασικής διαχείρισης ανθρώπινου δυναμικού και της σύγχρονης προσέγγισης που βασίζεται σε δεδομένα και εύρεση τάσεων προκειμένου να βγουν σημαντικά συμπεράσματα. Το σύνολο των δεδομένων βρίσκεται διαθέσιμο στην ιστοσελίδα Kaggle, έχοντας δημιουργηθεί από την εταιρεία IBM για σκοπούς ανάλυσης και εκπαίδευσης. Περιλαμβάνει 1470 εργαζόμενους και 32 στήλες με πληροφορίες για κάθε εργαζόμενο οι οποίες είτε είναι προσωπικές είτε αφορούν την εργασία του. Θα δούμε κάποια περιγραφικά μέτρα για τις μεταβλητές των δεδομένων μας και στη συνέχεια θα προσπαθήσουμε μέσω γραφικών απεικονίσεων να κατανοήσουμε και να προσδιορίσουμε τους λόγους που οδηγούν στην αποχώρηση των εργαζομένων. Όπως για παράδειγμα μελετώντας την αποχώρηση ανά φύλο, ηλικία, τμήμα, ανά απόσταση από την οικεία όπως μελετάνε οι Schwanen & Djist (2002) και άλλα. Στη συνέχεια, θα εφαρμόσουμε τεχνικές κατηγοριοποίησης χρησιμοποιώντας τέσσερις αλγορίθμους (Λογιστική Παλινδρόμηση, Δέντρα Απόφασης, Αλγόριθμος Τυχαίου Δάσους και Μηχανές Διανυσμάτων Υποστήριξης-SVM). Τέλος, θα συγκρίνουμε τα αποτελέσματα των προβλέψεων ώστε να καταλήξουμε στο καλύτερο μοντέλο. Η παρούσα διπλωματική εργασία στοχεύει στον ακαδημαϊκό διάλογο για την ενσωμάτωση των μεθόδων πρόβλεψης στη διαχείριση ανθρώπινου δυναμικού, όπως έχει αναφερθεί και από τους Garg et al. (2021).

1.5 Βιβλιογραφική επισκόπηση

Στη συγκεκριμένη ενότητα, θα αναφερθούμε στα υπάρχοντα επιστημονικά άρθρα τα οποία σχετίζονται με την παρούσα διπλωματική εργασία, όπως και στα συμπεράσματα που κατέληξαν.

Οι Schwanen, T., Dijst, M. (2002) ερεύνησαν την σχέση που υπάρχει μεταξύ του χρόνου μετακίνησης και του χρόνου διάρκειας της εργασίας. Τα δεδομένα πάνω στα οποία βασίζεται η έρευνά τους, έχουν παρθεί από την Ολλανδική Εθνική Ταξιδιωτική Έρευνα του 1998. Με τη χρήση της πολλαπλής ανάλυσης παλινδρόμησης, κατέληξαν στο συμπέρασμα πως οι εργαζόμενοι δαπανούν σε μετακινήσεις περίπου 28 λεπτά την ημέρα (μία διαδρομή) για να μεταβούν στην εργασία τους και να εργαστούν 8 ώρες. Επίσης, οι χρόνοι μετακινήσεων ποικίλλουν ανάλογα με τις κοινωνικές και δημογραφικές μεταβλητές, ενώ η αστική μορφή δεν έχει μεγάλη σημασία στο χρόνο των μετακινήσεων.

Ο Dicao Tang (2022) εξετάζει την βελτιστοποίηση των συστημάτων διαχείρισης ανθρώπινου δυναμικού με τη χρήση της εξόρυξης δεδομένων και του αλγόριθμου Random Forest σε συστήματα διαχείρισης ανθρώπινου δυναμικού. Οι στατιστικές μέθοδοι που χρησιμοποιεί είναι η πολλαπλή γραμμικής παλινδρόμηση, το δέντρο απόφασης, ο αλγόριθμος Naïve-Bayes και ο αλγόριθμος τυχαίου δάσους(Random Forest). Τα δεδομένα της προέρχονται από τον διαγωνισμό HR Analytics της πλατφόρμας Kaggle.

Ο Wei Kai (2022) ερευνά την έγκαιρη πρόληψη του κινδύνου σε θέματα διοίκησης ανθρώπινου δυναμικού μέσω του Δέντρου Απόφασης και των Support Vector Machines. Ο συγγραφέας, αναφέρει πως κάθε επιχείρηση θα πρέπει να μελετά και να αναλύει τους κινδύνους που σχετίζονται σε θέματα HR σε πρώιμο στάδιο. Επίσης, τονίζει ότι η διοίκηση ανθρώπινων πόρων διαφέρει σημαντικά από περιπτώσεις όπως η διαχείριση κεφαλαίων ή πρώτων υλών καθώς ο συγκεκριμένος κλάδος επηρεάζεται σε μεγάλο βαθμό από πολλούς παράγοντες. Έτσι, λόγω της ιδιαιτερότητας και των διαφορετικών χαρακτηριστικών που έχουν οι άνθρωποι, είναι αδύνατο για δύο ανθρώπους να είναι ακριβώς ίδιοι.

Η Sigal Alon (2003) στη δημοσίευσή της στο Research in Social Stratification and Mobility εξετάζει την επιρροή των φύλων στις δυσκολίες της απασχόλησης καθώς και την επίδραση των οικονομικών κύκλων στην αγορά εργασίας του Ισραήλ. Για την έρευνα έχει χρησιμοποιηθεί πολυμεταβλητή ανάλυση. Τα αποτελέσματα δείχνουν πως οι οικονομικοί κύκλοι διαμορφώνουν και ενισχύουν την ανισότητα των φύλων στις δυσκολίες εύρεσης απασχόλησης. Αν και οι άνδρες είναι πιο ευαίσθητοι από τις γυναίκες στους οικονομικούς κύκλους, και τα 2 φύλα επηρεάζονται ώστε να ξεφύγουν από την υποαπασχόληση. Τέλος, υπογραμμίζεται ότι το καθεστώς απασχόλησης ενός ατόμου επηρεάζεται από τα κοινωνικοδημογραφικά του χαρακτηριστικά.

Η Olivia Brinck και η Hanna Larsson (2019) ερευνούν τις αξίες στον εργασιακό χώρο, την βιώσιμη εργασία και τους λόγους για τους οποίους αποχωρούν οι εργαζόμενοι από την εταιρεία. Οι εργαζόμενοι που μελετούν είναι γεννημένοι από το 1979 έως το 1994 και χωρίστηκαν σε 3 ηλικιακά groups(κάτω των 24 ετών, 24 έως 39 ετών, άνω των 39 ετών. Τα

δεδομένα έχουν ληφθεί μέσω ερωτηματολογίου (Copenhagen Psychosocial Questionnaire II) από τους εργαζόμενους μιας Σουηδικής εταιρείας που δραστηριοποιείται στον τομέα των logistics. Ο αριθμός των εργαζομένων που απάντησαν στο ερωτηματολόγιο είναι 59 εκ των οποίων 44 είναι άνδρες και οι 15 γυναίκες. Οι στατιστικές μέθοδοι που χρησιμοποιήθηκαν ήταν η ANOVA και η ανάλυση παλινδρόμησης. Από την ανάλυση των απαντήσεων προέκυψαν δύο κύριοι λόγοι για τους οποίους αποχωρούν οι εργαζόμενοι. Ο πρώτος λόγος είναι ότι υπάρχει έλλειψη καθοδήγησης και συνεργασίας, πράγμα που απάντησε το 55% των ερωτηθέντων. Ο δεύτερος λόγος είναι η έλλειψη επαγγελματικής ανέλιξης μέσα στον οργανισμό.

Οι I. Setiawan et al (2020) αναλύουν το πρόβλημα αποχώρησης των εργαζομένων με τη χρήση της λογιστικής παλινδρόμησης. Τα δεδομένα είναι 261 εργάσιμων ημερών του 2015 και αφορούν 4410 εργαζόμενους. Από την έρευνά τους, προέκυψε πως οι εργαζόμενοι με μικρή εργασιακή εμπειρία είναι πιο πιθανό να αποχωρήσουν καθώς επιθυμούν να αποκτήσουν μεγαλύτερη εργασιακή εμπειρία. Ένα ακόμη ενδιαφέρον χαρακτηριστικό της έρευνάς τους είναι ότι ένας άγαμος εργαζόμενος έχει μεγαλύτερη πιθανότητα να αποχωρήσει σε σχέση με έναν έγγαμο ή έναν διαζευγμένο. Τέλος, για να αποφευχθεί η αποχώρηση των εργαζομένων, προτείνουν ότι η εταιρεία χρειάζεται να βελτιώσει το εργασιακό περιβάλλον, να μειώσει τον φόρτο εργασίας και να βελτιωθεί η σχέση μεταξύ των managers και των εργαζομένων.

ΚΕΦΑΛΑΙΟ 2

Μηχανική Μάθηση

2.1 Εισαγωγή

Η μηχανική μάθηση(machine learning, ML) είναι ένα παρακλάδι της τεχνητής νοημοσύνης που χρησιμοποιώντας αλγόριθμους και δεδομένα μαθαίνει μέσα από αυτά όπως οι άνθρωποι. Επινοήθηκε από τον Arthur Samuel το 1959 ο οποίος την όρισε ως «Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί». Αν και το όλο εγχείρημα ξεκίνησε από ένα παιχνίδι ντάμας στον υπολογιστή, πλέον στις μέρες μας η μηχανική μάθηση έχει εξελιχθεί σε τέτοιο βαθμό ώστε να χρησιμοποιείται σε αυτοοδηγούμενα οχήματα, στις ιατρικές διαγνώσεις, στη κατηγοριοποίηση ακολουθιών DNA και άλλα.

Γενικότερα, η μηχανική μάθηση είναι ένα σημαντικό εργαλείο της επιστήμης δεδομένων. Μέσω στατιστικών μεθόδων, οι αλγόριθμοι έχουν τη δυνατότητα να αναλύουν μεγάλους όγκους σύνθετων δεδομένων, να αποκαλύπτουν κρυμμένα μοτίβα και να προβλέπουν μελλοντικές τάσεις (Garg et al., 2021).

Στο πλαίσιο της ανάλυσης ανθρώπινου δυναμικού, η μηχανική μάθηση προσφέρει πρωτοφανείς ευκαιρίες για τη βελτιστοποίηση των διαδικασιών ανθρώπινου δυναμικού, συμπεριλαμβανομένης της κρίσιμης πτυχής της διαχείρισης της αποχώρησης των εργαζομένων (Garg et al., 2021).

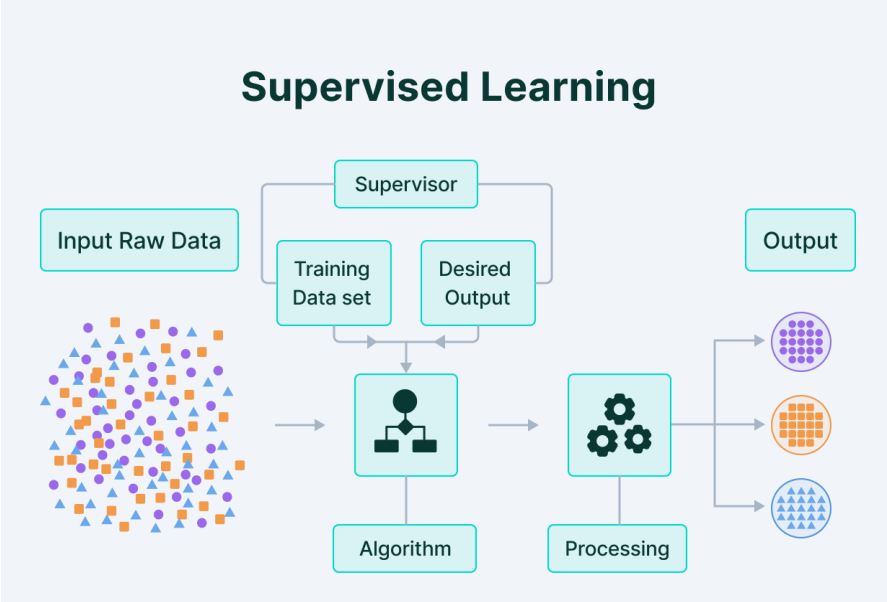
Η μηχανική μάθηση χωρίζεται σε τέσσερις κύριες κατηγορίες: την εποπτευόμενη μάθηση(Supervised Machine Learning), τη μη εποπτευόμενη μάθηση(Unsupervised Machine Learning), την ημι-εποπτευόμενη μάθηση (Semi-supervised Machine Learning) και την Ενισχυτική Μάθηση(Reinforcement Learning).

2.2 Κατηγορίες μηχανικής μάθησης

Αν και η μηχανική μάθηση χωρίζεται σε τέσσερις κύριες κατηγορίες, εμείς στη παρούσα διπλωματική εργασία θα αναφερθούμε στις δύο κυριότερες. Την εποπτευόμενη μηχανική μάθηση και την μη εποπτευόμενη μηχανική μάθηση.

2.2.1 Εποπτευόμενη μηχανική μάθηση (Supervised Machine Learning)

Το κύριο χαρακτηριστικό λειτουργίας της εποπτευόμενης μάθησης είναι η διαθεσιμότητα επισημασμένων(labelled) δεδομένων εκπαίδευσης. Ουσιαστικά, γνωρίζουμε εκ των προτέρων την κατηγοριοποίηση των δεδομένων και ο σκοπός είναι να εκπαιδευτεί ο αλγόριθμος πάνω σε αυτά τα δεδομένα. Τα δεδομένα εκπαίδευσης(training data), χρησιμοποιούνται από τον αλγόριθμο προκειμένου να τον βοηθήσουν να δημιουργήσει μοντέλα τα οποία στη συνέχεια όταν του ανατεθεί να ταξινομήσει μη επισημασμένα δεδομένα σε προκαθορισμένες κλάσεις, να το κάνει με τον καλύτερο δυνατό τρόπο. Η εποπτευόμενη μάθηση είναι εξαρτημένη από την ανθρώπινη παρέμβαση, σε αντίθεση με την μη εποπτευόμενη μάθηση όπως θα δούμε παρακάτω.



Εικόνα 1: Βήματα εποπτευόμενης μάθησης (Supervised Learning)

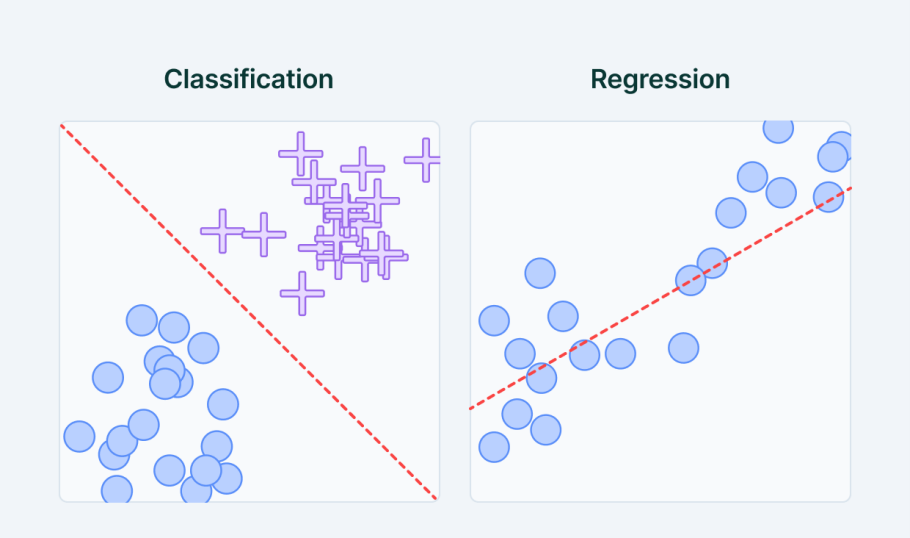
(Πηγή: https://www.v7labs.com/blog/supervised-vs-unsupervised-learning)

Η εποπτευόμενη μάθηση έχει κυρίως χρήση σε προβλήματα ταξινόμησης (classification) και παλινδρόμησης(regression):

- Στην ταξινόμηση(Classification), όπως προαναφέρθηκε, γίνεται ανάθεση ενός συνόλου δεδομένων εκπαίδευσης. Στη συνέχεια ο αλγόριθμος βγάζει συμπεράσματα σχετικά με τον τρόπο ταξινόμησης αυτών των μεταβλητών στις κλάσεις. Απαραίτητη προϋπόθεση είναι η μεταβλητή που επιθυμούμε να προβλέψουμε να είναι δίτιμη. Κάποιοι αλγόριθμοι classification είναι οι μηχανές διανυσμάτων υποστήριξης (SVM), το τυχαίο δάσος (Random Forest), τα δέντρα αποφάσεων (Decision Forests) και η

Λογιστική Παλινδρόμηση(Logistic Regression). Ένα πρόβλημα είναι για παράδειγμα η πρόβλεψη εάν κάποιος εργαζόμενος βάσει των συνθηκών εργασίας του, μπορεί να πάθει εργατικό ατύχημα.

- Στην παλινδρόμηση (Regression) η εποπτευόμενη μηχανική μάθηση χρησιμοποιείται προκειμένου να μας βοηθήσει να καταλάβουμε ποια είναι η σχέση μεταξύ των ανεξάρτητων και των εξαρτημένων μεταβλητών. Εάν η μεταβλητή που θέλουμε να προβλέψουμε είναι συνεχής τότε απαιτείται η χρήση αλγορίθμου παλινδρόμησης. Ένας κύριος αλγόριθμος παλινδρόμησης είναι η γραμμική παλινδρόμηση(linear regression). Ένα παράδειγμα πραγματικού προβλήματος στον κλάδο του ανθρώπινου δυναμικού(HRM) είναι η κατανόηση της σχέσης μεταξύ της επίδοσης των εργαζομένων(Employee Performance) με παράγοντες όπως η εκπαίδευση ή η ικανοποίηση των εργαζομένων.



Εικόνα 2: Γραφική αναπαράσταση ταξινόμησης και παλινδρόμησης

(Πηγή: https://www.v7labs.com/blog/supervised-vs-unsupervised-learning)

2.2.2 Μη εποπτευόμενη μηχανική μάθηση (Unsupervised Machine Learning)

Η μη εποπτευόμενη μάθηση πρόκειται για ένα κλάδο της μηχανικής μάθησης ο οποίος χαρακτηρίζεται από τη χρήση μη επισημασμένων δεδομένων (non labelled) τα οποία χρησιμοποιούνται για σκοπούς ανάλυσης και ομαδοποίησης του συνόλου των δεδομένων. Για να επιτευχθεί αυτό, οι αλγόριθμοι μαθαίνουν να αντιλαμβάνονται μοτίβα και τάσεις μεταξύ των δεδομένων, χωρίς να είναι απαιτητή η ανθρώπινη παρέμβαση. Σε αντίθεση με το εποπτευόμενο μοντέλο μάθησης το οποίο χρησιμοποιεί επισημασμένα δεδομένα εισόδου και

εξόδου για να προβλέψει ένα αποτέλεσμα για νέα δεδομένα, ένα μη εποπτευόμενο μοντέλο μάθησης μαθαίνει από το μη επισημασμένο μοντέλο εκπαίδευσης και προβλέπει πως θα ταξινομήσει τα δεδομένα. Έτσι, το μοντέλο της μη εποπτευόμενης μάθησης μπορεί να μας βοηθήσει να αντλήσουμε πληροφορίες από μεγάλου όγκους δεδομένων.

Η μη εποπτευόμενη μηχανική μάθηση έχει κυρίως χρήση σε προβλήματα κατηγοριοποίησης και μείωσης διαστάσεων:

- Στη συσταδοποίηση (Clustering) ο αλγόριθμος χρησιμοποιείται για διαχωρίσει τα ακατέργαστα δεδομένα σε ομάδες. Κάποιοι αλγόριθμοι κατηγοριοποίησης είναι ο K-Means, ο DBSCAN και ο BIRCH.
- Στη μείωση διαστάσεων (Dimensionality Reduction) ο αλγόριθμος χρησιμοποιεί τεχνικές προκειμένου να μειώσει των αριθμό των διαστάσεων σε προβλήματα που έχουμε μεγάλο όγκο δεδομένων. Κάποιοι αλγόριθμοι μείωσης διαστάσεων είναι η ανάλυση κυρίων συνιστωσών (Principal Components Analysis – PCA) και η Linear Discriminant Analysis (LDA).

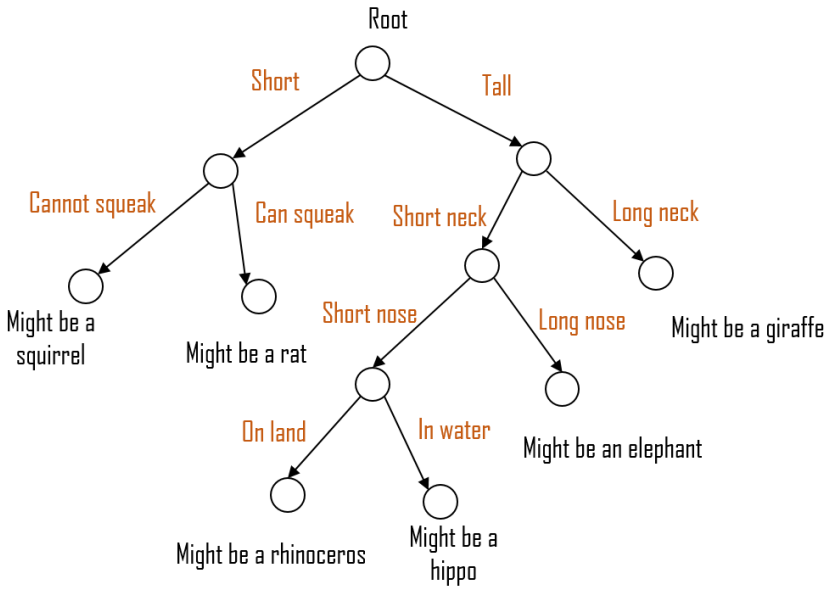
2.3 Αλγόριθμοι εποπτευόμενης μάθησης

Η μηχανική μάθηση περιλαμβάνει ένα μεγάλο πλήθος αλγορίθμων, οι οποίοι έχουν σχεδιαστεί για να εξυπηρετούν κάποιους σκοπούς. Για να επιλέξουμε κάποιον αλγόριθμο, θα πρέπει πρώτα να παρατηρηθούν σωστά τα δεδομένα που έχουμε διαθέσιμα προκειμένου να μπορούμε να μεγιστοποιήσουμε την απόδοσή τους. Παρακάτω θα δούμε τους αλγόριθμους εποπτευόμενης μάθησης οι οποίοι θα χρησιμοποιηθούν στην παρούσα διπλωματική εργασία για την ανάλυση του υπάρχοντος συνόλου δεδομένων (dataset).

2.3.1 Δέντρο Απόφασης (Decision Tree)

Το δέντρο απόφασης (Decision Tree) είναι ένας εποπτευόμενος αλγόριθμος μάθησης που μπορεί να χρησιμοποιηθεί για εργασίες ταξινόμησης αλλά και παλινδρόμησης. Έχει μια δομή δέντρου η οποία αποτελείται από τον κόμβο που βρίσκεται στην ρίζα (root node), τα κλαδιά (branches), τους εσωτερικούς κόμβους (Internal nodes) και τους κόμβους των φύλλων(Leaf nodes). Ένα δέντρο αποφάσεων ξεκινάει με τον ριζικό κόμβο, ο οποίος διαχωρίζει ένα χαρακτηριστικό του συνόλου δεδομένων με σκοπό να δημιουργήσει τους εσωτερικούς κόμβους (κόμβοι αποφάσεων). Στη συνέχεια, οι κόμβοι αποφάσεων με τη σειρά τους διαχωρίζουν το σύνολο των δεδομένων που έχουν σε υποσύνολα και η διαδικασία συνεχίζεται μέχρι το σύνολο των εγγραφών να κατανεμηθεί σε συγκεκριμένες κλάσεις.

Το μέγεθος των δέντρων παίζει μεγάλο ρόλο στην σωστή ταξινόμηση των δεδομένων .Τα δέντρα απόφασης προτιμώνται να παραμένουν μικρά σε μέγεθος καθώς όταν έχουν πολλούς κόμβους, αυξάνεται η πολυπλοκότητά τους και είναι δύσκολο να ερμηνευτούν.



Εικόνα 3: Το δέντρο απόφασης

(Πηγή: https://medium.com/geekculture/part-2-decision-trees-899894121249)

2.3.1.1 Αλγόριθμοι Δέντρων Απόφασης

Παρακάτω θα αναλυθούν οι πιο διαδεδομένοι αλγόριθμοι δέντρων αποφάσεων.

- **ID3**

Ο αλγόριθμος ID3 είναι ένας απλός αλγόριθμος δέντρου απόφασης που δημιουργήθηκε από τον Ross Quinlan το 1983. Ως κριτήριο διαμερισμού των κόμβων, ο ID3 χρησιμοποιεί το Information Gain. Η ανάπτυξη του δέντρου σταματάει όταν όλα τα δείγματα του δοσμένου κόμβου ανήκουν στην ίδια κατηγορία ή όταν δεν υπάρχουν άλλα γνωρίσματα ώστε να διαχωριστούν περαιτέρω. Ο συγκεκριμένος αλγόριθμος έχει μικρό υπολογιστικό κόστος και είναι απλός στην ερμηνεία. Παρόλα αυτά, στερείται τεχνικών κλαδέματος με αποτέλεσμα να μπορούν να προκύψουν μεγάλα δέντρα που είναι δύσκολο να ερμηνευτούν.

Ο υπολογισμός του Information Gain γίνεται από τον παρακάτω τύπο:

$$Information\ Gain = Εντροπία(κόμβου\ γονέα) - μέση\ Εντροπία(κόμβου\ παιδιών)$$

Ο υπολογισμός της εντροπίας (entropy) γίνεται από τον παρακάτω τύπο:

$$Εντροπία = \sum_{ι} -p_i \log_2(p_i)$$

όπου p_i είναι η πιθανότητα της κλάσης i σε ένα κόμβο.

Η εντροπία είναι ο τρόπος προσδιορισμού της ακαθαρσίας (impurity) ενός κόμβου. Εάν ένας κόμβος περιέχει μεγάλο αριθμό κλάσεων τότε ο κόμβος δεν είναι καθαρός. Εάν ένας κόμβος περιέχει μόνο μια κλάση τότε χαρακτηρίζεται ως καθαρός.

- **C4.5**

Ο C4.5 αναπτύχθηκε και αυτός από τον Ross Quinlan το1993 και πρόκειται για μια εξέλιξη του ID3. Σε αντίθεση με τον ID3, χρησιμοποιεί ως κριτήριο διαμερισμού το Gain Ratio. Επιπλέον ο συγκεκριμένος αλγόριθμος, επιδέχεται τεχνικών κλαδέματος οι οποίες αφαιρούν περιττά κλαδιά από το δέντρο και βοηθούν στην καλύτερη αποδοτικότητα του αλγόριθμου και στην αποφυγή του overfitting και της πολυπλοκότητας. Ο C4.5 μπορεί να χειριστεί διακριτές και συνεχείς μεταβλητές αλλά είναι κατάλληλος μόνο για μικρά datasets.

Ο υπολογισμός του Gain Ratio γίνεται από τον παρακάτω τύπο:

$$Gain\ Ratio = \frac{Information\ Gain}{-\sum_{i=2}^{n} D_i \log_2 Di}$$

Όπου n ο αριθμός των κόμβων διαχωρισμού της αρχικής μεταβλητής και Di ο αριθμός των εγγραφών που πηγαίνουν σε κόμβους παιδιά.

- **CART**

Ο αλγόριθμος CART (Classification And Regression Trees) αναπτύχθηκε από τους Leo Breiman, Jerome Friedman, Richard Olshen και Charles Stone το 1984. Ως κριτήριο διαχωρισμού ο CART χρησιμοποιεί το Gini impurity. Όπως και ο C4.5, ο αλγόριθμος CART χρησιμοποιεί και αυτός τεχνικές κλαδέματος προκειμένου να αποφευχθεί η υπερπροσαρμογή (overfitting) στα δεδομένα. Ως θετικό, ο συγκεκριμένος αλγόριθμος παράγει απλά δέντρα απόφασης και δεν επηρεάζεται από ακραίες τιμές (outliers). Παρόλα αυτά έχει την τάση να υπερπροσαρμόζεται στα δεδομένα.

Ο υπολογισμός του Gini impurity γίνεται από τον παρακάτω τύπο:

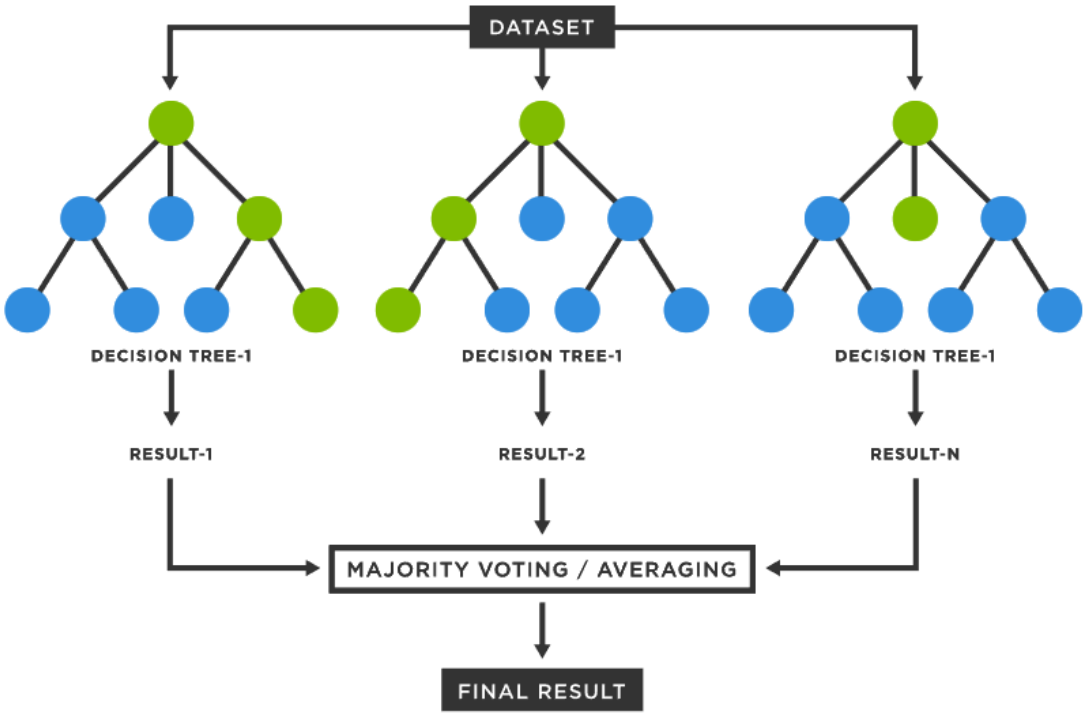
$$Gini\ impurity = \sum_{i=1}^{C} p(i)(1 - p(i))$$

Όπου C είναι ο αριθμός των κλάσεων και $p(i)$ είναι η πιθανότητα κατανομής ενός στοιχείου στην κλάση i.

2.3.2 Τυχαίο Δάσος (Random Forest)

Ο αλγόριθμος Random Forest είναι μια επέκταση της μεθοδολογίας των δέντρων αποφάσεων. Αρχικά δημιουργήθηκε το 1995 από τον Tin Kam Ho Και στη συνέχεια αναπτύχθηκε μια επέκταση του αλγορίθμου από τους Leo Breiman και Adele Cutler. Ο συγκεκριμένος αλγόριθμος, συνδυάζει πολλά δέντρα απόφασης και για την τελική πρόβλεψη, λαμβάνει υπόψιν τον μέσο όρο των προβλέψεων των δέντρων. Η παραπάνω διαδικασία έχει ως αποτέλεσμα τη δημιουργία ενός πιο σταθερού μοντέλου που δίνει μεγαλύτερη ακρίβεια στις προβλέψεις του σε σχέση με τα δέντρα απόφασης. Ο αλγόριθμος χρησιμοποιείται σε πολλούς τομείς όπως για παράδειγμα στον κλάδο του ανθρώπινου δυναμικού καθώς ο Random Forest μπορεί να καταγράψει αποτελεσματικά πολύπλοκες, μη γραμμικές σχέσεις μεταξύ αυτών των μεταβλητών και των αποτελεσμάτων των εργαζομένων, όπως η αποχώρηση ή η απόδοση (Tang, 2022).

Τα βήματα για τη λειτουργία του αλγόριθμου είναι τα ακόλουθα. Αρχικά, ο αλγόριθμος επιλέγει τυχαία δείγματα από το σύνολο δεδομένων. Στη συνέχεια κατασκευάζει ένα δέντρο απόφασης για κάθε δείγμα που έχει επιλεγεί στο προηγούμενο βήμα. Μετέπειτα, κάθε δέντρο παράγει μία πρόβλεψη. Τέλος, λαμβάνεται υπόψιν ο μέσος όρος των αποτελεσμάτων ώστε να καταλήξουμε στο αποτέλεσμα της ταξινόμησης.



Εικόνα 4: Βήματα εκτέλεσης του αλγορίθμου τυχαίου δάσους

(Πηγή: https://medium.com/@dishantkharkar9/about-random-forest-algorithms-62163357db25)

Ο αλγόριθμος Random Forest παρουσιάζει πολλά πλεονεκτήματα. Ένα από αυτά είναι ότι μπορεί να χρησιμοποιηθεί τόσο για προβλήματα κατηγοριοποίησης όσο και για προβλήματα παλινδρόμησης. Επίσης, έχει την τάση να μην υπερπροσαρμόζεται στα δεδομένα και παρέχει μεγάλο ποσοστό ακρίβειας στις προβλέψεις του. Τέλος, εκτελείται αποτελεσματικά σε μεγάλες βάσεις δεδομένων και μπορεί να παράγει πολύ καλές προβλέψεις εκτιμώντας τα δεδομένα που λείπουν (missing values).

Παρόλα τα πλεονεκτήματά του, ο συγκεκριμένος αλγόριθμος δεν παύει να έχει και κάποια μειονεκτήματα. Ένα από αυτά είναι ότι ένας μεγάλος αριθμός δέντρων μπορεί να κάνει τον αλγόριθμό αναποτελεσματικό και ταυτοχρόνως αργό. Τέλος, οι αλγόριθμοι μπορούν να εκπαιδεύονται με μεγάλη ταχύτητα αλλά στερούνται ταχύτητας στη φάση της πρόβλεψης.

2.3.3. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines – SVM) είναι ένας γραμμικός και μη γραμμικός αλγόριθμος εποπτευόμενης μάθησης ο οποίος χρησιμοποιείται για εργασίες ταξινόμησης και παλινδρόμησης. Αναπτύχθηκε από τον Vladimir Vapnik κατά τη διάρκεια της δεκαετίας του 1990. Κατά τη χρήση του αλγορίθμου, κάθε δεδομένο αναπαρίσταται ως ένα σημείο στο χώρο. Στη συνέχεια, κατά τη διαδικασία της ταξινόμησης, βρίσκουμε το μέγιστο υπερεπίπεδο (hyper plane), το οποίο διαχωρίζει καλύτερα τα δεδομένα μας σε δύο κλάσεις. Όπως είναι φυσικό μπορούν να δημιουργηθούν άπειρα υπερεπίπεδα που διαχωρίζουν τα δεδομένα. Για την εύρεση του καταλληλότερου υπερεπίπεδου, χρησιμοποιούμε το υπερεπίπεδο με το μεγαλύτερο περιθώριο (margin). Το περιθώριο είναι η μέγιστη απόσταση μεταξύ των πλησιέστερων σημείων κάθε κλάσης με το υπερεπίπεδο με την προϋπόθεση ότι τα δεδομένα ταξινομούνται σωστά. Τα πλησιέστερα στο υπερεπίπεδο σημεία, ονομάζονται διανύσματα στήριξης (support vectors).

Η εξίσωση του υπερεπίπεδου είναι της μορφής $w \cdot x + b = 0$ όπου το w και το x είναι οι παράμετροι του μοντέλου και $X = \{x_1, \dots, x_n\}$ το σύνολο των δεδομένων εκπαίδευσης. Οι κλάσεις ορίζονται ως $C = \{c_1, c_2\}$.

Επίσης ισχύει ότι εάν:

- $w \cdot x + b > 0$, τα σημεία δεδομένων βρίσκονται πάνω από το υπερεπίπεδο
- $w \cdot x + b < 0$, τα σημεία δεδομένων βρίσκονται κάτω από το υπερεπίπεδο

Τα παράλληλα υπερεπίπεδα εκφράζονται ως:

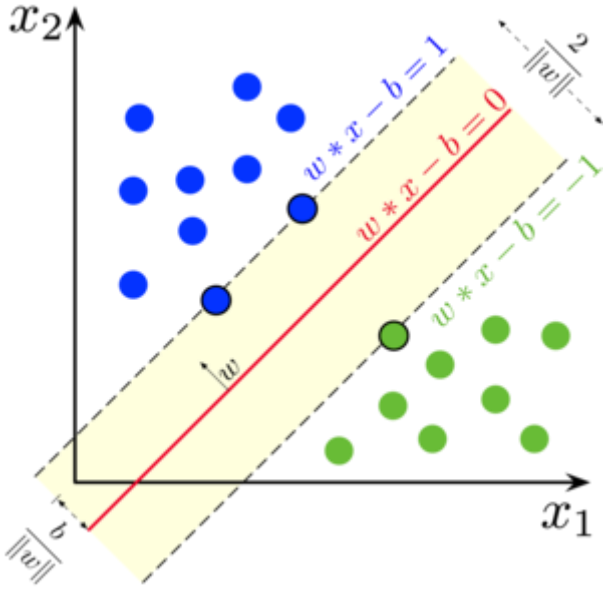
- Παράλληλο υπερεπίπεδο που ορίζει το πάνω όριο: $w \cdot x_1 + b = 1$
- Παράλληλο υπερεπίπεδο που ορίζει το κάτω όριο: $w \cdot x_2 + b = -1$

Για να προκύψει το περιθώριο, αφαιρούμε κατά μέλη τις δυο παραπάνω εξισώσεις και προκύπτει ότι

$$margin = \frac{2}{\|w\|}$$

Η απόσταση του υπερεπίπεδου από την αρχή των αξόνων δίνεται από τον τύπο $\frac{b_0}{\|w_0\|}$.

Όταν $b_0 > 0$, η αρχή των αξόνων είναι από τη θετική πλευρά του υπερεπίπεδου, όταν $b_0 < 0$ από την αρνητική ενώ όταν $b_0 = 0$ τότε το υπερεπίπεδο περνά από την αρχή των αξόνων



Εικόνα 5: Γραφική απεικόνιση λειτουργίας SVM

(Πηγή: https://en.wikipedia.org/wiki/Support_vector_machine)

Για να μπορεί ο γραμμικός ταξινομητής SVM να ταξινομεί και μη γραμμικά διαχωρίσιμα δεδομένα, επινοήθηκε το κόλπο πυρήνα (kernel trick). Ο συγκεκριμένος πυρήνας πρόκειται για συναρτήσεις που λαμβάνουν ως είσοδο χώρους μικρών διαστάσεων και τους μετατρέπει σε χώρο υψηλότερων διαστάσεων. Δηλαδή μετατρέπει το μη διαχωρίσιμο πρόβλημα σε διαχωρίσιμο (Ray, 2024).

Παρακάτω θα δούμε τις κυριότερες συναρτήσεις πυρήνα:

1. **Γραμμικός πυρήνας (Linear Kernel)**

Η χρήση του είναι για γραμμικά δεδομένα. Βοηθά στην αναπαράσταση των δεδομένων με τη χρήση γραμμικής σχέσης.

$$k(x, z) = x^T \cdot z$$

2. Πολυωνυμικός πυρήνας (Polynomial Kernel)

Μετασχηματίζει τα μη γραμμικά δεδομένα σε δεδομένα μεγαλύτερης διάστασης.

$$k(x, z) = (ax^T y + c)^d$$

όπου α είναι η κλίση, d ο πολυωνυμικός βαθμός και c μια σταθερά.

3. RBF πυρήνας (Radial Basis Function Kernel)

Είναι ο πιο ευρέως χρησιμοποιούμενος πυρήνας. Κι αυτός, όπως και ο πολυωνυμικός πυρήνας μετασχηματίζει τα μη γραμμικά δεδομένα σε δεδομένα μεγαλύτερης διάστασης.

$$k(x, z) = \exp(-\gamma \|x - z\|)$$

όπου $\gamma = \frac{1}{2\sigma^2}$.

4. Σιγμοειδής πυρήνας (Sigmoid Kernel)

Ο συγκεκριμένος πυρήνας ονομάζεται και ως πυρήνας υπερβολικής εφαπτομένης και προέρχεται από το πεδίο των νευρωνικών δικτύων.

$$k(x, z) = \tanh(ax^T z + c)$$

όπου α η κλίση και c μια σταθερά.

Κεφάλαιο 3

Λογιστική Παλινδρόμηση

3.1 Εισαγωγή

Μέσω της διαδικασίας που λέγεται παλινδρόμηση, μας δίνεται η δυνατότητα να ερευνούμε τη συσχέτιση που έχει μια μεταβλητή απόκρισης (Y) με μία ή περισσότερες επεξηγηματικές μεταβλητές (X_i , $i = 1,2 \dots, k$).

Η πιο βασική μέθοδος παλινδρόμησης είναι η γραμμική παλινδρόμηση. Για να μπορέσουμε να την χρησιμοποιήσουμε όμως βασική προϋπόθεση είναι η μεταβλητή απόκρισης και τα σφάλματα να ακολουθούν την κανονική κατανομή. Υπάρχουν όμως περιπτώσεις που δεν ικανοποιούνται τα παραπάνω κριτήρια. Έτσι λοιπόν προέκυψε το λογιστικό μοντέλο, το οποίο έχει διακριτή (ή δίτιμη) μεταβλητή απόκρισης και τα σφάλματα δεν ακολουθούν την κανονική κατανομή.

Η λογιστική παλινδρόμηση, που αποτελεί γενίκευση της απλής γραμμικής παλινδρόμησης, χρησιμοποιείται όταν θέλουμε να προβλέψουμε την εμφάνιση ή όχι ενός συγκεκριμένου γεγονότος. Όπως για παράδειγμα στην περίπτωσή μας, εάν αποχώρησε ή συνεχίζει να εργάζεται στην εταιρεία ένας εργαζόμενος.

Το γραμμικό μοντέλο είναι αδύνατο να χρησιμοποιηθεί, όταν η μεταβλητή Y είναι δυαδική και έχουμε τα εξής τρία προβλήματα:

1. Τα σφάλματα δεν είναι κανονικά.
2. Τα σφάλματα έχουν άνισες διασπορές.
3. Περιορισμός στη συνάρτηση απόκρισης (η προβλεπόμενη πιθανότητα θα πρέπει να ανήκει στα διάστημα (0,1))

Παρόλο που στα δύο πρώτα προβλήματα είναι δυνατό σε κάποιες περιπτώσεις να τα παραλείψουμε και να χρησιμοποιήσουμε την γραμμική παλινδρόμηση, εφαρμόζοντας κάποιες άλλες τεχνικές, το τρίτο πρόβλημα μας το απαγορεύει ρητά, γιατί δεν μπορεί να αντιμετωπιστεί με διαφορετικό τρόπο.

Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική), στη δε δεύτερη αποκλειστικά ποσοτική και συνεχής. Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων α και β_i γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η

εκτίμηση των παραμέτρων γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας, δηλαδή επιλέγονται οι πιο πιθανοφανείς εκτιμήσεις των παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα. Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων, ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν.

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης, ανάλογα με την φύση της εξαρτημένης κατηγορικής μεταβλητής η οποία μπορεί να είναι:

1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία.

2. Διατάξιμη (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες, μεταξύ των οποίων ισχύει η έννοια της διάταξης, όπως π.χ. σε μια ερώτηση της κλίμακας διαφωνώ : καθόλου, λίγο, μέτρια, αρκετά, πολύ.

3. Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες, χωρίς κάποια φυσική διαβάθμιση, όπως π.χ. το χρώμα αντικειμένων.

(Καλλιακμάνης. 2020)

3.2 Το μοντέλο logit

Η λογιστική παλινδρόμηση είναι το γενικευμένο γραμμικό μοντέλο για δίτιμες αποκρίσεις με συνάρτηση σύνδεσης την logit.

Αν Y μια δίτιμη απόκριση με $P(Y = 1) = p = E(Y)$ το μοντέλο λογιστική παλινδρόμησης εκφράζεται ως εξής:

$$logit(p) \equiv \log\left(\frac{p}{1-p}\right) \equiv b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

όπου $x_1, x_2, \dots, x_k$ είναι οι επεξηγηματικές μεταβλητές του μοντέλου.

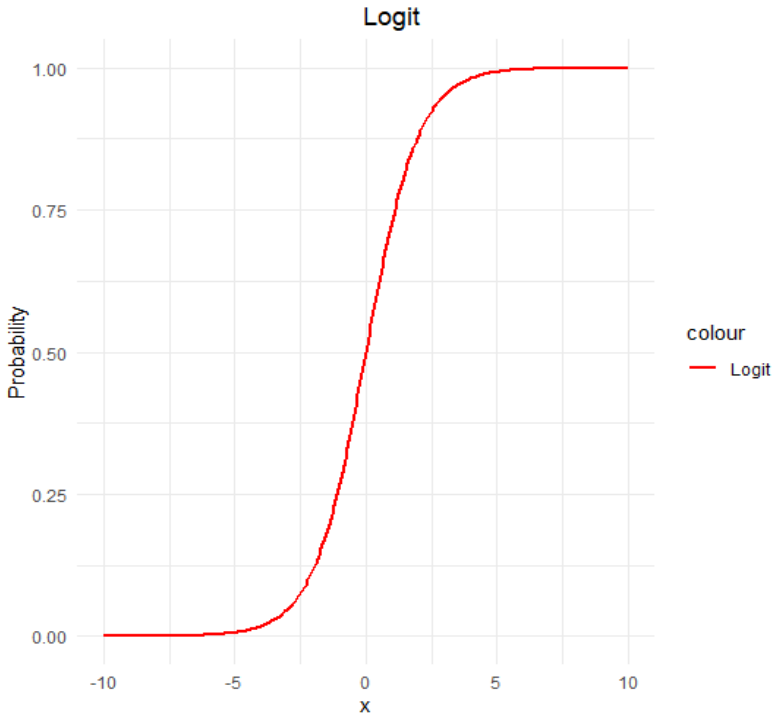
Εάν αντιστρέψουμε την logit, μπορούμε να παρατηρήσουμε πως το μοντέλο της λογιστικής παλινδρόμησης εκφράζει την πιθανότητα:

$$p = \frac{e^{b_0+b_1\cdot x_1+\cdots+b_k x_k}}{1+e^{b_0+b_1\cdot x_1+\cdots+b_k x_k}}$$

ή ισοδύναμα τη σχετική πιθανότητα (odds) ως:

$$\frac{p}{1-p} = e^{b_0+b_1\cdot x_1+\cdots+b_k x_k}$$

Η ονομασία του προέρχεται από το log-arithmic un-it. Η συνάρτηση logit, είναι ο λογάριθμος της σχετικής πιθανότητας για ένα γεγονός. Δηλαδή ο λογάριθμος της πιθανότητας να συμβεί ένα γεγονός προς την πιθανότητα να μη συμβεί αυτό το γεγονός. Οι συντελεστές b μας δείχνουν τη μεταβολή της τιμής του logit. Συγκεκριμένα, εάν αυξηθεί ο συντελεστής x_1 κατά μία μονάδα τότε η σχετική πιθανότητα επιτυχίας θα αυξηθεί πολλαπλασιαστικά κατά e^{b_1}. Λόγω της εύκολης ερμηνείας των αποτελεσμάτων με βάση τη σχετική πιθανότητα, η logit προτιμάται σε σχέση με τις άλλες συναρτήσεις σύνδεσης που θα δούμε παρακάτω. (Πολίτης, 2023)



Εικόνα 6: Συνάρτηση logit

Όπως φαίνεται και στην παραπάνω γραφική παράσταση, η Logit είναι γνησίως αύξουσα συνάρτηση.

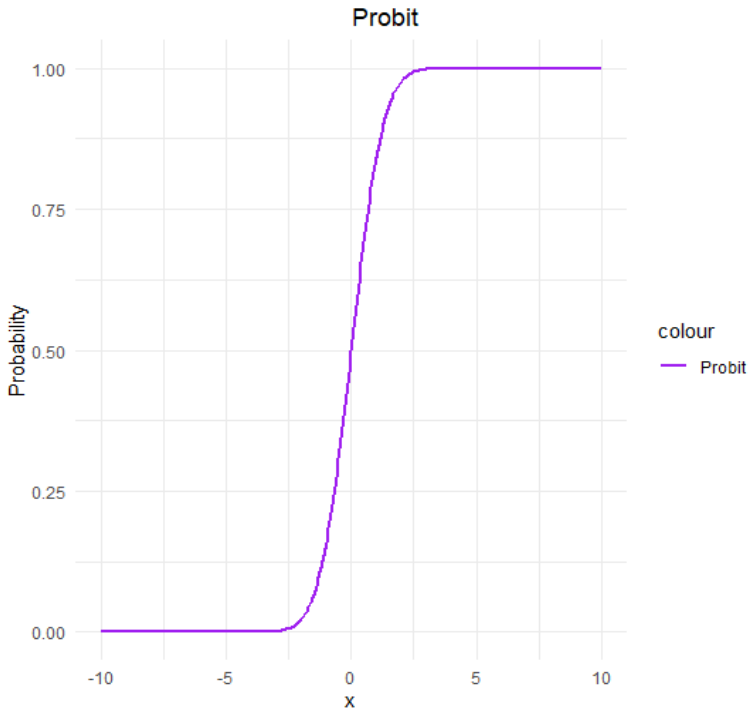
3.3 Το μοντέλο probit

Η ιδέα της δημιουργίας του probit δημοσιεύθηκε από τον Chester Ittner Bliss το 1034 στο περιοδικό Science και αφορούσε τον τρόπο αντιμετώπισης δεδομένων όπως το ποσοστό παρασίτων που σκοτώθηκαν από ένα φυτοφάρμακο. Η ονομασία του προκύπτει από τον συνδυασμό των λέξεων prob-ability un-it. Τα μοντέλα probit είναι ευρέως χρησιμοποιούμενα στον τομέα της οικονομετρίας όπου χρειάζεται να παρθεί η απόφαση για αφορά ενός προϊόντος. Επίσης, τα μοντέλα probit χρησιμοποιούνται στον τομέα της βιοστατιστικής όπου χρειάζεται να γνωρίζουμε εάν ένας ασθενής αντιδρά καλά σε ένα φάρμακο ή όχι. Η προσαρμογή των μοντέλων probit, δεν έχει μεγάλη διαφορά από την προσαρμογή των μοντέλων logit. Η πιο σημαντική διαφορά είναι ότι το μοντέλο probit χρησιμοποιεί την αθροιστική Gaussian κατανομή ενώ τα logit χρησιμοποιούν την λογιστική συνάρτηση.

Υποθέτουμε ότι η εξαρτημένη μεταβλητή Y είναι δίτιμη και λαμβάνει τιμές 0 και 1. Η X είναι ένα διάνυσμα μεταβλητών η οποία επηρεάζει το αποτέλεσμα της Y με συντελεστές $b_0, b_1, \dots, b_k$.
Η μορφή του μοντέλου θα είναι:

$$P(Y = 1|X) = \Phi(b_0 + b_1 \cdot x_1 + \cdots + b_k \cdot x_k)$$

όπου Φ η αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής.



Εικόνα 7: Συνάρτηση probit

3.4 Το μοντέλο clog-log

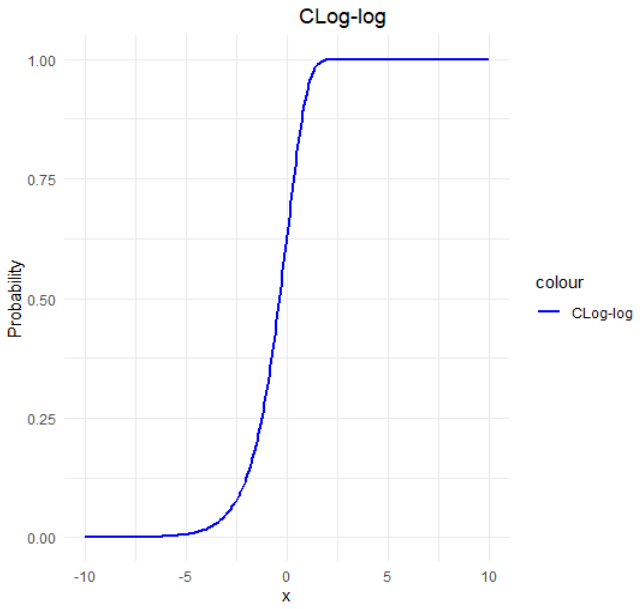
Το complementary log log (clog-log) μοντέλο είναι μια επέκταση του μοντέλου της λογιστικής παλινδρόμησης και είναι χρήσιμη όταν η πιθανότητα του συμβάντος είναι πολύ μικρή ή πολύ μεγάλη. Το μοντέλο συνήθως χρησιμοποιείται για την αντιμετώπιση γεγονότων που τα δεδομένα δεν είναι συμμετρικά εντός του διαστήματος [0,1] και αυξάνονται αργά σε μικρές έως μέτριες τιμές αλλά απότομα κοντά στο 1.

Το μοντέλο clog-log έχει τη μορφή:

$$c\,log\,log(p) \equiv log(-log(1-p_x)) = b_0 + b_1x_1 + \cdots + b_kx_k$$

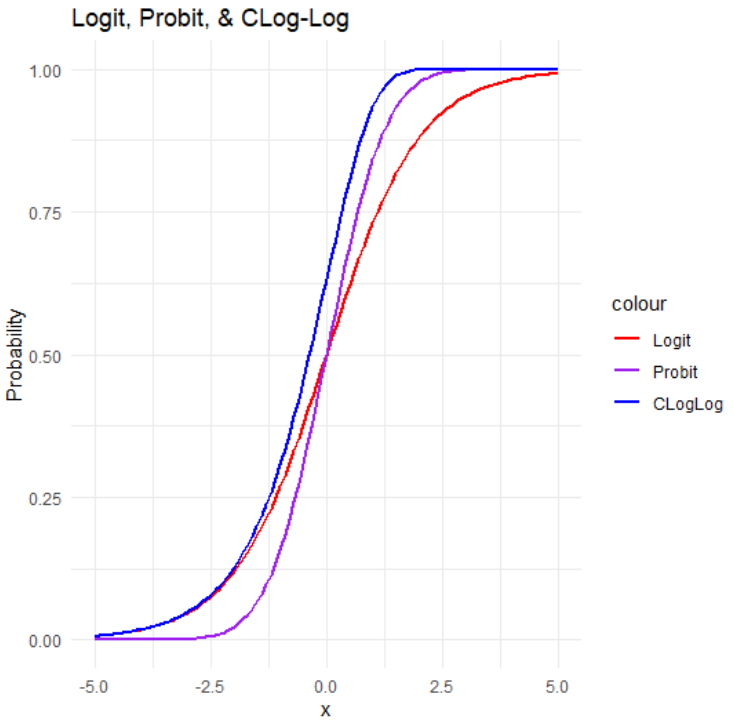
όπου $p_x = P(Y = 1|X = x)$

Στο παρακάτω γράφημα, μπορούμε πράγματι να δούμε ότι το $P(Y = 1)$ προσεγγίζει το αργά το 0 και πλησιάζει γρήγορά το 1.



Εικόνα 8: Συνάρτηση clog-log

Παρακάτω παρουσιάζεται διάγραμμα και των τριών μοντέλων μαζί.



Εικόνα 9: Διάγραμμα συναρτήσεων logit, probit & C Log-log

Στο παραπάνω διάγραμμα είναι προφανές πως εκτός από την logit που αναφέρθηκε και νωρίτερα, γνησίως αύξουσες συναρτήσεις είναι και η probit αλλά και η clog-log.

Κεφάλαιο 4

Παρουσίαση των δεδομένων, περιγραφικά μέτρα και διαγραμματικές απεικονίσεις

4.1 Παρουσίαση των δεδομένων

Τα σύνολο δεδομένων το οποίο θα χρησιμοποιηθεί στην παρούσα διπλωματική εργασία περιλαμβάνει 1470 εγγραφές στις οποίες υπάρχουν 32 στήλες με πληροφορίες για κάθε εγγραφή. Κάθε εγγραφή περιγράφει έναν εργαζόμενο της εταιρείας IBM για τον οποίο υπάρχουν πληροφορίες σχετικά με το εργασιακό-προσωπικό του προφίλ. Το σύνολο των δεδομένων έχει δημιουργηθεί από τους data scientists της IBM και τα δεδομένα έχουν αναρτηθεί στην ιστοσελίδα Kaggle. Το σύνολο των δεδομένων είναι διαθέσιμο προς χρήση στα πλαίσια πρακτικής εξάσκησης σε μεθόδους πρόβλεψης.

Παρακάτω παρουσιάζονται οι μεταβλητές του συνόλου δεδομένων.

Μεταβλητή	Περιγραφή
AGE	Ηλικία
ATTRITION	Ένδειξη εάν ο εργαζόμενος αποχώρησε από την εταιρεία (0=No, 1=Yes)
BUSINESS TRAVEL	Συχνότητα επαγγελματικών ταξιδιών (0=No Travel, 1=Travel Frequently, 2=Tavel Rarely)
DEPARTMENT	Τμήμα Απασχόλησης (0=HR, 1=R&D, 2=Sales)
DISTANCE FROM HOME	Απόσταση τόπου παροχής εργασίας από το σπίτι(σε μίλια).
EDUCATION	Επίπεδο εκπαίδευσης (1=Below College, 2=College, 0=Bachelor, 4=Master, 3=Doctor)
EDUCATION FIELD	Πεδίο εκπαίδευσης (0=HR, 1=LIFE SCIENCES, 2=MARKETING, 3=MEDICAL SCIENCES, 4=OTHER, 5= TECHNICAL)
EMPLOYEE COUNT	Μετρητής εργαζόμενου
EMPLOYEE NUMBER	Κωδικός εργαζόμενου
ENVIRONMENT SATISFACTION	Επίπεδο ευχαρίστησης με το περιβάλλον στην εργασία (1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)
GENDER	Φύλο (0=FEMALE, 1=MALE)
JOB INVOLVEMENT	Επίπεδο εργασιακής συμμετοχής (1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)
JOB LEVEL	Επίπεδο εργασιακής θέσης (4=VERY LOW, 1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)
JOB ROLE	Θέση εργασίας (0=HC REP, 1=HR, 2=LAB TECHNICIAN, 3=MANAGER, 4= MANUFACTORING DIRECTOR, 5= RESEARCH DIRECTOR, 6= RESEARCH SCIENTIST, 7=SALES EXECUTIVE, 8= SALES REPRESENTATIVE)
JOB SATISFACTION	Ευχαρίστηση με την εργασία (1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)

MARITAL STATUS	Οικογενειακή κατάσταση (0=DIVORCED, 1=MARRIED, 2=SINGLE)
MONTHLY INCOME	Μηνιαίος μισθός
NUMCOMPANIES WORKED	Αριθμός εταιρειών στις οποίες έχει εργαστεί
OVER 18	Ένδειξη εάν είναι άνω των 18 ετών (1=YES, 2=NO)
OVERTIME	Ένδειξη εάν εργάζεται υπερωρίες(0=NO, 1=YES)
PERCENT SALARY HIKE	Ποσοστό αύξησης του μισθού σε σχέση με τον αρχικό μισθό στην εταιρεία
PERFORMANCE RATING	Βαθμολογία επίδοσης στην εργασία (3=LOW, 2=MEDIUM, 0=HIGH, 1=VERY HIGH)
RELATIONSHIP SATISFACTION	Επίπεδο ικανοποίησης εργασιακών σχέσεων (1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)
STANDARD HOURS	Ώρες εργασίας
STOCK OPTIONS LEVEL	Επίπεδο δυνατότητας αγοράς μετοχών (1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)
TOTAL WORKING YEARS	Συνολικά χρόνια εργασίας
TRAINING TIMES LAST YEAR	Φορές εκπαίδευσης κατά το τελευταίο έτος
WORK LIFE BALANCE	Επίπεδο ισορροπίας μεταξύ εργασιακού και προσωπικού χρόνου (1=LOW, 2=MEDIUM, 0=HIGH, 3=VERY HIGH)
YEARS AT COMPANY	Πόσα χρόνια εργάζεται στην εταιρεία
YEARS IN CURRENT ROLE	Πόσα χρόνια ο εργαζόμενος βρίσκεται σε αυτή τη θέση
YEARS SINCE LAST PROMOTION	Πόσα χρόνια έχουν περάσει από την τελευταία προαγωγή
YEARS WITH CURRENT MANAGER	Πόσα χρόνια έχουν περάσει έχοντας τον ίδιο Manager

Πίνακας 1: Περιγραφή των μεταβλητών

4.2 Έλεγχος για ελλιπείς τιμές και περιγραφικά μέτρα

4.2.1 Έλεγχος για ελλιπείς τιμές (missing values)

Ένα από τα συνηθέστερα εμπόδια που μπορούμε να συναντήσουμε κατά την έναρξη της ανάλυσης συνόλων δεδομένων είναι η ύπαρξη ελλιπών τιμών (missing values). Πρόκειται για τιμές οι οποίες απουσιάζουν από το σύνολο των δεδομένων και αν δεν αντιμετωπιστούν, μπορεί να προκαλέσουν προβλήματα ως προς την αποτελεσματικότητα των αλγορίθμων και των αποτελεσμάτων. Υπάρχουν διάφορες μέθοδοι αντιμετώπισης αυτού του προβλήματος. Μερικές από αυτές είναι οι παρακάτω:

- **Διαγραφή**: Στη συγκεκριμένη μέθοδο διαγράφεται η γραμμή ή η στήλη του συνόλου δεδομένων που έχει ελλιπείς τιμές. Ο κίνδυνος με τη χρήση αυτής της μεθόδου είναι η απώλεια αρκετών δεδομένων ιδίως όταν υπάρχουν πολλές ελλιπείς τιμές.
- **Αντικατάσταση με τιμή**: Αυτή η μέθοδος περιλαμβάνει την αντικατάσταση των ελλιπών τιμών με μια τιμή βάσει των διαθέσιμων δεδομένων. Αυτή η τιμή μπορεί να είναι η διάμεσος, η μέση τιμή και άλλες.

- **Ξεχωριστή κατηγορία**: Εάν το σύνολο δεδομένων μας έχει κατηγορικές μεταβλητές οι οποίες περιλαμβάνουν κατηγορικές μεταβλητές τότε αυτές οι τιμές μπορούν να αντιμετωπιστούν ως ξεχωριστή κατηγορία.

  Το σύνολο των δεδομένων που θα χρησιμοποιηθεί στην παρούσα διπλωματική εργασία, δεν περιλαμβάνει ελλιπείς τιμές.

4.2.2 Περιγραφικά μέτρα

Στον παρακάτω πίνακα παρουσιάζονται τα περιγραφικά μέτρα για τις αριθμητικές μεταβλητές. Συγκεκριμένα, παρουσιάζονται για κάθε μεταβλητή κατά σειρά, ο συνολικός αριθμός των εγγραφών, η μέση τιμή, η τυπική απόκλιση, η ελάχιστη τιμή, το πρώτο τεταρτημόριο, η διάμεσος, το τρίτο τεταρτημόριο και η μέγιστη τιμή. Έχουν διαγραφεί οι μεταβλητές StandardHours, Over18 και EmployeeCounter καθώς λαμβάνουν μόνο μία τιμή. Επίσης, έχει διαγραφεί και η μεταβλητή EmployeeNumber καθώς περιλαμβάνει τον μοναδικό αριθμό μητρώου κάθε εργαζόμενου.

	count	mean	std	min	25%	50%	75%	max
Age	1470,00	36,92	9,14	18,00	30,00	36,00	43,00	60,00
DistanceFromHome	1470,00	9,19	8,11	1,00	2,00	7,00	14,00	29,00
MonthlyIncome	1470,00	6502,93	4707,96	1009,00	2911,00	4919,00	8379,00	19999,00
NumCompaniesWorked	1470,00	2,69	2,50	0,00	1,00	2,00	4,00	9,00
PercentSalaryHike	1470,00	15,21	3,66	11,00	12,00	14,00	18,00	25,00
TotalWorkingYears	1470,00	11,28	7,78	0,00	6,00	10,00	15,00	40,00
TrainingTimesLastYear	1470,00	2,80	1,29	0,00	2,00	3,00	3,00	6,00
YearsAtCompany	1470,00	7,01	6,13	0,00	3,00	5,00	9,00	40,00
YearsInCurrentRole	1470,00	4,23	3,62	0,00	2,00	3,00	7,00	18,00
YearsSinceLastPromotion	1470,00	2,19	3,22	0,00	0,00	1,00	3,00	15,00
YearsWithCurrManager	1470,00	4,12	3,57	0,00	2,00	3,00	7,00	17,00

Πίνακας 2: Περιγραφικά μέτρα αριθμητικών μεταβλητών

Στη συνέχεια, παρατίθενται τα περιγραφικά μέτρα για τις κατηγορικές μεταβλητές.

	count	unique	top	freq
Attrition	1470	2	No	1233
BusinessTravel	1470	3	Travel_Rarely	1043
Department	1470	3	Research & Development	961
Education	1470	5	Bachelor	572

EducationField	1470	6	Life Sciences	606
EnvironmentSatisfaction	1470	4	High	453
Gender	1470	2	Male	882
JobInvolvement	1470	4	High	868
JobLevel	1470	5	Very Low	543
JobRole	1470	9	Sales Executive	326
JobSatisfaction	1470	4	Very High	459
MaritalStatus	1470	3	Married	673
OverTime	1470	2	No	1054
PerformanceRating	1470	2	High	1244
RelationshipSatisfaction	1470	4	High	459
StockOptionLevel	1470	4	Low	631
WorkLifeBalance	1470	4	High	893

Πίνακας 3: Περιγραφικά μέτρα κατηγορικών μεταβλητών

Πιο συγκεκριμένα, ο Πίνακας 3 περιλαμβάνει μετρητή των συνολικών τιμών, ένδειξη για το πόσες μοναδικές τιμές μπορεί να έχει κάθε κατηγορική μεταβλητή, την τιμή με την μεγαλύτερη εμφάνιση και την συχνότητα εμφάνισης της επικρατούσας τιμής της μεταβλητής.

Column	Value	Count
Attrition	**Yes**	237
	No	1233
BusinessTravel	**Non-Travel**	150
	Travel_Frequently	277
	Travel_Rarely	1043
Department	**Human Resources**	63
	Sales	446
	Research & Development	961
Education	**Doctor**	48
	Below College	170
	College	282
	Master	398
	Bachelor	572
EducationField	**Human Resources**	27
	Other	82
	Technical Degree	132
	Marketing	159
	Medical	464
	Life Sciences	606
EnvironmentSatisfaction	**Low**	284
	Medium	287
	Very High	446
	High	453

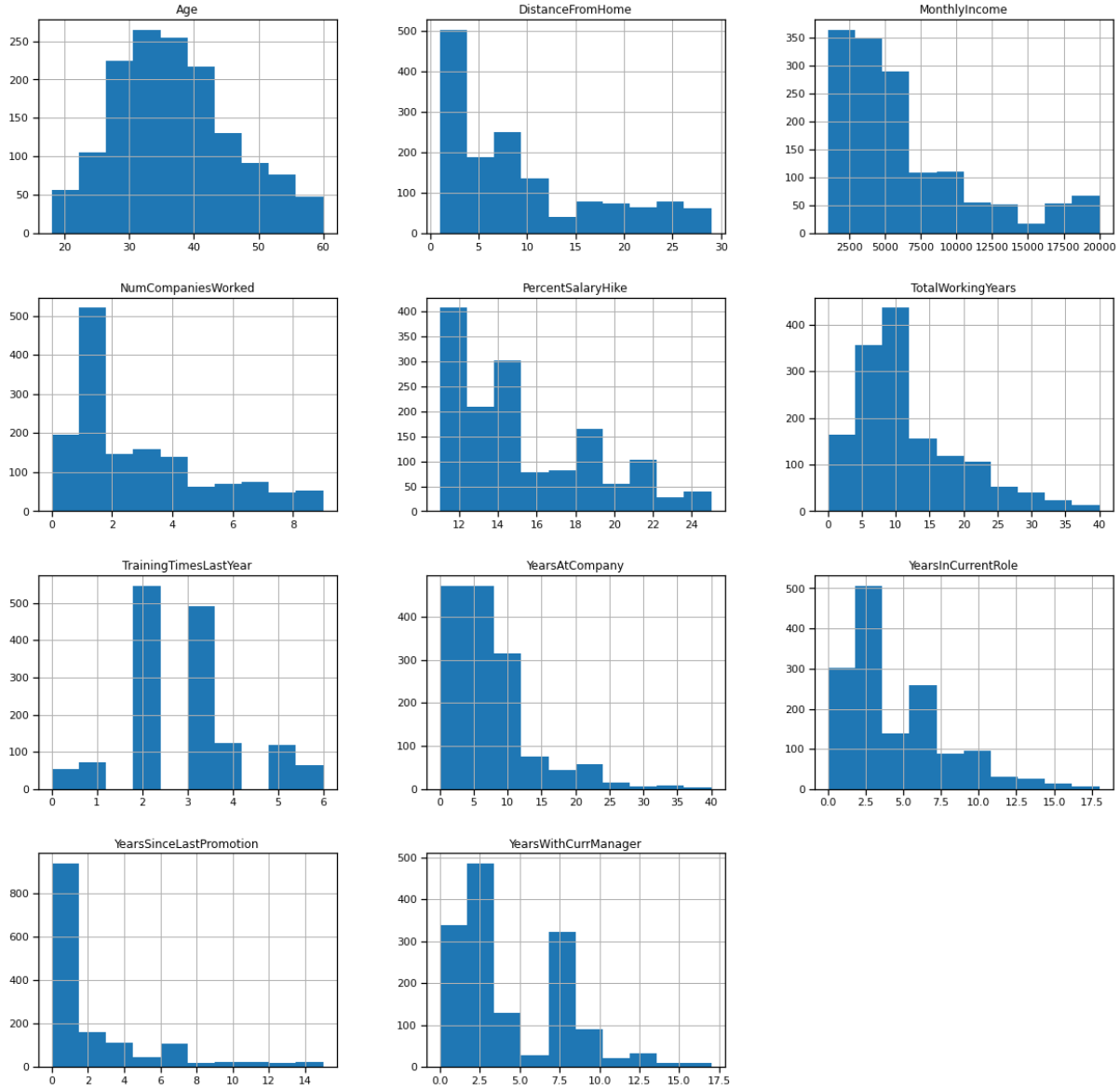
Gender	**Female**	588
	Male	882
JobInvolvement	**Low**	83
	Very High	144
	Medium	375
	High	868
JobLevel	**Very High**	69
	High	106
	Medium	218
	Low	534
	Very Low	543
JobRole	**Human Resources**	52
	Research Director	80
	Sales Representative	83
	Manager	102
	Healthcare Representative	131
	Manufacturing Director	145
	Laboratory Technician	259
	Research Scientist	292
	Sales Executive	326
JobSatisfaction	**Medium**	280
	Low	289
	High	442
	Very High	459
MaritalStatus	**Divorced**	327
	Single	470
	Married	673
OverTime	**Yes**	416
	No	1054
PerformanceRating	**VERY HIGH**	226
	HIGH	1244
RelationshipSatisfaction	**Low**	276
	Medium	303
	Very High	432
	High	459
StockOptionLevel	**Very High**	85
	High	158
	Medium	596
	Low	631
WorkLifeBalance	**Low**	80
	Very High	153
	Medium	344
	High	893

Πίνακας 4: Μέτρηση τιμών ανά κατηγορική μεταβλητή

Ο παραπάνω πίνακας 4 περιλαμβάνει τον αριθμό εμφάνισης των τιμών των κατηγορικών μεταβλητών.

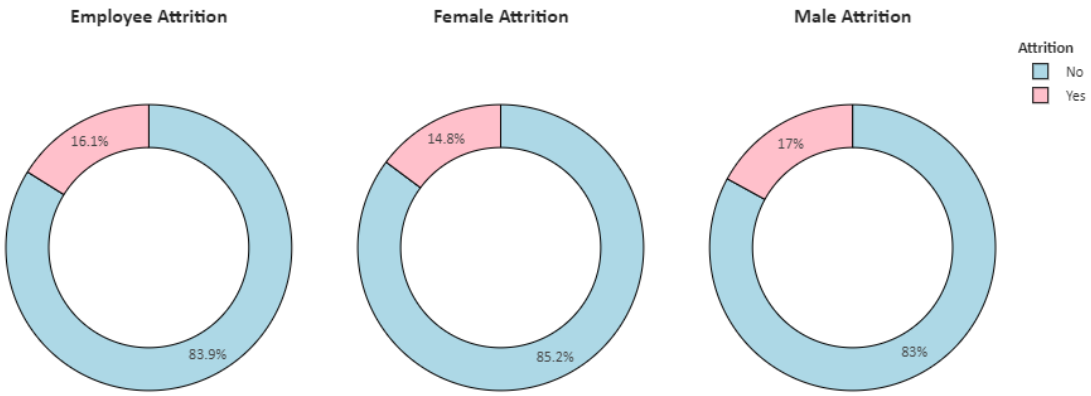
4.3 Διαγραμματικές απεικονίσεις

Παρακάτω, παρουσιάζονται ιστογράμματα που έχουν δημιουργηθεί για τις αριθμητικές μεταβλητές του συνόλου των δεδομένων.



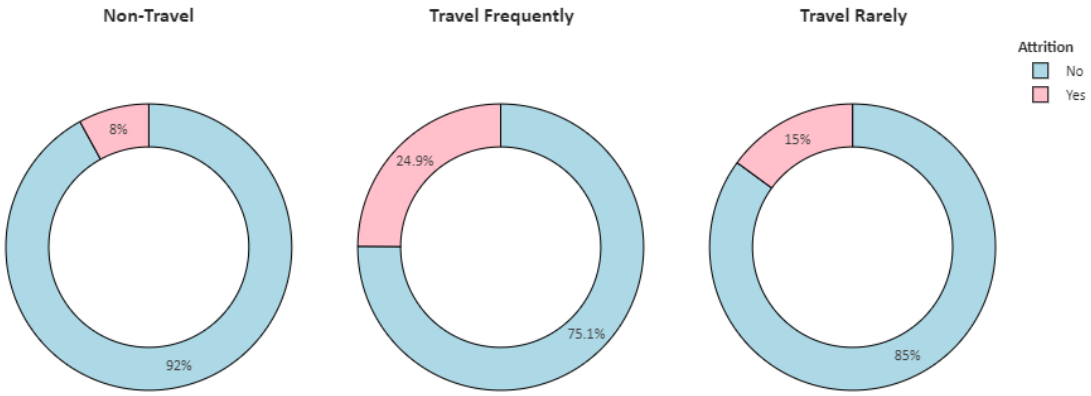
Εικόνα 10: Ιστογράμματα αριθμητικών μεταβλητών

Από τα παραπάνω ιστογράμματα παρατηρούμε πως για παράδειγμα οι μεταβλητές YearsAtCompany, MonthlyIncome και DistanceFromHome δεν κατανέμονται συμμετρικά αλλά λοξά δεξιά. Σχετικά με την ηλικία (μεταβλητή Age) μπορούμε να πούμε πως το μεγαλύτερο μέρος του προσωπικού φαίνεται να είναι μεταξύ 25 και 45 ετών. Αντίστοιχα, φαίνεται πως η πλειοψηφία του προσωπικού έχει εργαστεί μόνο σε 1 εταιρεία πριν προσληφθεί στην εταιρεία που μελετάμε.



Εικόνα 11: Pie charts: Attrition ανά Gender

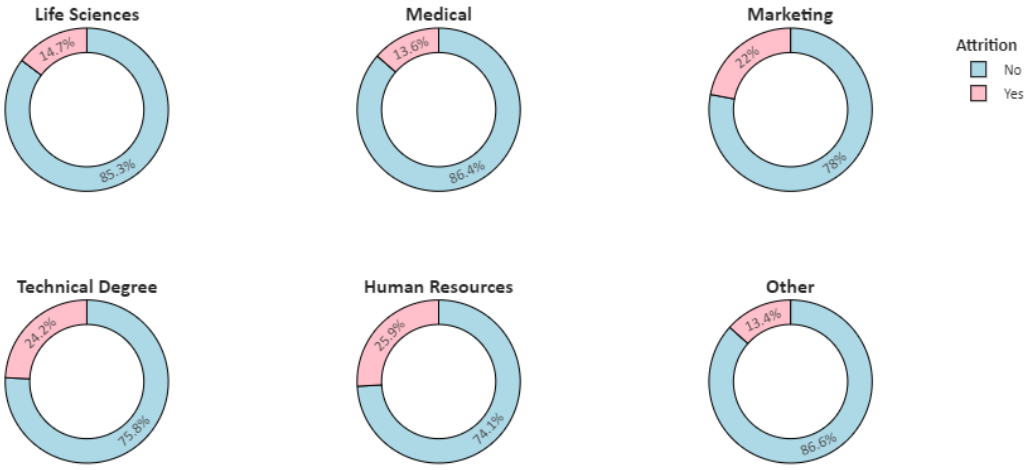
Στα παραπάνω pie charts μπορούμε να παρατηρήσουμε πως το ποσοστό αποχώρησης των εργαζομένων είναι 16%. Πιο συγκεκριμένα, οι άνδρες υπερέχουν σε ποσοστό αποχώρησης με 17% έναντι των γυναικών που έχουν ποσοστό αποχώρησης 14,8%.



Εικόνα 12: Pie charts: Attrition ανά BusinessTravel

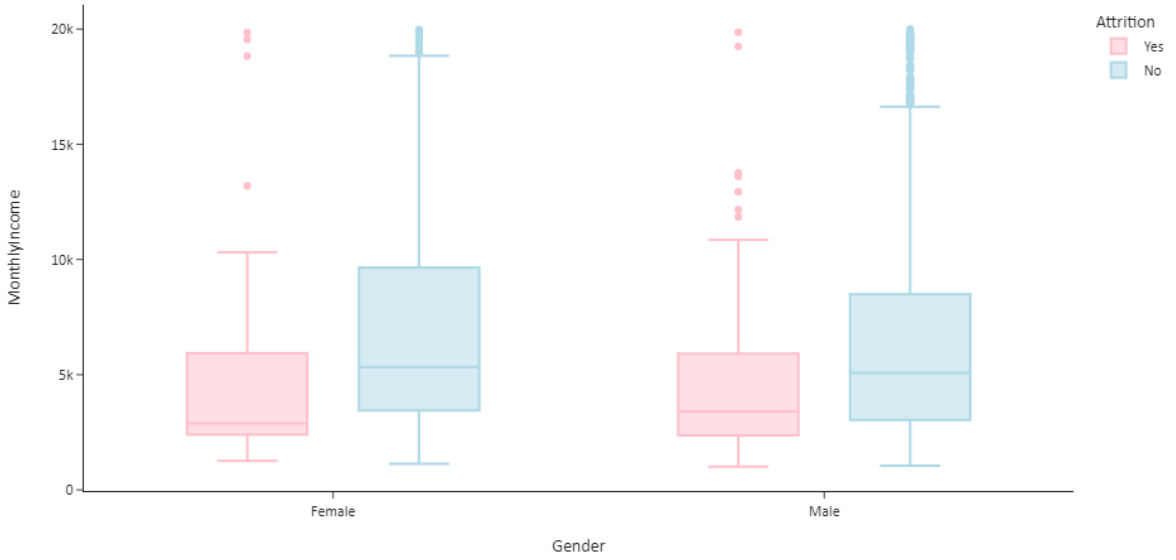
Στην Εικόνα 12, παρουσιάζονται pie charts για την αποχώρηση και την συχνότητα που χρειάζεται να ταξιδεύουν οι εργαζόμενοι. Το 24,9% των εργαζομένων που χρειάζεται να ταξιδεύουν συχνά για επαγγελματικούς σκοπούς, τείνει να αποχωρεί από την εταιρεία.

Ποσοστό πολύ μεγάλο συγκριτικά με τα αντίστοιχα ποσοστά για όσους δεν ταξιδεύουν καθόλου (8%) ή όσους ταξιδεύουν σπάνια (15%).



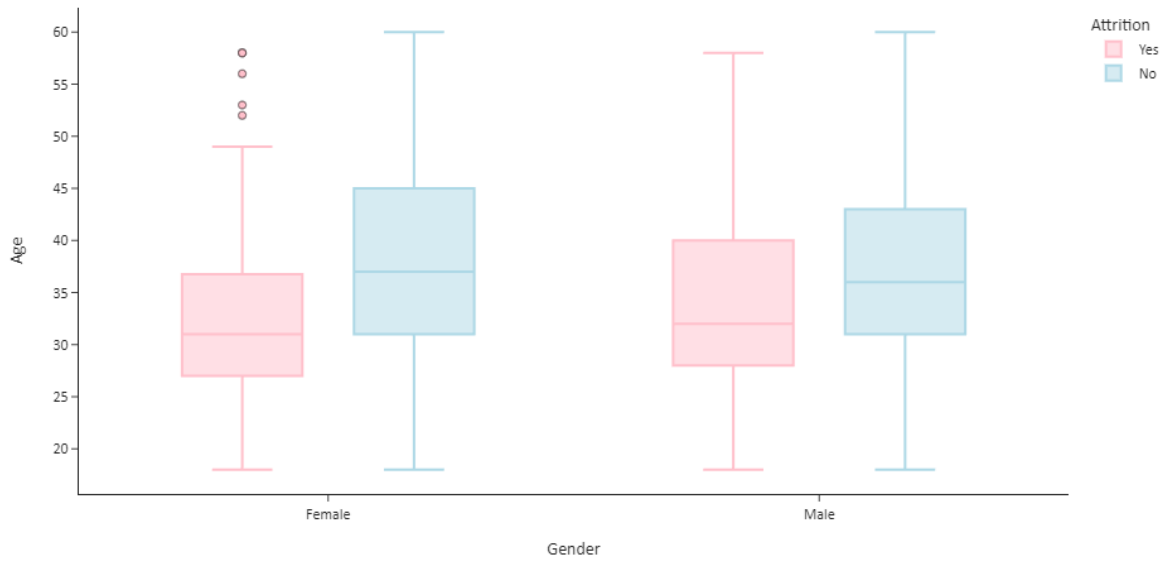
Εικόνα 13: Pie charts: Attrition ανά EducationField

Στην Εικόνα 13 μπορούμε να παρατηρήσουμε πως είναι πιο πιθανό να αποχωρήσει κάποιος εργαζόμενος ο οποίος είτε έχει σπουδάσει Marketing (22%), είτε Human Resources (25,9%), είτε είναι κάτοχος Technical Degree(24,2%).



Εικόνα 14: Boxplots: MonthlyIncome - Attrition ανά Gender

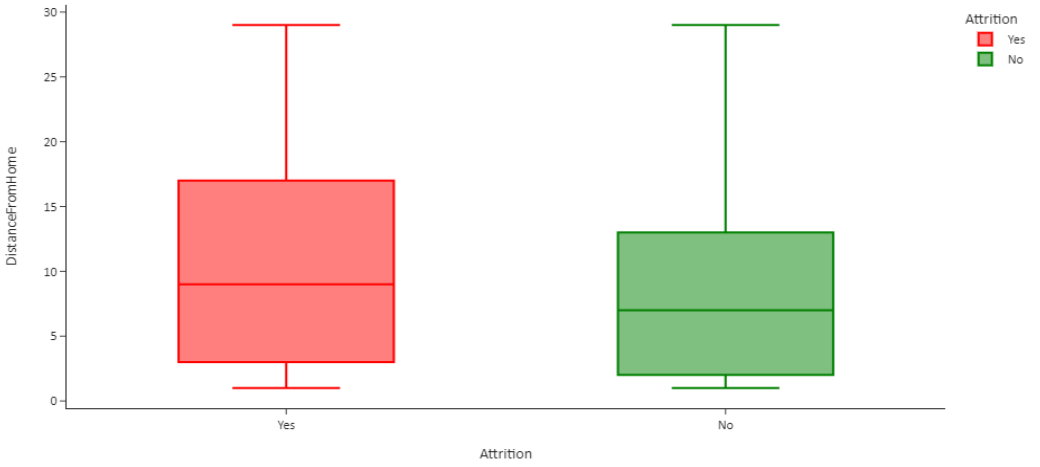
Στην Εικόνα 14 παρουσιάζονται κάποια πολύ ενδιαφέροντα θηκογράμματα (boxplots). Τα boxplots είναι γραφήματα στα οποία είναι μπορούμε να παρατηρήσουμε την μέγιστη και την ελάχιστη τιμή, το ενδοτεταρτημοριακό εύρος, τη διάμεσο, το πρώτο τεταρτημόριο, το τρίτο τεταρτημόριο καθώς και την ύπαρξη ακραίων τιμών (extreme values). Στην περίπτωσή μας, μπορεί εύκολα να παρατηρηθεί πως οι εργαζόμενοι με τους χαμηλότερους μισθούς τείνουν να αποχωρούν. Επίσης, είναι εμφανές πως οι γυναίκες που παραμένουν στην εταιρεία, έχουν μεγαλύτερους μισθούς σε σχέση με τους άνδρες που παραμένουν στην εταιρεία. Αντίστοιχα οι γυναίκες που παραμένουν στην εταιρεία έχουν οριακά μεγαλύτερη διάμεσο, μεγαλύτερο ενδοτεταρτημοριακό εύρος, ενώ έχουν και μεγαλύτερη μέγιστη τιμή μισθού. Αντίθετα, παρατηρείται πως οι γυναίκες που αποχώρησαν από την εταιρεία έχουν μικρότερη διάμεσο και μικρότερη μέγιστη τιμή σε σχέση με τους αντίστοιχους άνδρες.



Εικόνα 15: Boxplots: Age - Gender ανά Attrition

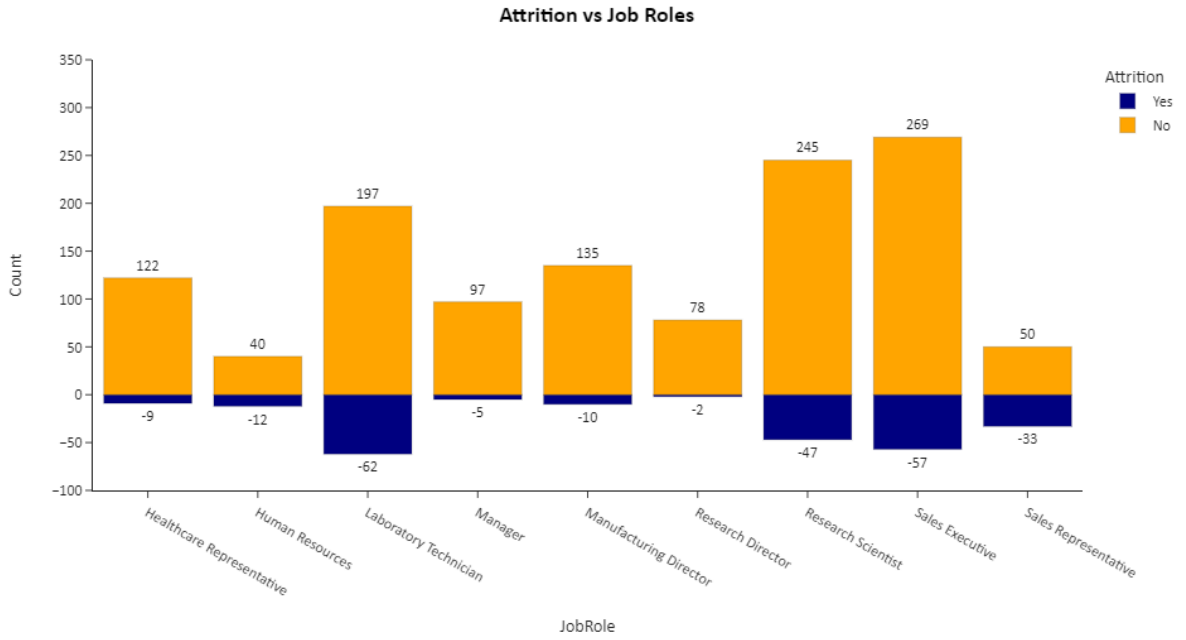
Στα παραπάνω boxplots, παρατηρούμε πως οι αποχωρήσαντες και των δύο φύλων είναι μικρότεροι ηλικιακά σε σχέση με αυτούς που παρέμειναν στην εταιρεία. Αυτό είναι εμφανές από τις διαμέσους, τις μέγιστες τιμές καθώς και το ενδοτεταρτημοριακό εύρος. Επίσης, παρατηρείται πως οι γυναίκες που αποχώρησαν περιέχουν και τέσσερα extreme values. Συνεπώς, φαίνεται πως υπάρχει μια τάση των νεότερων σε ηλικία να αλλάζουν εργοδότες πράγμα που φαντάζει λογικό καθώς αναζητούν περισσότερη εμπειρία και ευκαιρίες επαγγελματικής ανέλιξης.

Στη συνέχεια, θα μελετήσουμε τον ρόλο που παίζει η απόσταση του τόπου παροχής εργασίας σε σχέση με την οικεία του κάθε εργαζόμενου.



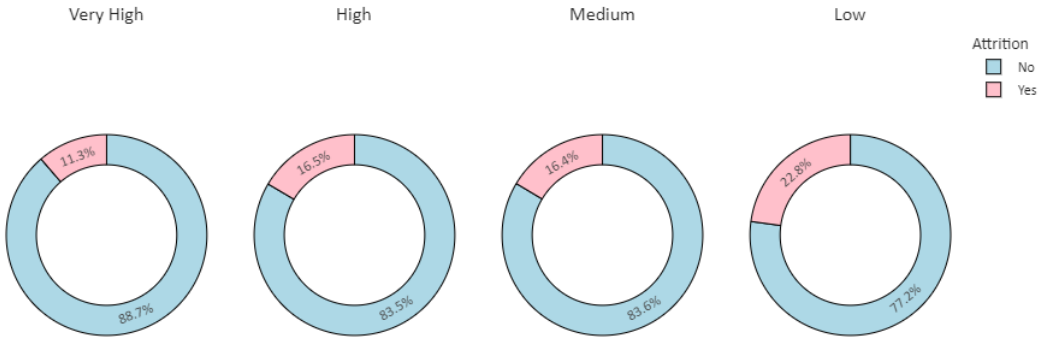
Εικόνα 16: Boxplots: DistanceFromHome - Attrition

Στα παραπάνω boxplots μπορούμε να παρατηρήσουμε αρχικά πως οι αποχωρήσαντες και αυτοί που έχουν παραμείνει, έχουν ίδια μέγιστη και ελάχιστη τιμή απόστασης της εργασίας από την οικία τους. Το πρώτο τεταρτημόριο όσων αποχώρησαν είναι 3 μίλια ενώ όσων παρέμειναν λαμβάνει την τιμή 1. Αναφορικά με το τρίτο τεταρτημόριο, οι αποχωρήσαντες έχουν την τιμή 17 ενώ όσοι παρέμειναν 13. Τέλος, και η διάμεσος όσων αποχώρησαν είναι οριακά μεγαλύτερη σε σχέση με όσους παρέμειναν 9 και 7 αντίστοιχα.



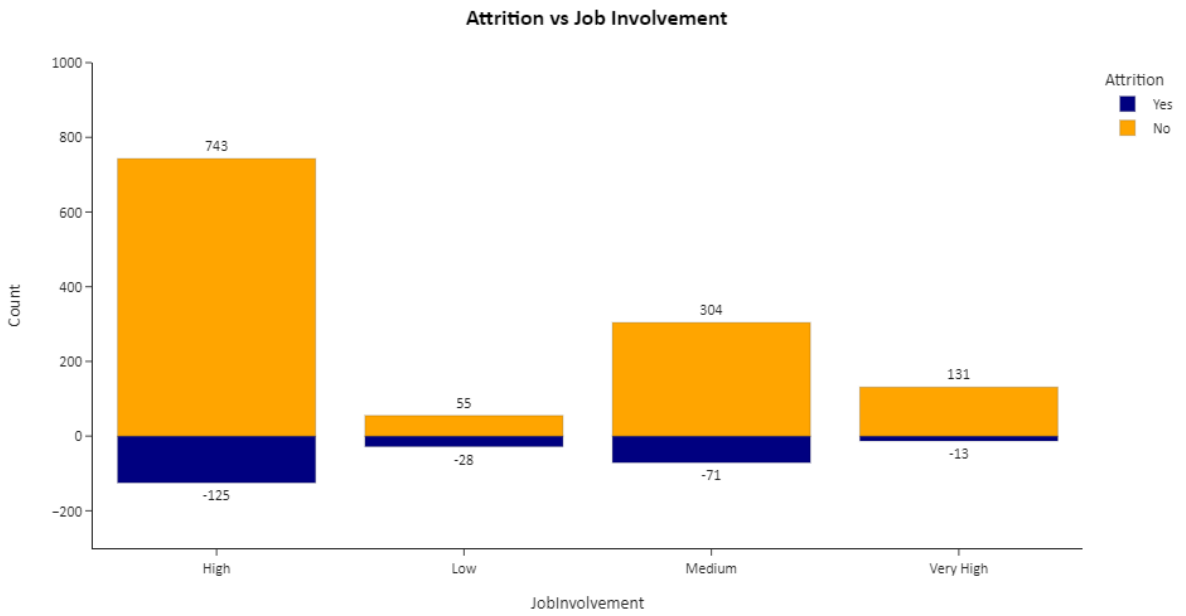
Εικόνα 17: Barcharts: Job Roles - Attrition

Από τα παραπάνω ραβδογράμματα (barcharts) ανάμεσα στις μεταβλητές JobRoles και Attrition, μπορούμε να παρατηρήσουμε πως οι θέσεις όπως Laboratory Technician, Sales Executive και Research Scientist καταλαμβάνουν τις πρώτες θέσεις σε αποχωρήσεις με 62, 57 και 47 αποχωρήσεις αντίστοιχα. Όλες οι παραπάνω θέσεις, εκτός από το μεγαλύτερο αριθμό αποχωρήσεων, είναι και οι θέσεις με τους περισσότερους απασχολούμενους. Ένα πολύ ενδιαφέρον στοιχείο μας δίνει η θέση του Sales Representative όπου αναλογικά με τις άλλες θέσεις παρουσιάζει το μεγαλύτερο ποσοστό αποχωρήσεων με 39,75%. Αντίθετα, φαίνεται πως οι Research Directors παρουσιάζουν το χαμηλότερο ποσοστό αποχώρησης το οποίο κυμαίνεται περίπου στο 2,5%.



Εικόνα 18: Pie Charts: JobSatisfaction - Attrition

Στην Εικόνα 18 παρατηρούμε τη σχέση της μεταβλητής Attrition με την μεταβλητή JobSatisfaction. Όσοι εργαζόμενοι έχουν χαμηλή ικανοποίηση από την εργασία τους, παρατηρούμε πως είναι πιο πιθανό να αποχωρήσουν σε σχέση με άλλους εργαζόμενους που βρίσκονται σε υψηλότερα επίπεδα εργασιακής ικανοποίησης. Παρόλα αυτά, τα ποσοστά δεν έχουν κάποια μεγάλη διαφορά που πιθανώς θα περιμέναμε. Για παράδειγμα, όσοι έχουν μέση και υψηλή ικανοποίηση φαίνεται πως έχουν το ίδιο ποσοστό παραμονής-αποχώρησης από την εταιρεία.
Στην συνέχεια θα παρουσιάσουμε ραβδογράμματα ανάμεσα στη μεταβλητή JobInvolvement και Attrition ώστε να έχουμε μια εικόνα εάν η επίδραση της εργασίας κάθε εργαζόμενου παίζει ρόλο στην απόφασή του να αποχωρήσει.



Εικόνα 19: Barcharts: Job Involvement - Attrition

Όπως μπορούμε να παρατηρήσουμε και παραπάνω, τις περισσότερες αποχωρήσεις έχουν όσοι έχουν μέση (medium) και υψηλή (high) επίδραση μέσω της εργασίας τους. Παρόλα αυτά, οι αριθμοί αυτοί είναι φυσιολογικοί καθώς οι κατηγορίες medium και high περιλαμβάνουν και τον μεγαλύτερο αριθμό εργαζομένων. Εάν τώρα λάβουμε υπόψιν τις αναλογίες όσων έχουν αποχωρήσει σε σχέση με αυτούς που έχουν παραμείνει, μπορούμε εύκολα να συμπεράνουμε πως ένας στους τρείς εργαζόμενους που η εργασία του έχει χαμηλή (low) επίδραση τείνει να αποχωρεί από την εταιρεία. Αντίθετα, από τους εργαζόμενους που έχουν πολύ υψηλή (very high) επίδραση μέσω της εργασίας τους, τείνει να αποχωρεί περίπου το 10% εξ αυτών.

Κεφάλαιο 5

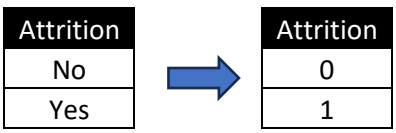
Εφαρμογή ταξινόμησης με χρήση της λογιστικής παλινδρόμησης στο σύνολο δεδομένων

5.1 Εισαγωγή

Στο κεφάλαιο αυτό θα εφαρμόσουμε ταξινόμηση (classification) στο σύνολο των δεδομένων με την χρήση της λογιστικής παλινδρόμησης ως ταξινομητή. Συγκεκριμένα θα αναλυθούν 2 μοντέλα. Ένα με τη χρήση όλων των μεταβλητών του συνόλου δεδομένων με σκοπό να προβλεφθεί η μεταβλητή Attrition καθώς και ένα μοντέλο με τη χρήση ενός υποσυνόλου των διαθέσιμων μεταβλητών του συνόλου δεδομένων. Φυσικά δεν θα πρέπει να παραληφθεί και η διαδικασία της προεπεξεργασίας δεδομένων που θα προηγηθεί της ταξινόμησης.

5.2 Προεπεξεργασία Δεδομένων

Ένα από τα κρισιμότερα βήματα προκειμένου να προχωρήσουμε στην ανάλυση των δεδομένων είναι η προεπεξεργασία των δεδομένων. Στην συγκεκριμένη ενότητα, τα κατηγορικά δεδομένα του συνόλου δεδομένων έχουν κωδικοποιηθεί (Label Encoding). Αυτή η διαδικασία περιλαμβάνει τον μετασχηματισμό των κατηγορικών μεταβλητών σε ακέραιες τιμές. Συγκεκριμένα, κάθε μοναδική εγγραφή κάθε μεταβλητής λαμβάνει τιμές ξεκινώντας από το 0 έως τον συνολικό αριθμό μοναδικών εγγραφών αφαιρώντας ένα. Για παράδειγμα, στο υπάρχον σύνοδο δεδομένων η μεταβλητή Attrition που δείχνει εάν ένας εργαζόμενος έχει παραμείνει ή έχει αποχωρήσει από την εταιρεία λαμβάνει την τιμή 0 εάν έχει παραμείνει στην εταιρεία ενώ εάν έχει αποχωρήσει ο εργαζόμενος λαμβάνει την τιμή 1.



Εικόνα 20: Κωδικοποίηση κατηγορικών μεταβλητών

5.3. Εφαρμογή ταξινόμησης με χρήση λογιστικής παλινδρόμησης

Για να εφαρμόσουμε την ταξινόμηση (classification) στο σύνολο των δεδομένων μας, αρχικά θα χρειαστεί να διασπάσουμε το σύνολο των δεδομένων σε δύο νέα υποσύνολα. Το ένα υποσύνολο θα ονομαστεί Χ και θα περιλαμβάνει τις τιμές των ανεξάρτητων μεταβλητών ενώ το δεύτερο υποσύνολο θα ονομαστεί Υ και θα περιλαμβάνει τις τιμές της εξαρτημένης μεταβλητής την οποία και θέλουμε να προβλέψουμε. Στη συνέχεια, θα χρειαστεί να εφαρμόσουμε τον διαχωρισμό του συνόλου των δεδομένων.

Ο διαχωρισμός του συνόλου των δεδομένων εφαρμόζεται κυρίως σε δύο σύνολα. Το ένα ονομάζεται σύνολο εκπαίδευσης (training set) το οποίο τροφοδοτεί το μοντέλο προκειμένου να μπορέσει να εκπαιδευτεί βρίσκοντας τα μοτίβα των δεδομένων. Το δεύτερο ονομάζεται σύνολο δοκιμής (test set) το οποίο χρησιμεύει στην αξιολόγηση της απόδοσης του μοντέλου και κατά πόσο είναι ικανό να ταξινομεί νέα δεδομένα. Ο σκοπός της παραπάνω διαδικασίας διαχωρισμού είναι η αποφυγή της υπερπροσαρμογής (overfitting) στα υπάρχοντα δεδομένα καθώς ο αλγόριθμος θα δυσκολεύεται να επεξηγεί νέα δεδομένα.

Υπάρχουν διάφορες αναλογίες ως προς τον διαχωρισμό των συνόλων εκπαίδευσης και των συνόλων δοκιμής. Όλες όμως έχουν το μεγαλύτερο εύρος των δεδομένων για σκοπούς εκπαίδευσης. Οι δύο πιο συνήθεις αναλογίες είναι 70% / 30% και 80% / 20% του συνόλου των δεδομένων.

Πριν από την παρουσίαση των αποτελεσμάτων της ταξινόμησης θα χρειαστεί να αναφερθούμε πρώτα στις τεχνικές αξιολόγησης των μοντέλων ταξινόμησης και συγκεκριμένα στον πίνακα σύγχυσης.

Ο πίνακας σύγχυσης (Confusion Matrix) είναι ένας πίνακας που χρησιμοποιείται προκειμένου να αξιολογήσουμε την απόδοση ενός μοντέλου ταξινόμησης. Ένας πίνακας σύγχυσης περιλαμβάνει τις παρακάτω πληροφορίες:

- **Ορθά θετικό (TP):** Είναι ο αριθμός των περιπτώσεων που προέβλεψε σωστά ως θετικές το μοντέλο.

- **Ορθά αρνητικό (TN):** Είναι ο αριθμός των περιπτώσεων που προέβλεψε σωστά ως αρνητικές το μοντέλο.

- **Ψευδώς θετικό (FP):** Είναι ο αριθμός των περιπτώσεων που εσφαλμένα το μοντέλο προέβλεψε ως θετικές παρόλο που στην πραγματικότητα ήταν αρνητικές.

- **Ψευδώς αρνητικό (FN):** Είναι ο αριθμός των περιπτώσεων που εσφαλμένα το μοντέλο προέβλεψε ως αρνητικές παρόλο που στην πραγματικότητα ήταν θετικές.

Όπως μπορεί να παρατηρηθεί και στον παρακάτω πίνακα (Πίνακας 5), τα στοιχεία της διαγωνίου του πίνακα μας δίνουν τις σωστές προβλέψεις ενώ τα στοιχεία που δεν βρίσκονται στη διαγώνιο, μας δίνουν τις λάθος προβλέψεις.

	Πρόβλεψη	
Πραγματική τιμή	**Θετικό**	**Αρνητικό**
Θετικό	Ορθά θετικό	Ψευδώς αρνητικό
Αρνητικό	Ψευδώς θετικό	Ορθά αρνητικό

Πίνακας 5: Μορφή πίνακα σύγχυσης (Confusion Matrix)

Με βάση τις τιμές στον πίνακα σύγχυσης, μπορούν να υπολογιστούν διάφορες μετρικές για την αξιολόγηση της απόδοσης του μοντέλου, όπως

- **Ορθότητα (Accuracy):** Αυτή μετρά τη συνολική ορθότητα του μοντέλου και υπολογίζεται ως:

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

- **Ακρίβεια (Precision):** Αυτό μετρά το ποσοστό των αληθώς θετικών μεταξύ όλων των θετικών προβλέψεων και υπολογίζεται ως:

$$\frac{TP}{TP + FP}$$

- **Ανάκληση (Recall):** Αυτό μετρά την αναλογία των αληθώς θετικών μεταξύ όλων των πραγματικών θετικών αποτελεσμάτων και υπολογίζεται ως:

$$\frac{TP}{TP + FN}$$

- **Βαθμολογία F1 (F1-Score):** Πρόκειται για τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης και υπολογίζεται ως:

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

(Ελευθεράκου Ο. , 2023)

5.3.1 Εφαρμογή ταξινόμησης με χρήση λογιστικής παλινδρόμησης για το πλήρες μοντέλο

Εφαρμόζοντας ταξινόμηση με τη χρήση της λογιστικής παλινδρόμησης για το σύνολο των δεδομένων και με διαχωρισμό 70% / 30% προέκυψαν τα εξής αποτελέσματα:

Testing Results:					
Accuracy Score:	0.8617				
Classification Report:					
	0	1	accuracy	macro avg	weighted avg
precision	0.874384	0.714286	0.861678	0.794335	0.846794
recall	0.972603	0.328947	0.861678	0.650775	0.861678
f1-score	0.920882	0.450450	0.861678	0.685666	0.839810

Πίνακας 6: Αποτελέσματα ταξινόμησης (Classification Report)

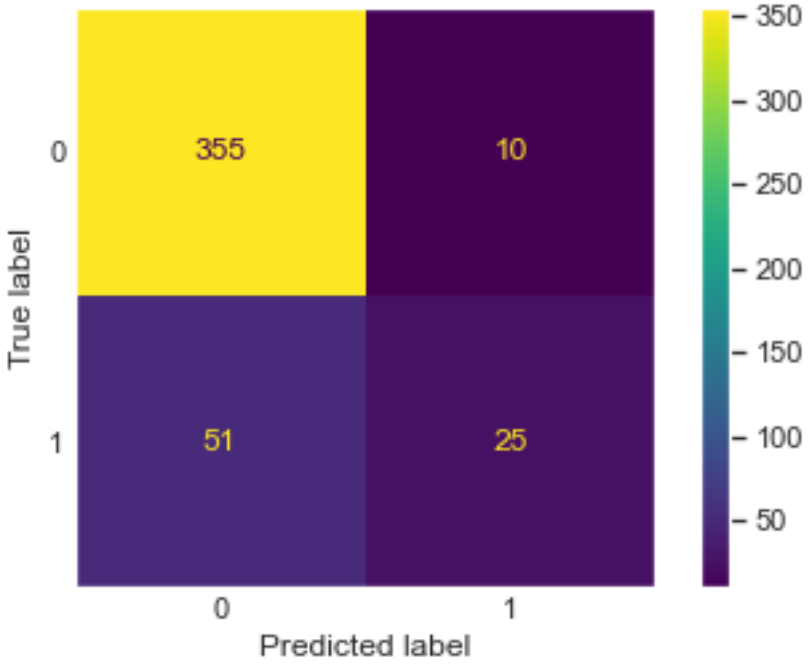
Παρατηρούμε ότι η ορθότητα (Accuracy) είναι 86,17%. Δηλαδή το μοντέλο μας ταξινομεί σωστά το 86,17% όλων των εγγραφών.

Η ακρίβεια (Precision) είναι 87,43% για την κλάση 0 και 71,42% για την κλάση 1. Δηλαδή το μοντέλο προβλέπει σωστά στο 87,43% των περιπτώσεων που ανήκουν στην κλάση 0 ανάμεσα σε όσα έχει προβλέψει στην κλάση 0. Αντίστοιχα, προβλέπει σωστά το 71,42% όσων ανήκουν στην κλάση 1 απ' όσους έχουν προβλεφθεί στην κλάση 1.

Αναφορικά με την ανάκληση (Recall) που μας δείχνει ότι το μοντέλο μας αναγνωρίζει σωστά το 97,26% όλων των εγγραφών που ανήκουν πραγματικά στην κλάση 0 ενώ αναγνωρίζει σωστά μόνο το 32,89% απ' όλες τις εγγραφές που ανήκουν πραγματικά στην κλάση 1.

Το F1-Score για την κλάση 0 είναι 92,08% ενώ για την κλάση 1 είναι 45,04%. Αυτό σημαίνει ότι για την πρόβλεψη της κλάσης 0, δηλαδή των εργαζομένων που έχουν παραμείνει στην εταιρεία, ο αλγόριθμος έχει καλή απόδοση στην ταξινόμηση αυτών των εγγραφών.

Στη συνέχεια, παρουσιάζεται ο πίνακας σύγχυσης του μοντέλου:

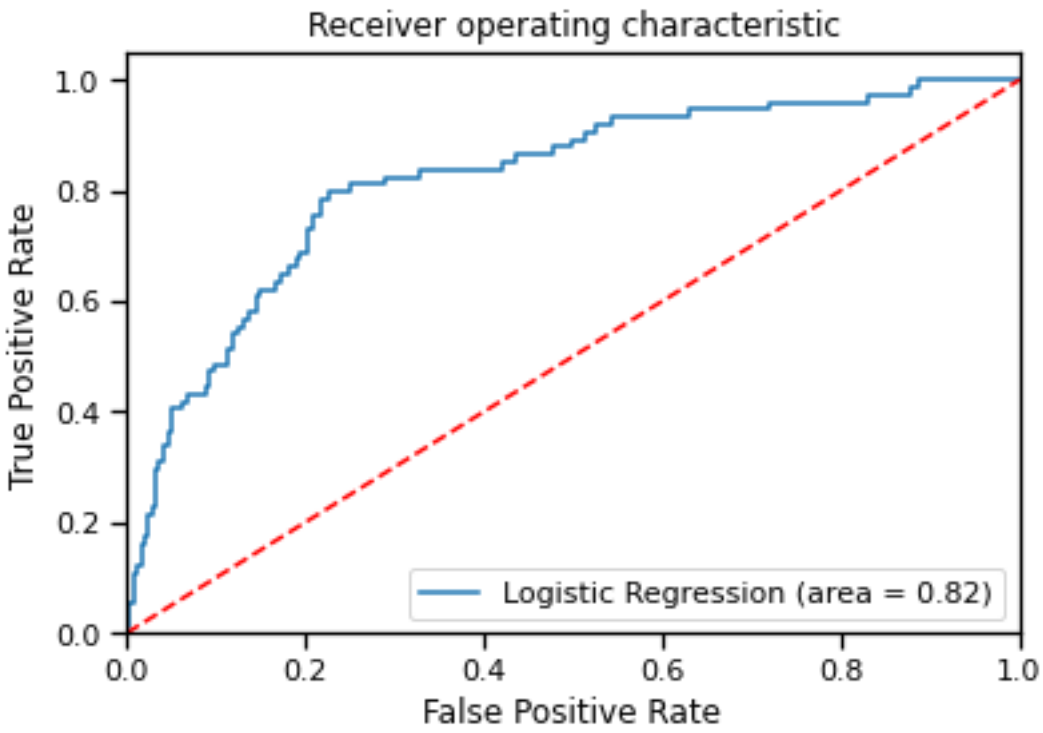


Εικόνα 21: Πίνακας σύγχυσης ταξινόμησης με χρήση λογιστικής παλινδρόμησης (πλήρες μοντέλο)

Από τον παραπάνω πίνακα σύγχυσης λαμβάνουμε τις εξής πληροφορίες:

- 355 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 0.
- 25 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 1.
- 10 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 1 ενώ στην πραγματικότητα ανήκουν στην κλάση 0.
- 51 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 0 ενώ στην πραγματικότητα ανήκουν στην κλάση 1.

Παρακάτω, παρουσιάζεται η καμπύλη ROC του μοντέλου:



Εικόνα 22: Καμπύλη ROC πλήρους μοντέλου με χρήση λογιστικής παλινδρόμησης

Όπως φαίνεται και στην παραπάνω εικόνα (Εικόνα 22), η τιμή AUC είναι 0,82. Γενικά, το AUC (Area Under Curve) χρησιμοποιείται ως ένα μέτρο το οποίο μας βοηθάει στην συνολική αξιολόγηση του μοντέλου και λαμβάνει τιμές από 0 έως 1. Σε κάθε έλεγχο θέλουμε η τιμή AUC να είναι μεγαλύτερη του 0,5 καθώς μόνο τότε οι σωστές θετικές εκτιμήσεις είναι περισσότερες από τις λανθασμένες θετικές εκτιμήσεις. Στην περίπτωσή μας η τιμή του AUC θα μπορούσε να χαρακτηριστεί αρκετά καλή καθώς είναι άνω του 0,8.

Τέλος, παρουσιάζεται και ο πίνακας ANOVA με τις εκτιμήσεις των μεταβλητών του πλήρους μοντέλου.

Generalized Linear Model Regression Results

Dep. Variable:	Attrition	No. Observations:	1470
Model:	GLM	Df Residuals:	1442
Model Family:	Binomial	Df Model:	27
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-488.75
Date:	Sat, 01 Jun 2024	Deviance:	977.51
Time:	16:46:57	Pearson chi2:	1.67e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.1746	0.892	-3.558	0.000	-4.923	-1.426
Age	-0.0295	0.013	-2.351	0.019	-0.054	-0.005
BusinessTravel	0.0330	0.126	0.262	0.794	-0.214	0.280
Department	1.2220	0.255	4.788	0.000	0.722	1.722
DistanceFromHome	0.0404	0.010	4.002	0.000	0.021	0.060
Education	-0.0235	0.052	-0.452	0.651	-0.125	0.078
EducationField	0.0250	0.062	0.402	0.688	-0.097	0.147
EnvironmentSatisfaction	-0.1202	0.068	-1.775	0.076	-0.253	0.013
Gender	0.2706	0.170	1.588	0.112	-0.063	0.605
JobInvolvement	0.0274	0.075	0.366	0.714	-0.119	0.174
JobLevel	0.4453	0.079	5.651	0.000	0.291	0.600
JobRole	-0.1222	0.048	-2.543	0.011	-0.216	-0.028
JobSatisfaction	-0.2217	0.068	-3.241	0.001	-0.356	-0.088
MaritalStatus	0.6872	0.130	5.276	0.000	0.432	0.943
MonthlyIncome	-5.213e-05	3.5e-05	-1.489	0.136	-0.000	1.65e-05
NumCompaniesWorked	0.1652	0.035	4.709	0.000	0.096	0.234
OverTime	1.6717	0.173	9.690	0.000	1.334	2.010
PercentSalaryHike	-0.0419	0.036	-1.161	0.246	-0.113	0.029
PerformanceRating	0.3868	0.368	1.051	0.293	-0.335	1.108
RelationshipSatisfaction	-0.0827	0.068	-1.214	0.225	-0.216	0.051
StockOptionLevel	-0.0451	0.130	-0.347	0.729	-0.300	0.210
TotalWorkingYears	-0.0268	0.026	-1.024	0.306	-0.078	0.024
TrainingTimesLastYear	-0.1451	0.067	-2.177	0.029	-0.276	-0.014
WorkLifeBalance	0.1311	0.073	1.791	0.073	-0.012	0.275
YearsAtCompany	0.0695	0.034	2.035	0.042	0.003	0.136
YearsInCurrentRole	-0.1193	0.042	-2.835	0.005	-0.202	-0.037
YearsSinceLastPromotion	0.1370	0.038	3.583	0.000	0.062	0.212
YearsWithCurrManager	-0.1041	0.043	-2.436	0.015	-0.188	-0.020

Πίνακας 7: Πίνακας ANOVA reduced model

Παρατηρούμε ότι σε επίπεδο σημαντικότητας 5%, οι μεταβλητές Age, Department, DistanceFromHome, Joblevel, JobRole, JobSatisfaction, MaritalStatus, NumCompaniesWorked, Overtime, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager είναι στατιστικά σημαντικές.
Αντιθέτως, οι μεταβλητές WorkLifeBalance, TotalWorkingYears, StockOptionsLevel, RelationshipSatisfaction, PerformanceRating, PercentSalaryHike, MonthlyIncome, Gender, JobInvolvement, Education, EducationField, EnvironmentSatisfaction και BusinessTravel φαίνεται πως δεν είναι στατιστικά σημαντικές.

5.3.2 Εφαρμογή ταξινόμησης με χρήση λογιστικής παλινδρόμησης για υποσύνολο του πλήρους μοντέλου

Για την εφαρμογή της ταξινόμησης με τη χρήση υποσυνόλου του πλήρους μοντέλου έχει γίνει επιλογή μεταβλητών με τη χρήση της εντολής SequentialFeatureSelection με οπίσθια τροφοδότηση και με χρήση score το αρνητικό μέσο τετραγωνικό σφάλμα.

Από την παραπάνω διαδικασία προέκυψαν οι εξής μεταβλητές:

- Age
- DistanceFromHome
- JobLevel
- MaritalStatus
- MonthlyIncome
- OverTime
- TotalWorkingYears
- YearsAtCompany
- YearsInCurrentRole
- YearsWithCurrManager

Εφαρμόζοντας ταξινόμηση με τη χρήση της λογιστικής παλινδρόμησης για τα παραπάνω δεδομένα και με διαχωρισμό 70% / 30% προέκυψαν τα εξής αποτελέσματα:

Testing Results					
Accuracy Score	0.8481				
Classification Report:					
	0	1	accuracy	macro avg	weighted avg
precision	0.862319	0.629630	0.848073	0.745974	0.823273
recall	0.972752	0.229730	0.848073	0.601241	0.848073
f1-score	0.914213	0.336634	0.848073	0.625423	0.817295

Πίνακας 8: Αποτελέσματα ταξινόμησης – reduced model(Classification Report)

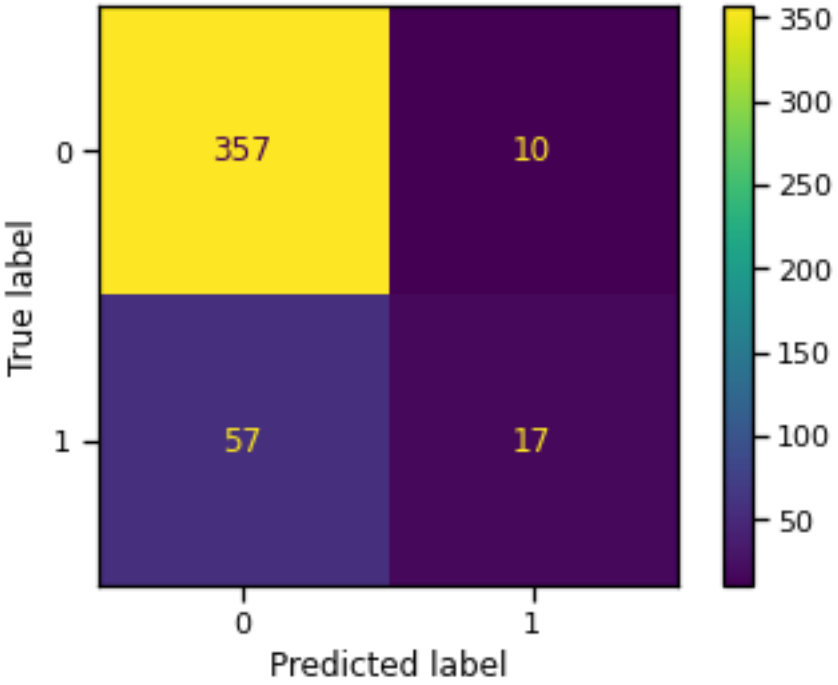
Αρχικά, παρατηρούμε ότι η ορθότητα (Accuracy) είναι 84,81%. Δηλαδή το μοντέλο μας ταξινομεί σωστά το 84,81% όλων των εγγραφών.

Η ακρίβεια (Precision) είναι 86,23% για την κλάση 0 και 62,96% για την κλάση 1. Δηλαδή το μοντέλο προβλέπει σωστά στο 86,23% των περιπτώσεων που ανήκουν στην κλάση 0 ανάμεσα σε όσα έχει προβλέψει στην κλάση 0. Αντίστοιχα, προβλέπει σωστά το 62,96% όσων ανήκουν στην κλάση 1 απ' όσους έχουν προβλεφθεί στην κλάση 1.

Αναφορικά με την ανάκληση (Recall) που μας δείχνει ότι το μοντέλο μας αναγνωρίζει σωστά το 97,27% όλων των εγγραφών που ανήκουν πραγματικά στην κλάση 0 ενώ αναγνωρίζει σωστά μόνο το 22,97% απ' όλες τις εγγραφές που ανήκουν πραγματικά στην κλάση 1.

Το F1-Score για την κλάση 0 είναι 91,42% ενώ για την κλάση 1 είναι 33,66%. Αυτό σημαίνει ότι για την πρόβλεψη της κλάσης 0, δηλαδή των εργαζομένων που έχουν παραμείνει στην εταιρεία, ο αλγόριθμος έχει καλή απόδοση στην ταξινόμηση αυτών των εγγραφών.

Στη συνέχεια παρουσιάζεται ο πίνακας σύγχυσης:



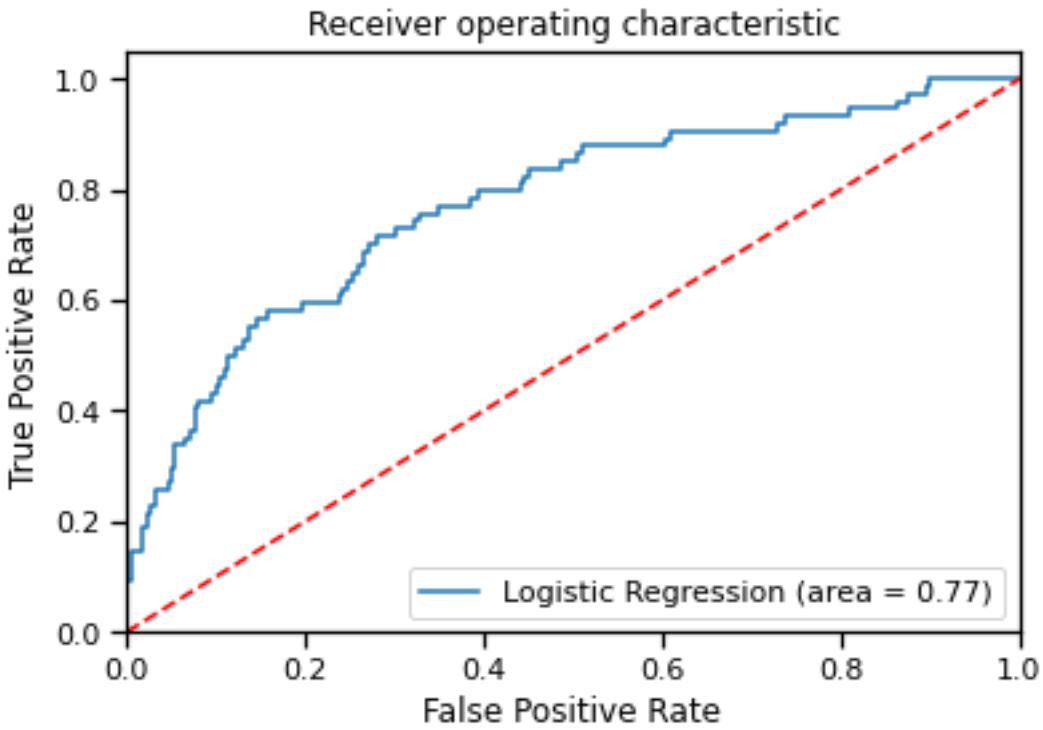
Εικόνα 23: Πίνακας σύγχυσης ταξινόμησης με χρήση λογιστικής παλινδρόμησης (reduced model)

Από τον παραπάνω πίνακα σύγχυσης λαμβάνουμε τις εξής πληροφορίες:

- 357 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 0.
- 17 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 1.

- 10 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 1 ενώ στην πραγματικότητα ανήκουν στην κλάση 0.
- 57 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 0 ενώ στην πραγματικότητα ανήκουν στην κλάση 1.

Παρακάτω παρουσιάζεται η καμπύλη ROC του μοντέλου:



Εικόνα 24: Καμπύλη ROC reduced μοντέλου με χρήση λογιστικής παλινδρόμησης

Σε αυτή τη περίπτωση το AUC είναι 0,77. Είναι ικανοποιητικό το score καθώς αν κρίνουμε ότι στο full model με όλες τις ερμηνευτικές μεταβλητές είχαμε AUC = 0,82 τότε στο μοντέλο με μόνο 10 ερμηνευτικές είχαμε σχετικά μικρή απώλεια με 0,77.

Κεφάλαιο 6

Εφαρμογή ταξινόμησης με χρήση αλγορίθμων μηχανικής μάθησης στο σύνολο δεδομένων

6.1 Εισαγωγή

Στο κεφάλαιο αυτό θα εφαρμόσουμε ταξινόμηση (classification) στο σύνολο των δεδομένων με την χρήση αλγορίθμων της μηχανικής μάθησης. Συγκεκριμένα, θα χρησιμοποιηθεί ο αλγόριθμος Decision Tree (Δέντρα Απόφασης), ο αλγόριθμος Random Forest (Τυχαίου Δάσους) και τέλος ο Support Vector Machine (Μηχανές Διανυσμάτων Υποστήριξης). Επιπλέον, θα παρουσιαστούν οι παρακάτω μετρικές για την αξιολόγηση της απόδοσης των μοντέλων όπως ορίστηκαν και στο προηγούμενο κεφάλαιο:

- Accuracy (Ορθότητα)
- Precision (Ακρίβεια)
- Recall (Απόκλιση)
- F1 Score (Βαθμολογία F1)

6.2 Αλγόριθμος Δέντρου Απόφασης (Decision Tree Algorithm)

Εφαρμόζουμε ταξινόμηση με τη χρήση του αλγορίθμου Decision Tree για το σύνολο των δεδομένων και με διαχωρισμό 70% / 30% και κριτήριο διαχωρισμού το Gini impurity. Προέκυψαν τα εξής αποτελέσματα:

Testing Results					
Accuracy Score:	0.8095				
Classification Report:					
	0	1	accuracy	macro avg	weighted avg
precision	0.879357	0.426471	0.809524	0.652914	0.803362
recall	0.893733	0.391892	0.809524	0.642812	0.809524
f1-score	0.886486	0.408451	0.809524	0.647469	0.806272

Πίνακας 9: Αποτελέσματα ταξινόμησης – Decision Tree (Classification Report)

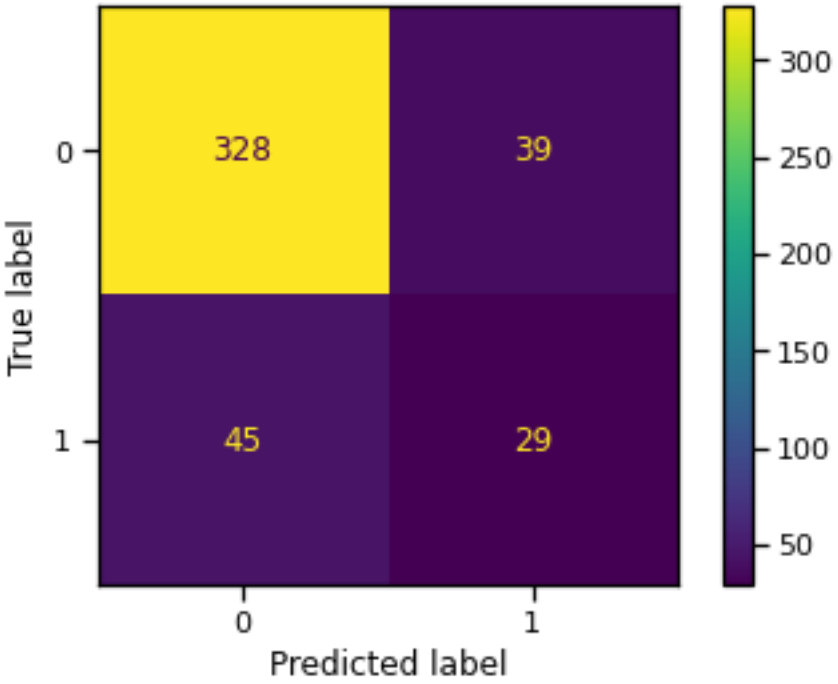
Αρχικά, παρατηρούμε ότι η ορθότητα (Accuracy) είναι 80,95%. Δηλαδή το μοντέλο μας ταξινομεί σωστά το 80,95% όλων των εγγραφών.

Η ακρίβεια (Precision) είναι 87,93% για την κλάση 0 και 42,64% για την κλάση 1. Δηλαδή το μοντέλο προβλέπει σωστά στο 87,93% των περιπτώσεων που ανήκουν στην κλάση 0 ανάμεσα σε όσα έχει προβλέψει στην κλάση 0. Αντίστοιχα, προβλέπει σωστά το 42,64% όσων ανήκουν στην κλάση 1 απ' όσους έχουν προβλεφθεί στην κλάση 1.

Αναφορικά με την ανάκληση (Recall) που μας δείχνει ότι το μοντέλο μας αναγνωρίζει σωστά το 89,37% όλων των εγγραφών που ανήκουν πραγματικά στην κλάση 0 ενώ αναγνωρίζει σωστά μόνο το 39,18% απ' όλες τις εγγραφές που ανήκουν πραγματικά στην κλάση 1.

Το F1-Score για την κλάση 0 είναι 88,64% ενώ για την κλάση 1 είναι 40,84%. Αυτό σημαίνει ότι για την πρόβλεψη της κλάσης 0, δηλαδή των εργαζομένων που έχουν παραμείνει στην εταιρεία, ο αλγόριθμος έχει καλή απόδοση στην ταξινόμηση αυτών των εγγραφών.

Στη συνέχεια παρουσιάζεται ο πίνακας σύγχυσης (Confusion Matrix):



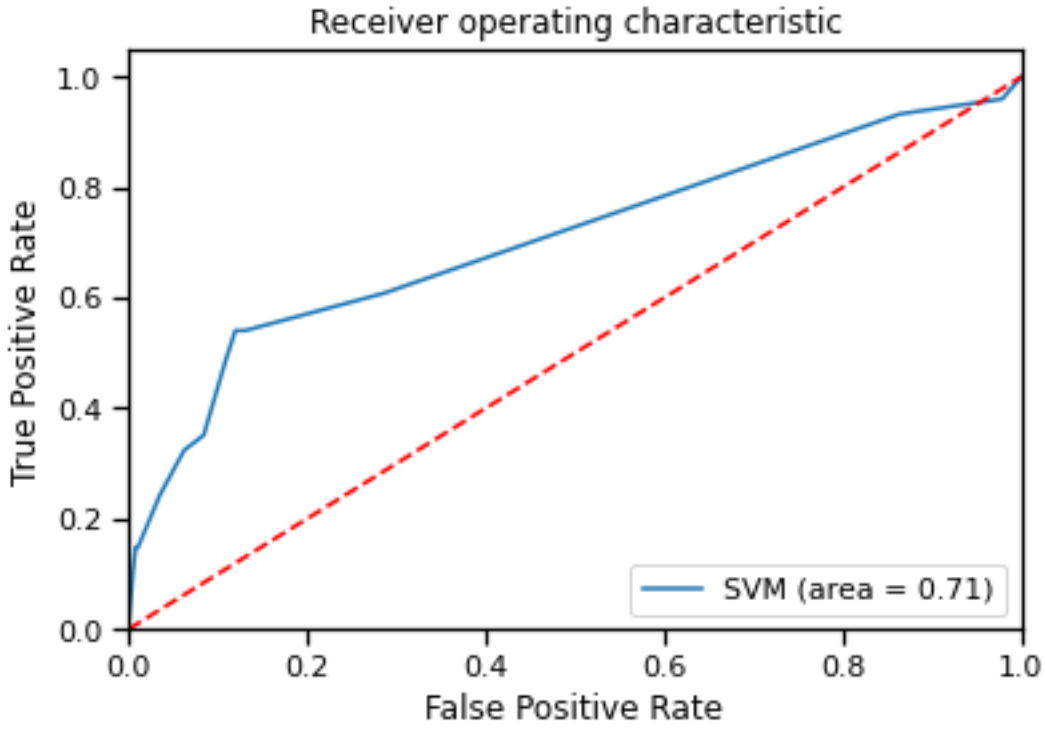
Εικόνα 25: Πίνακας σύγχυσης ταξινόμησης με χρήση δέντρων απόφασης

Από τον παραπάνω πίνακα σύγχυσης λαμβάνουμε τις εξής πληροφορίες:

- 328 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 0.
- 29 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 1.
- 39 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 1 ενώ στην πραγματικότητα ανήκουν στην κλάση 0.

- 45 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 0 ενώ στην πραγματικότητα ανήκουν στην κλάση 1.

Στην παρακάτω εικόνα, παρουσιάζεται η καμπύλη ROC καθώς και η τιμή AUC. Όπως μπορούμε να παρατηρήσουμε, η τιμή AUC είναι 0,71.



Εικόνα 26: Καμπύλη ROC μοντέλου με χρήση δέντρων απόφασης

6.3 Αλγόριθμος Τυχαίου Δάσους (Random Forest Algorithm)

Στη συγκεκριμένη ενότητα, θα εφαρμόσουμε ταξινόμηση με τη χρήση του αλγορίθμου Random Forest για το σύνολο των δεδομένων. Ο διαχωρισμός που θα κάνουμε στα δεδομένα και σε αυτή τη περίπτωση θα είναι 70% / 30%. Τα παρακάτω αποτελέσματα προέκυψαν από την ανάλυση:

Testing Results:					
Accuracy Score:	0.8594				
Classification Report:					
	0	1	accuracy	macro avg	weighted avg
precision	0.871046	0.700000	0.85941	0.785523	0.842345
recall	0.975477	0.283784	0.85941	0.629630	0.859410
f1-score	0.920308	0.403846	0.85941	0.662077	0.833646

Πίνακας 10: Αποτελέσματα ταξινόμησης – Random Forest (Classification Report)

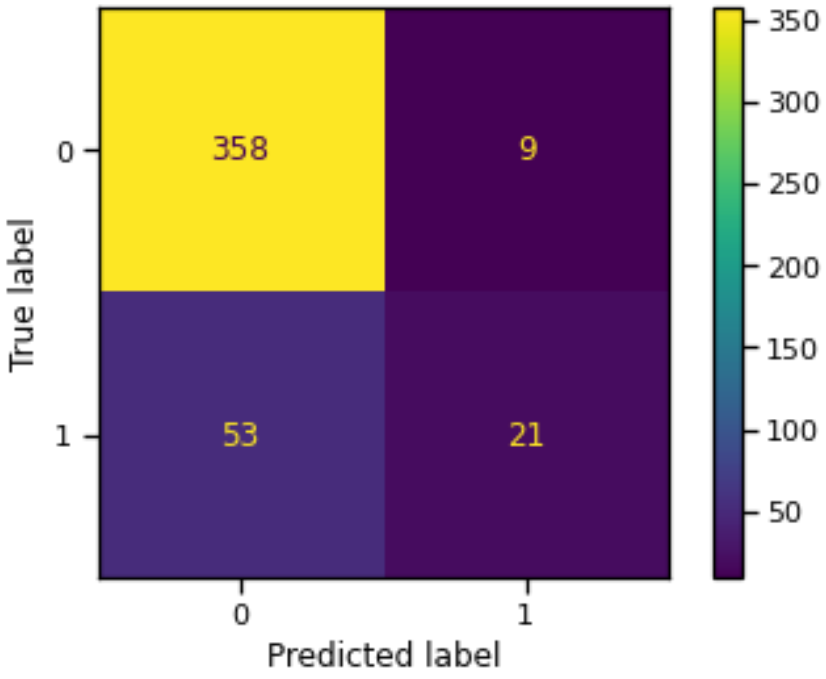
Από τον παραπάνω πίνακα παρατηρούμε ότι, η ορθότητα (Accuracy) είναι 85,94%. Δηλαδή το μοντέλο μας ταξινομεί σωστά το 85,94% όλων των εγγραφών.

Η ακρίβεια (Precision) είναι 87,10% για την κλάση 0 και 70,00% για την κλάση 1. Δηλαδή το μοντέλο προβλέπει σωστά στο 87,10% των περιπτώσεων που ανήκουν στην κλάση 0 ανάμεσα σε όσα έχει προβλέψει στην κλάση 0. Αντίστοιχα, προβλέπει σωστά το 70,00% όσων ανήκουν στην κλάση 1 απ' όσους έχουν προβλεφθεί στην κλάση 1.

Αναφορικά με την ανάκληση (Recall) που μας δείχνει ότι το μοντέλο μας αναγνωρίζει σωστά το 97,54% όλων των εγγραφών που ανήκουν πραγματικά στην κλάση 0 ενώ αναγνωρίζει σωστά μόνο το 28,37% απ' όλες τις εγγραφές που ανήκουν πραγματικά στην κλάση 1.

Το F1-Score για την κλάση 0 είναι 92,03% ενώ για την κλάση 1 είναι 40,38%. Αυτό σημαίνει ότι για την πρόβλεψη της κλάσης 0, δηλαδή των εργαζομένων που έχουν παραμείνει στην εταιρεία, ο αλγόριθμος έχει πολύ καλή απόδοση στην ταξινόμηση αυτών των εγγραφών.

Στη συνέχεια παρουσιάζεται ο πίνακας σύγχυσης (Confusion Matrix):



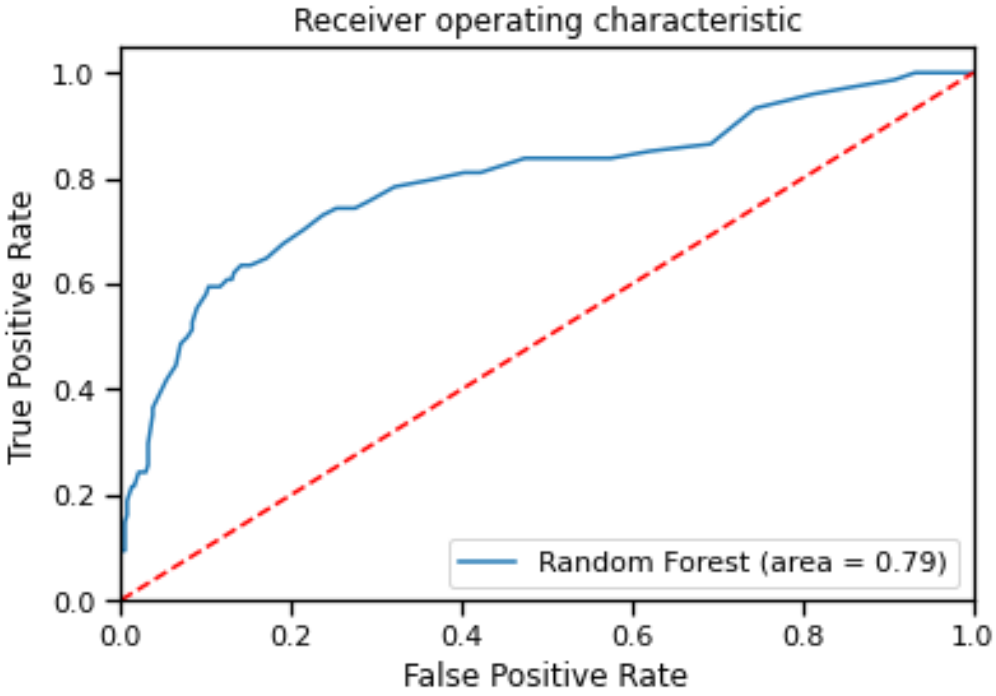
Εικόνα 27: Πίνακας σύγχυσης ταξινόμησης με χρήση αλγορίθμου τυχαίου δάσους

Από τον πίνακα σύγχυσης της εικόνας 27 λαμβάνουμε τις εξής πληροφορίες:

- 358 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 0.

- 21 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 1.
- 9 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 1 ενώ στην πραγματικότητα ανήκουν στην κλάση 0.
- 53 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 0 ενώ στην πραγματικότητα ανήκουν στην κλάση 1.

Στη συνέχεια παρουσιάζεται η καμπύλη ROC του μοντέλου:



Εικόνα 28: Καμπύλη ROC μοντέλου με χρήση αλγόριθμου τυχαίου δάσους

Όπως μπορούμε να παρατηρήσουμε από την παραπάνω εικόνα, το AUC είναι και σε αυτή τη περίπτωση ικανοποιητικό και πιο συγκεκριμένα είναι 0,79.

6.4 Αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines Algorithm)

Σε αυτή την ενότητα, θα εφαρμόζουμε ταξινόμηση με τον αλγόριθμο Support Vector Machines για το σύνολο των δεδομένων και με διαχωρισμό 70% / 30% και τη χρήση του γραμμικού πυρήνα (Linear Kernel). Από την ανάλυση, προέκυψαν τα εξής αποτελέσματα:

Testing Results	
Accuracy Score	0.8481
Classification Report:	

	0	1	accuracy	macro avg	weighted avg
precision	0.875000	0.585366	0.848073	0.730183	0.826399
recall	0.953678	0.324324	0.848073	0.639001	0.848073
f1-score	0.912647	0.417391	0.848073	0.665019	0.829543

Πίνακας 11: Αποτελέσματα ταξινόμησης – SVM (Classification Report)

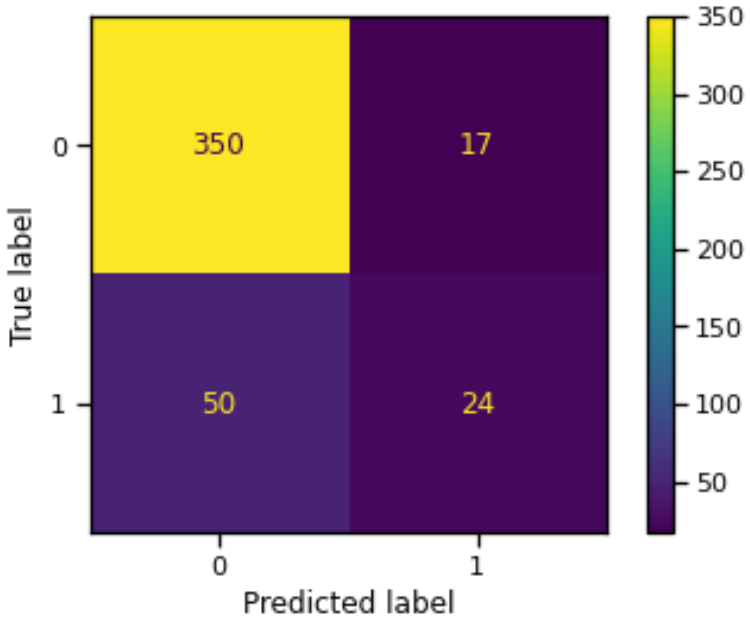
Από τον παραπάνω πίνακα παρατηρούμε ότι, η ορθότητα (Accuracy) είναι 84,81%. Δηλαδή το μοντέλο μας ταξινομεί σωστά το 84,81% όλων των εγγραφών.

Η ακρίβεια (Precision) είναι 87,50% για την κλάση 0 και 58,53% για την κλάση 1. Δηλαδή το μοντέλο προβλέπει σωστά στο 87,50% των περιπτώσεων που ανήκουν στην κλάση 0 ανάμεσα σε όσα έχει προβλέψει στην κλάση 0. Αντίστοιχα, προβλέπει σωστά το 58,53% όσων ανήκουν στην κλάση 1 απ' όσους έχουν προβλεφθεί στην κλάση 1.

Αναφορικά με την ανάκληση (Recall) που μας δείχνει ότι το μοντέλο μας αναγνωρίζει σωστά το 95,36% όλων των εγγραφών που ανήκουν πραγματικά στην κλάση 0 ενώ αναγνωρίζει σωστά μόνο το 32,43% απ' όλες τις εγγραφές που ανήκουν πραγματικά στην κλάση 1.

Το F1-Score για την κλάση 0 είναι 91,26% ενώ για την κλάση 1 είναι 41,73%. Αυτό σημαίνει ότι για την πρόβλεψη της κλάσης 0, δηλαδή των εργαζομένων που έχουν παραμείνει στην εταιρεία, ο αλγόριθμος έχει πολύ καλή απόδοση στην ταξινόμηση αυτών των εγγραφών.

Στη συνέχεια παρουσιάζεται ο πίνακας σύγχυσης (Confusion Matrix) του μοντέλου:

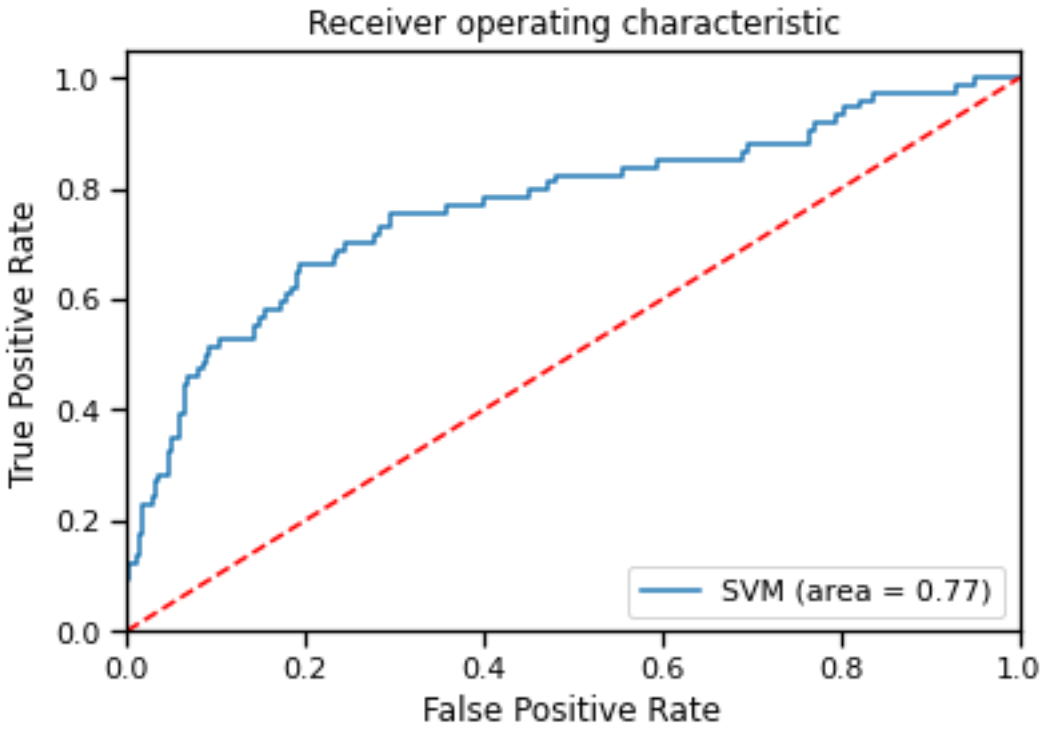


Εικόνα 29: Πίνακας σύγχυσης ταξινόμησης με χρήση αλγορίθμου SVM

Από τον πίνακα σύγχυσης της εικόνας 26 λαμβάνουμε τις εξής πληροφορίες:

- 350 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 0.
- 24 εργαζόμενοι έχουν προβλεφθεί σωστά ότι ανήκουν στην κλάση 1.
- 17 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 1 ενώ στην πραγματικότητα ανήκουν στην κλάση 0.
- 50 εργαζόμενοι έχουν προβλεφθεί ότι ανήκουν στην κλάση 0 ενώ στην πραγματικότητα ανήκουν στην κλάση 1.

Παρακάτω παρουσιάζεται η καμπύλη ROC:



Εικόνα 30: Καμπύλη ROC μοντέλου με χρήση αλγόριθμου SVM

Μπορούμε εύκολα να παρατηρήσουμε ότι η τιμή AUC είναι 0,77 η οποία μπορεί να θεωρηθεί καλή καθώς βρίσκεται κοντά στο 0,80.

Κεφάλαιο 7

Συμπεράσματα

Στην παρούσα εργασία, χρησιμοποιήθηκαν δεδομένα από την ιστοσελίδα Kaggle τα οποία αφορούν 32 στήλες με πληροφορίες για 1470 εργαζόμενους. Τα δεδομένα έχουν δημιουργηθεί από τους data scientists της εταιρείας IBM για σκοπούς ανάλυσης και εκπαίδευσης.

Στο πρώτο κεφάλαιο παρουσιάστηκαν οι λόγοι που οδηγούν στην αποχώρηση των εργαζομένων ενώ αναλύθηκαν και οι επιπτώσεις που έχουν οι αποχωρήσεις των εργαζομένων στις επιχειρήσεις. Τέλος, παρουσιάστηκαν οι στόχοι της διπλωματικής εργασίας καθώς και η βιβλιογραφική επισκόπηση εργασιών οι οποίες έχουν ασχοληθεί με το θέμα της παρούσας εργασίας.

Στο δεύτερο κεφάλαιο παρουσιάστηκε το θεωρητικό υπόβαθρο της μηχανικής μάθησης καθώς αναπτύχθηκε και το θεωρητικό υπόβαθρο και των αλγορίθμων Δέντρων Απόφασης, Τυχαίου Δάσους και Μηχανών Διανυσμάτων Υποστήριξης (SVM).

Στο τρίτο κεφάλαιο αναπτύξαμε εκτενώς το θεωρητικό υπόβαθρο της λογιστικής παλινδρόμησης.

Στο τέταρτο κεφάλαιο προχωρήσαμε σε μια παρουσίαση του συνόλου των δεδομένων που χρησιμοποιήθηκαν στην εργασία. Πιο συγκεκριμένα, προχωρήσαμε στην επεξήγηση των μεταβλητών, στην παρουσίαση των περιγραφικών μέτρων, ενώ είδαμε διάφορες διαγραμματικές απεικονίσεις των δεδομένων που έδωσαν πολύ χρήσιμα αποτελέσματα.

Στο πέμπτο κεφάλαιο προχωρήσαμε σε προεπεξεργασία των δεδομένων ώστε να είναι έτοιμο για την κατηγοριοποίηση, ενώ δεν παραλείψαμε να θέσουμε το θεωρητικό υπόβαθρο των τεχνικών αξιολόγησης των μοντέλων πρόβλεψης. Στη συνέχεια, διαχωρίσαμε το σύνολο των δεδομένων σε train set και test set με αναλογία 70% / 30% και προχωρήσαμε σε χρήση της λογιστικής παλινδρόμησης με χρήση δύο μοντέλων. Ένα μοντέλο λογιστικής παλινδρόμησης με χρήση όλων των μεταβλητών και ένα μοντέλο με χρήση ενός μοντέλου με τις δέκα πιο σημαντικές επεξηγηματικές μεταβλητές. Τέλος, παρουσιάστηκαν τα αποτελέσματα των αλγορίθμων και αξιολογήθηκαν βάσει των τεχνικών αξιολόγησης.

Στο έκτο κεφάλαιο χρησιμοποιήθηκαν οι αλγόριθμοι μηχανικής μάθησης όπου αναπτύχθηκαν στο δεύτερο κεφάλαιο ενώ αξιολογήθηκαν βάσει των τεχνικών αξιολόγησης που ορίσαμε στο πέμπτο κεφάλαιο. Και για τους τρείς αλγορίθμους της μηχανικής μάθησης, η διάσπαση του συνόλου δεδομένων είναι σε αναλογία 70% / 30%.

Τα αποτελέσματα όλων των αλγορίθμων, παρουσιάζονται συγκεντρωτικά στον παρακάτω πίνακα:

Algorithm	Precision		Recall		F1 - score		Accuracy	AUC
	0	1	0	1	0	1		
Logistic Regression (Full Model)	0.874384	0.714286	0.972603	0.328947	0.920882	0.450450	0.8614	0.82
Logistic Regression (Reduced Model)	0.862319	0.629630	0.972752	0.229730	0.914213	0.336634	0.8481	0.77
Decision Tree	0.879357	0.426471	0.893733	0.391892	0.886486	0.408451	0.8095	0.71
Random Forest	0.871046	0.7	0.975477	0.283784	0.920308	0.403846	0.8594	0.79
SVM	0.875	0.585366	0.953678	0.324324	0.912647	0.417391	0.8481	0.77

Πίνακας 12: Σύγκριση αποτελεσμάτων αλγορίθμων ταξινόμησης

Από τον πίνακα 12 φαίνεται πως από τους αλγορίθμους που χρησιμοποιήσαμε για την εφαρμογή ταξινόμησης ξεχωρίζει για την απόδοσή του ο αλγόριθμος της λογιστικής παλινδρόμησης με τη χρήση όλου του συνόλου μεταβλητών. Αξίζει να σημειωθεί πως ως τιμή 1 στην μεταβλητή Attrition έχουν οριστεί οι εργαζόμενοι οι οποίοι έχουν αποχωρήσει και είναι στο σύνολο 237 ενώ ως τιμή 0 οι εργαζόμενοι που έχουν παραμείνει στην εταιρεία και είναι στο σύνολο 1233. Στο σύνολο δεδομένων μόλις το 16,12% των εγγραφών έχουν την τιμή 1 ενώ το υπόλοιπο 83,88% έχει την τιμή 0. Είναι φυσικό πως οι αλγόριθμοι εκπαιδεύονται καλύτερα στην ταξινόμηση της τιμής 0 λόγω περισσότερων δεδομένων εκπαίδευσης πάνω στη συγκεκριμένη τιμή.

Στο υπάρχον σύνολο δεδομένων, υπάρχει πληθώρα δεδομένων τα οποία μπορούν να χρησιμοποιηθούν για περαιτέρω μελέτες και για πρόβλεψη οποιασδήποτε μεταβλητής, όχι μόνο της παραμονής και της αποχώρησης από μια εταιρεία.

Βιβλιογραφία

Ελληνική

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

Γκλώτσος, Δ., & Κάβουρας, Δ. (2023). Μηχανική Μάθηση [Κεφάλαιο 12]. Επεξεργασία Ιατρικής Εικόνας [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/11933

Ελευθεράκου, Ό. (2023, Μάρτιος). Εφαρμογές μηχανικής μάθησης στη νόσο του Parkinson, Διπλωματική εργασία, ΠΜΣ Εφαρμοσμένη Στατιστική, Πανεπιστήμιο Πειραιώς

Καλλιακμάνης, Δ. (2020). Στατιστικά μοντέλα για την απόδοση μιας ομάδας στο μπάσκετ: Ποια στατιστικά στοιχεία είναι καθοριστικά για τη απόδοση της ομάδας, σε ετήσια βάση, Διπλωματική εργασία, ΠΜΣ Εφαρμοσμένη Στατιστική, Πανεπιστήμιο Πειραιώς

Καμίτσης, Α. (2023, Ιούνιος). Δείκτες για την αξιολόγηση της απόδοσης παικτών και ομάδων σε αγώνες μπάσκετ και παράγοντες που τους επηρεάζουν. Διπλωματική εργασία, ΠΜΣ Εφαρμοσμένη Στατιστική, Πανεπιστήμιο Πειραιώς

Παναγιωτακόπουλος, Χ., Τσαλίδης, Χ., Γάκης, Π., & Κόκκινος, Θ. (2023). Μηχανική Μάθηση [Κεφάλαιο 10]. Υπολογιστική γλωσσολογία [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. https://hdl.handle.net/11419/9416

Πολίτης, Κ. (2021). Διαφάνειες μαθήματος «Γενικευμένα Γραμμικά Μοντέλα», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Πολίτης, Κ. (2022). Σημειώσεις μαθήματος «Γενικευμένα Γραμμικά Μοντέλα: Η έννοια της αλληλεπίδρασης», ΠΜΣ «Εφαρμοσμένη Στατιστική», Πανεπιστήμιο Πειραιώς.

Ξένη

Alon, Sigal. (2004). The Gender Stratification of Employment Hardship: Queuing, Opportunity Structure and Economic Cycles. Research in Social Stratification and Mobility

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Brink O. & Larsson H. (2019). Workplace values, sustainable employment and turnover intention: A general perspective, UMEA University

Cai, W. (2022). HRM risk early warning based on a hybrid solution of decision tree and support vector machine. Wireless Communications and Mobile Computing, 2022, 1-7.

Garg, S., Sinha, S., Kar, A.K. and Mani, M. (2021), "A review of machine learning applications in human resource management", International Journal of Productivity and Performance Management, https://doi.org/10.1108/IJPPM-08- 2020-0427

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3 (pp. 246-252). Springer Berlin Heidelberg.

Maertz, C. P., & Campion, M. A. (1998). "25 years of voluntary turnover research: A review and critique." International Review of Industrial and Organizational Psychology, 13, 49-81.

O'Connell M. & Kung M.-C. (2007). "Employee Turnover & Retention: Understanding the True Costs and Reducing them through Improved Selection Processes." Industrial Management Article

Ren, Q., Cheng, H., & Han, H. (2017, March). Research on machine learning framework based on random forest algorithm. In AIP conference proceedings (Vol. 1820, No. 1). AIP Publishing.

Schwanen, T., Dijst, M. (2002) Travel-time ratios for visits to the workplace: the relationship between commuting time and work duration.Trans. Research Part A, 36, 573-592.

Setiawan, I. A., Suprihanto, S., Nugraha, A. C., & Hutahaean, J. (2020, April). HR analytics: Employee attrition analysis using logistic regression. In IOP Conference Series: Materials Science and Engineering (Vol. 830, No. 3, p. 032001). IOP Publishing

Singer, G., & Cohen, I. (2020). An objective-based entropy approach for interpretable decision tree models in support of human resource management: The case of absenteeism at work. Entropy, 22(8), 821.

Tang, D. (2022). Optimization of Human Resource Management System Based on Data Mining Technology and Random Forest Algorithm. Wireless Communications and Mobile Computing, 2022.

Σύνδεσμοι

https://en.wikipedia.org/wiki/Kernel_method

https://en.wikipedia.org/wiki/Support_vector_machine

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

https://en.wikipedia.org/wiki/Probit

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

https://en.wikipedia.org/wiki/Logit

https://towardsdatascience.com/a-gentle-introduction-to-complementary-log-log-regression-8ac3c5c1cd83

https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

https://thedatascientist.com/understanding-tree-based-models-a-simple-guide/

https://plotly.com/python-api-reference/generated/plotly.graph_objects.Figure.html

https://hvplot.holoviz.org/reference/tabular/hist.html

https://scikit-learn.org/stable/modules/preprocessing.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

https://www.ey.com/en_gr/workforce/work-reimagined-survey

https://www.bls.gov/news.release/pdf/tenure.pdf

https://news.microsoft.com/2024/05/08/microsoft-and-linkedin-release-the-2024-work-trend-index-on-the-state-of-ai-at-work/

https://www.europeandatajournalism.eu/cp_data_news/the-great-turnover-record-resignations-and-job-vacancies-in-europe/

https://www.adecco-jobs.com/-/media/project/adeccogroup/articles/article-doc/adecco-2022--exploring-workers-professional-aspirations-report-executive-summary.pdf/?modified=20220125113246

https://cdn2.hubspot.net/hubfs/478187/2017%20Retention%20Report%20Campaign/Work%20Institute%202017%20-Retention%20Report.pdf

https://medium.com/dataseries/radial-basis-functions-rbf-kernels-rbf-networks-explained-simply-35b246c4b76c

https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Παράρτημα

Κώδικας Python

Κεφάλαιο 4

Εισαγωγή βιβλιοθηκών

```
!pip install -q hvplot

import warnings

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import plotly.express as px

import plotly.graph_objects as go

import scipy

import hvplot

import hvplot.pandas

from scipy.stats import chi2_contingency

from plotly.subplots import make_subplots

from plotly.offline import init_notebook_mode

from statistics import stdev

from pprint import pprint

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import RobustScaler, StandardScaler
```

```
from sklearn.model_selection import RandomizedSearchCV

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, roc_auc_score

warnings.filterwarnings("ignore")

import plotly.figure_factory as ff

init_notebook_mode(connected=True)

sns.set_context("notebook")
```

Import του dataset και έλεγχος missing values

```
data = pd.read_csv('C:/Users/kyrko/Desktop/IBM Dataset/hr ibm dataset2.csv', delimiter=',')

print("Υπάρχουν {:,} γραμμές και {} στήλες στο αρχείο δεδομένων.".format(data.shape[0], data.shape[1]))

print("Αριθμός ελλειπών δεδομένων {}.".format(data.isnull().sum().sum()))

data.head()
```

Διαγραφή στηλών EmployeeCount, EmployeeNumber, StandardHours, Over18

```
data = data.drop(['EmployeeCount' ,'EmployeeNumber', 'StandardHours', 'Over18'], axis=1)

data.head()


def ReplaceIt(i,st=""):
    data[st]=data[st].replace(1-i,'Low')
    data[st]=data[st].replace(2-i,'Medium')
    data[st]=data[st].replace(3-i,'High')
    data[st]=data[st].replace(4-i,'Very High')
```

Αλλαγή κωδικοποίησης κατηγορικών μεταβλητών

```
CatList = ["EnvironmentSatisfaction",
"JobSatisfaction","JobInvolvement","RelationshipSatisfaction","WorkLifeBalance","StockOp
tionLevel"]
i=0
for st in CatList:
    if st =="StockOptionLevel":
        i = i+1
    ReplaceIt(i,st)


data["JobLevel"] = data["JobLevel"].replace(
    {1:"Very Low",
     2:"Low",
     3:"Medium",
     4:"High",
     5:"Very High"}
)
data["Education"] = data["Education"].replace(
    {1:"Below College",
     2:"College",
     3:"Bachelor",
     4:"Master",
     5:"Doctor"}
)
data['PerformanceRating']=data['PerformanceRating'].replace(3,'High')
data['PerformanceRating']=data['PerformanceRating'].replace(4,'Very High')
data.head()
```

Περιγραφικά μέτρα αριθμητικών μεταβλητών

```
data_transp=data.describe()
data_transp.transpose()
```

Περιγραφικά μέτρα κατηγορικών μεταβλητών

```
kmet_cols=data.select_dtypes(include=object).columns.tolist()
kmet_df=pd.DataFrame(data[kmet_cols].melt(var_name='Column', value_name='Value')
                     .value_counts()).rename(columns={0: 'Count'}).sort_values(by=['Column',
'Count'])
display(data.select_dtypes(include=object).describe().transpose())
pd.set_option('display.max_rows', None)
display(kmet_df)
```

Δημιουργία ιστογραμμάτων

```
data.hist(figsize=(20,20))
plt.grid(False)
plt.axis('off')
plt.show()
```

Δημιουργία pie chart attrition vs Gender

```
diagr_fylo=data.groupby('Attrition',as_index=False)['Age'].count()
diagr_fylo['Count']=diagr_fylo['Age']
diagr_fylo.drop('Age',axis=1,inplace=True)
diagr_fylo2=data.groupby(['Gender','Attrition'],as_index=False)['Age'].count()
diagr_fylo2['Count']=diagr_fylo2['Age']
diagr_fylo2.drop('Age',axis=1,inplace=True)
fig=go.Figure()
fig=make_subplots(rows=1,cols=3)
fig = make_subplots(rows=1, cols=3, specs=[[{"type": "pie"}, {"type": "pie"}, {"type":
"pie"}]],subplot_titles=('<b>Employee Attrition', '<b>Female Attrition','<b>Male Attrition'))
fig.add_trace(go.Pie(values=diagr_fylo['Count'],labels=diagr_fylo['Attrition'],hole=0.7,marke
r_colors=['LightBlue','Pink'],name='Employee Attrition',showlegend=False),row=1,col=1)
```

```
fig.add_trace(go.Pie(values=diagr_fylo2[(diagr_fylo2['Gender']=='Female')]['Count'],labels=
diagr_fylo2[(diagr_fylo2['Gender']=='Female')]['Attrition'],hole=0.7,marker_colors=['DeepBl
ue','Pink'],name='Female Attrition',showlegend=False),row=1,col=2)

fig.add_trace(go.Pie(values=diagr_fylo2[(diagr_fylo2['Gender']=='Male')]['Count'],labels=dia
gr_fylo2[(diagr_fylo2['Gender']=='Male')]['Attrition'],hole=0.7,marker_colors=['DeepBlue','P
ink'],name='Male Attrition',showlegend=True),row=1,col=3)

fig.update_layout(title_x=0,template='simple_white',showlegend=True,legend_title_text="<b
style=\"font-size:90%;\">Attrition",title_text='<b style="color:black; font-
size:120%;"></b>',font_family="Calibri",title_font_family="Calibri")

fig.update_traces(marker=dict(line=dict(color='#000000', width=1)))
```

Δημιουργία pie chart attrition vs Business Travel

```
btr=data.groupby(['BusinessTravel','Attrition'],as_index=False)['Age'].count()

btr.rename(columns={'Age':'Count'},inplace=True)

fig=go.Figure()

fig=make_subplots(rows=1,cols=3)

fig = make_subplots(rows=1, cols=3, specs=[[{"type": "pie"}, {"type": "pie"}, {"type":
"pie"}]],subplot_titles=('<b>Non-Travel', '<b>Travel Frequently','<b>Travel Rarely'))

fig.add_trace(go.Pie(values=btr[btr['BusinessTravel']=='Non-
Travel']['Count'],labels=btr[btr['BusinessTravel']=='Non-
Travel']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Non-
Travel',showlegend=False),row=1,col=1)

fig.add_trace(go.Pie(values=btr[btr['BusinessTravel']=='Travel_Frequently']['Count'],labels=
btr[btr['BusinessTravel']=='Travel_Frequently']['Attrition'],hole=0.7,marker_colors=['LightBl
ue','Pink'],name='Travel Frequently',showlegend=False),row=1,col=2)

fig.add_trace(go.Pie(values=btr[btr['BusinessTravel']=='Travel_Rarely']['Count'],labels=btr[b
tr['BusinessTravel']=='Travel_Rarely']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'
],name='Travel Rarely',showlegend=True),row=1,col=3)

fig.update_layout(title_x=0.5,template='simple_white',showlegend=True,legend_title_text="
<b style=\"font-size:90%;\">Attrition",title_text='<b style="color:black; font-
size:120%;"></b>',font_family="Calibri")

fig.update_traces(marker=dict(line=dict(color='#000000', width=1)))
```

Δημιουργία pie chart attrition vs Education Field

```
edf=data.groupby(['EducationField','Attrition'],as_index=False)['Age'].count()

edf.rename(columns={'Age':'Count'},inplace=True)

fig=go.Figure()

fig = make_subplots(rows=2, cols=3, specs=[[{"type": "pie"}, {"type": "pie"}, {"type": "pie"}],[{"type": "pie"}, {"type": "pie"}, {"type": "pie"}]],subplot_titles=('<b>Life Sciences', '<b>Medical','<b>Marketing','<b>Technical Degree','<b>Human Resources','<b>Other'))


fig.add_trace(go.Pie(values=edf[edf['EducationField']=='Life Sciences']['Count'],labels=edf[edf['EducationField']=='Life Sciences']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Life Sciences',showlegend=False),row=1,col=1)

fig.add_trace(go.Pie(values=edf[edf['EducationField']=='Medical']['Count'],labels=edf[edf['EducationField']=='Medical']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Medical',showlegend=False),row=1,col=2)

fig.add_trace(go.Pie(values=edf[edf['EducationField']=='Marketing']['Count'],labels=edf[edf['EducationField']=='Marketing']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Marketing',showlegend=True),row=1,col=3)

fig.add_trace(go.Pie(values=edf[edf['EducationField']=='Technical Degree']['Count'],labels=edf[edf['EducationField']=='Technical Degree']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Technical Degree',showlegend=False),row=2,col=1)

fig.add_trace(go.Pie(values=edf[edf['EducationField']=='Human Resources']['Count'],labels=edf[edf['EducationField']=='Human Resources']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Human Resources',showlegend=False),row=2,col=2)

fig.add_trace(go.Pie(values=edf[edf['EducationField']=='Other']['Count'],labels=edf[edf['EducationField']=='Other']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Other',showlegend=False),row=2,col=3)

fig.update_layout(title_x=0.5,template='simple_white',showlegend=True,legend_title_text="<b>Attrition",font_family="Calibri",title_font_family="Calibri")

fig.update_traces(marker=dict(line=dict(color='#000000', width=1)))
```

Δημιουργία boxplots monthlyincome vs gender ανά attrition

```
boxplot1=px.box(data,x='Gender',y='MonthlyIncome',color='Attrition',template='simple_white',color_discrete_sequence=['Pink','LightBlue'])

boxplot1.update_xaxes(visible=True)

boxplot1.update_yaxes(visible=True)

boxplot1.update_layout(title_x=0.5,template='simple_white',showlegend=True,font_family="Calibri",title_font_family="Calibri")

boxplot1.show()
```

Δημιουργία boxplots age vs gender ανά attrition

```
boxplot2=px.box(data,x='Gender',y='Age',color='Attrition',template='simple_white',color_discrete_sequence=['Pink','LightBlue'])

boxplot2.update_traces(marker=dict(line=dict(color='#000000', width=0.5)))

boxplot2.update_xaxes(visible=True)

boxplot2.update_yaxes(visible=True)

boxplot2.update_layout(title_x=0.5,template='simple_white',showlegend=True,font_family="Calibri",title_font_family="Calibri")

boxplot2.show()
```

Δημιουργία boxplots Attrition vs DistanceFromHome

```
boxplot2=px.box(data,x='Attrition',y='DistanceFromHome',color='Attrition',template='simple_white',color_discrete_sequence=['Red','Green'])

boxplot2.update_traces(marker=dict(line=dict(color='#000000', width=0.5)))

boxplot2.update_xaxes(visible=True)

boxplot2.update_yaxes(visible=True)

boxplot2.update_layout(title_x=0.5,template='simple_white',showlegend=True,font_family="Calibri",title_font_family="Calibri")

boxplot2.show()
```

Δημιουργία barplots Attrition vs JobRole

```
jr=data.groupby(['JobRole','Attrition'],as_index=False)['Age'].count()

a=jr[jr['Attrition']=='Yes']
```

```
b=jr[jr['Attrition']=='No']

a['Age']=a['Age'].apply(lambda x: -x)

jr=pd.concat([a,b],ignore_index=True)

jr['Count']=jr['Age']

fig=px.bar(jr,x='JobRole',y='Count',color='Attrition',template='simple_white',text='Count',color_discrete_sequence=['Navy','Orange'])

fig.update_yaxes(range=[-100,350])

fig.update_traces(marker=dict(line=dict(color='#000000', width=0.1)),textposition = "outside")

fig.update_xaxes(visible=True)

fig.update_yaxes(visible=True)

fig.update_layout(title_x=0.5,template='simple_white',showlegend=True,title_text='<b style="color:black; font-size:105%;"> Attrition vs Job Roles</b>',font_family="Calibri",title_font_family="Calibri")

fig.show()
```

Δημιουργία pie chart attrition vs JobSatisfaction

```
js=data.groupby(['JobSatisfaction','Attrition'],as_index=False)['Age'].count()

js.rename(columns={'Age':'Count'},inplace=True)

fig=go.Figure()

fig = make_subplots(rows=1, cols=4, specs=[[{"type": "pie"}, {"type": "pie"},{"type": "pie"}, {"type": "pie"}]],subplot_titles=('Very High', 'High','Medium','Low'))

fig.add_trace(go.Pie(values=js[js['JobSatisfaction']=='Very High']['Count'],labels=js[js['JobSatisfaction']=='Very High']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Very High',showlegend=False),row=1,col=1)

fig.add_trace(go.Pie(values=js[js['JobSatisfaction']=='High']['Count'],labels=js[js['JobSatisfaction']=='High']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='High',showlegend=False),row=1,col=2)

fig.add_trace(go.Pie(values=js[js['JobSatisfaction']=='Medium']['Count'],labels=js[js['JobSatisfaction']=='Medium']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Medium',showlegend=False),row=1,col=3)
```

```
fig.add_trace(go.Pie(values=js[js['JobSatisfaction']=='Low']['Count'],labels=js[js['JobSatisfaction']=='Low']['Attrition'],hole=0.7,marker_colors=['LightBlue','Pink'],name='Low',showlegend=True),row=1,col=4)

fig.update_layout(title_x=0.5,template='simple_white',showlegend=True,legend_title_text="Attrition",font_family="Calibri",title_font_family="Calibri")

fig.update_traces(marker=dict(line=dict(color='#000000', width=1)))
```

Δημιουργία barplots Attrition vs JobInvolvement

```
jin=data.groupby(['JobInvolvement','Attrition'],as_index=False)['Age'].count()

c=jin[jin['Attrition']=='Yes']

d=jin[jin['Attrition']=='No']

c['Age']=c['Age'].apply(lambda x: -x)

jin=pd.concat([c,d],ignore_index=True)

jin['Count']=jin['Age']

fig=px.bar(jin,x='JobInvolvement',y='Count',color='Attrition',template='simple_white',text='Count',color_discrete_sequence=['Navy','Orange'])

fig.update_yaxes(range=[-300,1000])

fig.update_traces(marker=dict(line=dict(color='#000000', width=0.1)),textposition = "outside")

fig.update_xaxes(visible=True)

fig.update_yaxes(visible=True)

fig.update_layout(title_x=0.5,template='simple_white',showlegend=True,title_text='<b style="color:black; font-size:105%;"> Attrition vs Job Involvement</b>',font_family="Calibri",title_font_family="Calibri")

fig.show()
```

Κεφάλαιο 5

Αντιγραφή του αρχικού dataset

```
data_new=data.copy()
```

Μετασχηματισμός Attrition σε δίτιμη 0=No, 1=Yes

```
data_new['Attrition']=data_new['Attrition'].apply(lambda x: 0 if x == 'No' else 1)
```

Κωδικοποίηση όλων των κατηγορικών μεταβλητών

```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
le = LabelEncoder()
le_count = 0
for col in data_new.columns[1:]:
    if data_new[col].dtype == 'object':
        if len(list(data_new[col].unique())) <= 9:
            le.fit(data_new[col])
            data_new[col] = le.transform(data_new[col])
            le_count += 1
print('{} columns were label encoded.'.format(le_count))
```

Διαχωρισμός μεταβλητών σε μεταβλητή στόχο και επεξηγηματικών μεταβλητών

```
from sklearn.model_selection import train_test_split
X = data_new.drop('Attrition', axis=1)
y = data_new.Attrition
```

Διαχωρισμός σε train set & test set

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=27)
```

Εφαρμογή λογιστικής παλινδρόμησης (full model)

```
from sklearn.linear_model import LogisticRegression
lr_clf = LogisticRegression(solver='liblinear')
lr_clf.fit(X_train, y_train)
```

```
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report,
roc_auc_score


def evaluate(model, X_train, X_test, y_train, y_test):
    y_test_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)


    print("TRAINING RESULTS: \n===============================")
    clf_report = pd.DataFrame(classification_report(y_train, y_train_pred, output_dict=True))
    print(f"CONFUSION MATRIX:\n{confusion_matrix(y_train, y_train_pred)}")
    print(f"ACCURACY SCORE:\n{accuracy_score(y_train, y_train_pred):.4f}")
    print(f"CLASSIFICATION REPORT:\n{clf_report}")


    print("TESTING RESULTS: \n===============================")
    clf_report = pd.DataFrame(classification_report(y_test, y_test_pred, output_dict=True))
    print(f"CONFUSION MATRIX:\n{confusion_matrix(y_test, y_test_pred)}")
    print(f"ACCURACY SCORE:\n{accuracy_score(y_test, y_test_pred):.4f}")
    print(f"CLASSIFICATION REPORT:\n{clf_report}")
evaluate(lr_clf, X_train, X_test, y_train, y_test)
```

Confusion Matrix – LR (full model)

```
y_pred = lr_clf.predict(X_test)
from sklearn.metrics import ConfusionMatrixDisplay
reg_cm = confusion_matrix(y_test, y_pred, labels=lr_clf.classes_)
reg_disp = ConfusionMatrixDisplay(confusion_matrix=reg_cm,
display_labels=lr_clf.classes_)
reg_disp.plot(values_format='')
plt.show()
```

Διάγραμμα ROC Curve / AUC

```
from sklearn.metrics import roc_auc_score
```

```
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, lr_clf.predict_proba(X_test)[:,1])
fpr, tpr, thresholds = roc_curve(y_test, lr_clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

Πίνακας ANOVA – LR (full model)

```
formula = "Attrition ~ Age + BusinessTravel+ Department + DistanceFromHome +
Education + EducationField + EnvironmentSatisfaction + Gender + JobInvolvement +
JobLevel + JobRole + JobSatisfaction + MaritalStatus + MonthlyIncome +
NumCompaniesWorked + OverTime + PercentSalaryHike + PerformanceRating +
RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear
+ WorkLifeBalance + YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
YearsWithCurrManager"
```

Επιλογή επεξηγηματικών μεταβλητών LR (Reduced model)

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import SelectKBest, chi2
selector = SelectKBest(chi2, k=10)
X_new = selector.fit_transform(X, y)
```

```
print(selector.get_support())

print(X.columns.values)
```

Πίνακας ANOVA – LR (reduced model)

```
formula_rm = "Attrition ~ Age + DistanceFromHome + JobLevel + MaritalStatus +
MonthlyIncome + OverTime + TotalWorkingYears+ YearsAtCompany + YearsInCurrentRole
+ YearsWithCurrManager"

mod_rm = smf.glm(formula=formula_rm, data=data_new,
family=sm.families.Binomial()).fit()

print(mod_rm.summary())
```

Εφαρμογή λογιστικής παλινδρόμησης (reduced model)

```
X_rm =
data_new[['Age','DistanceFromHome','JobLevel','MaritalStatus','MonthlyIncome','OverTime',
'TotalWorkingYears','YearsAtCompany','YearsInCurrentRole', 'YearsWithCurrManager']]

Xrm_train, Xrm_test, yrm_train, yrm_test = train_test_split(X_rm, y, test_size=0.3,
random_state=27)

from sklearn.linear_model import LogisticRegression

lr_clf_rm = LogisticRegression(solver='liblinear')

lr_clf_rm.fit(Xrm_train, yrm_train)

evaluate(lr_clf_rm, Xrm_train, Xrm_test, yrm_train, yrm_test)
```

Confusion Matrix – LR (reduced model)

```
y_pred_rm = lr_clf_rm.predict(Xrm_test)

from sklearn.metrics import ConfusionMatrixDisplay

reg_cm_rm = confusion_matrix(yrm_test, y_pred_rm, labels=lr_clf_rm.classes_)

reg_disp_rm = ConfusionMatrixDisplay(confusion_matrix=reg_cm_rm,
display_labels=lr_clf_rm.classes_)

reg_disp_rm.plot(values_format='')

plt.show()
```

Διάγραμμα ROC Curve / AUC

```
from sklearn.metrics import roc_auc_score
```

```
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(yrm_test, lr_clf_rm.predict_proba(Xrm_test)[:,1])
fpr, tpr, thresholds = roc_curve(yrm_test, lr_clf_rm.predict_proba(Xrm_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

Κεφάλαιο 6

Εφαρμογή αλγορίθμου random forest

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=27)
from sklearn.ensemble import RandomForestClassifier
rf_clf = RandomForestClassifier(n_estimators=100, bootstrap=False)
rf_clf.fit(X_train, y_train)
evaluate(rf_clf, X_train, X_test, y_train, y_test)
print(classification_report(y_test, y_pred_rf, digits=4))
print('AUC score:', roc_auc_score(y_test, rf_clf.predict_proba(X_test)[:,1]))
```

Confusion Matrix – Random Forest

```
y_pred_rf = rf_clf.predict(X_test)
from sklearn.metrics import ConfusionMatrixDisplay
```

```
reg_cm = confusion_matrix(y_test, y_pred_rf, labels=rf_clf.classes_)
reg_disp = ConfusionMatrixDisplay(confusion_matrix=reg_cm,
display_labels=rf_clf.classes_)
reg_disp.plot(values_format='')
plt.show()
```

Διάγραμμα ROC Curve / AUC

```
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
logit_roc_auc = roc_auc_score(y_test, rf_clf.predict_proba(X_test)[:,1])
fpr, tpr, thresholds = roc_curve(y_test, rf_clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Random Forest (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

Εφαρμογή αλγορίθμου SVM

```
from sklearn.svm import SVC
svm_clf = SVC(kernel='linear', random_state=42,probability=True)
svm_clf.fit(X_train, y_train)
evaluate(svm_clf, X_train, X_test, y_train, y_test)
print(classification_report(y_test, y_pred_svm, digits=4))
logit_roc_auc = roc_auc_score(y_test, svm_clf.predict_proba(X_test)[:,1])
```

```
fpr, tpr, thresholds = roc_curve(y_test, svm_clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='SVM (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

Confusion Matrix – SVM

```
y_pred_svm = svm_clf.predict(X_test)
from sklearn.metrics import ConfusionMatrixDisplay
reg_cm = confusion_matrix(y_test, y_pred_svm, labels=svm_clf.classes_)
reg_disp = ConfusionMatrixDisplay(confusion_matrix=reg_cm,
display_labels=svm_clf.classes_)
reg_disp.plot(values_format='')
plt.show()
```

Διάγραμμα ROC Curve / AUC

```
print(classification_report(y_test, y_pred_svm, digits=4))
logit_roc_auc = roc_auc_score(y_test, svm_clf.predict_proba(X_test)[:,1])
fpr, tpr, thresholds = roc_curve(y_test, svm_clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='SVM (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
```

```
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```

Εφαρμογή αλγορίθμου Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
dt_clf = DecisionTreeClassifier(criterion='gini',max_depth = 4,random_state=42)
dt_clf.fit(X_train, y_train)
evaluate(dt_clf, X_train, X_test, y_train, y_test)
print(classification_report(y_test, y_pred_dt, digits=4))
```

Confusion Matrix – Decision Tree

```
y_pred_dt = dt_clf.predict(X_test)
from sklearn.metrics import ConfusionMatrixDisplay
reg_cm = confusion_matrix(y_test, y_pred_dt, labels=dt_clf.classes_)
reg_disp = ConfusionMatrixDisplay(confusion_matrix=reg_cm,
display_labels=dt_clf.classes_)
reg_disp.plot(values_format='')
plt.show()
```

Διάγραμμα ROC Curve / AUC

```
logit_roc_auc = roc_auc_score(y_test, dt_clf.predict_proba(X_test)[:,1])
fpr, tpr, thresholds = roc_curve(y_test, dt_clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='SVM (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
```

```
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```