ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ

(M.Sc. in Business Administration, Analytics and

Information Systems)

Διπλωματική Εργασία

Διερεύνηση αιτιών εργασιακής φθοράς με χρήση μεθόδων ταξινόμησης

Investigating the reasons of employee attrition with classification methods

Μαρία Γιαγκλή

Επιβλέπων Καθηγητής: Ευάγγελος Βασιλείου

Αθήνα, Απρίλιος 2019

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ

(M.Sc. in Business Administration, Analytics and

Information Systems)

Διπλωματική Εργασία

Διερεύνηση αιτιών εργασιακής φθοράς με χρήση μεθόδων ταξινόμησης

Investigating the reasons of employee attrition with classification methods

Μαρία Γιαγκλή

Επιβλέπων Καθηγητής: Ευάγγελος Βασιλείου

Αθήνα, Απρίλιος 2019

Διπλωματική εργασία υποβληθείσα στο Τμήμα Οικονομικών Επιστημών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, ως μέρους των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Διοίκηση, Αναλυτική και Πληροφοριακά Συστήματα Επιχειρήσεων.

Η Γιαγκλή Μαρία βεβαιώνει ότι η υποβληθείσα εργασία είναι προσωπική εκτός από τα σημεία όπου γίνεται αναφορά στις εργασίες άλλων.

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Ευάγγελο Βασιλείου για την εξαιρετική συνεργασία που είχαμε, για την καθοδήγηση και βοήθεια που μου παρείχε καθ' όλη την διάρκεια εκπόνησης της παρούσας εργασίας, καθώς και για την επιμονή και το κίνητρο που μου καλλιέργησε, προκειμένου να δώσω το μέγιστο των δυνατοτήτων μου. Ακολούθως, ένα μεγαλύτερο ευχαριστώ οφείλω στους γονείς μου, για την υποστήριξή τους, ψυχολογική και οικονομική και για την παρότρυνσή τους, να ολοκληρώσω έναν ακόμη απαιτητικό κύκλο σπουδών. Τέλος ευχαριστώ τους φίλους μου, για την ψυχολογική υποστήριξη και βοήθεια που μου παρείχαν όντας δίπλα μου, σε όλη την διάρκεια της συγκεκριμένης απαιτητικής διαδικασίας.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΠΕΡΙΛΗΨΗ

Η εργασιακή φθορά αποτελεί ένα από τα σοβαρότερα προβλήματα που αντιμετωπίζουν οι οργανισμοί στην παγκόσμια οικονομία. Το ανθρώπινο δυναμικό μπορεί να αποτελέσει καθοριστικό παράγοντα ανταγωνιστικού πλεονεκτήματος για τους οργανισμούς, αλλά ταυτόχρονα και μια δαπανηρή πρόκληση για αυτούς. Για τον λόγο αυτό οι εταιρίες επιθυμούν να κατανοήσουν καλύτερα τα κύρια ζητήματα, από τα οποία πηγάζει το φαινόμενο της εργασιακής φθοράς. Η επιλογή κάποιου εργαζομένου να παραιτηθεί, μπορεί να οφείλεται σε πολλούς διαφορετικούς λόγους. Ωστόσο είναι σημαντικό για τις εταιρίες να εντοπίσουν τους πιο βαρυσήμαντους και να προβλέψουν ποιοι είναι οι εργαζόμενοι, που έχουν υψηλή πιθανότητα να φύγουν, ώστε να προβούν σε αλλαγές και να βελτιώσουν την διαχείριση του ανθρώπινου δυναμικού τους .

Στην παρούσα εργασία επιλέξαμε να μελετήσουμε τους αλγορίθμους στατιστικής μάθησης, προκειμένου να εντοπίσουμε τους κύριους λόγους της εργασιακής φθοράς, μέσω των μεθόδων εξόρυξης δεδομένων. Συγκεκριμένα, αναλύουμε τους αλγορίθμους Logistic Regression, Decision Trees και Random Forest χρησιμοποιώντας το περιβάλλον της R και R studio. Η μελέτη διεξάγεται σε ένα σύνολο δεδομένων που παρέχεται από το kaggle.com, το οποίο αποτελεί ιστοσελίδα διαγωνισμών εξόρυξης δεδομένων. Το συγκεκριμένο σύνολο δεδομένων αποτελείται από περίπου 30.000 εγγραφές και εξετάζει ορισμένα από τα χαρακτηριστικά των εργαζομένων, όπως το επίπεδο ικανοποίησης τους, τον βαθμό τελευταίας αξιολόγησης τους, τον μισθό και ούτω καθ’ εξής.

Η μελέτη εστιάζει αρχικά στην θεωρητική αποτύπωση των αλγορίθμων και την συνέχεια στην πρακτική εφαρμογή και ανάλυση τους, στο σύνολο των δεδομένων μας. Ύστερα από επαναληπτική παραμετροποίηση και αξιολόγηση του κάθε αλγορίθμου, καταλήγουμε στο καλύτερο δυνατό μοντέλο για τον καθένα από αυτούς. Συγκρίνοντας τις επιδόσεις τους στα άγνωστα δεδομένα, επιλέγουμε ένα τελικό μοντέλο μέσω του οποίου εξάγουμε τα συμπεράσματα μας. Τέλος παραθέτουμε κάποιες προτάσεις για περαιτέρω έρευνα στο μέλλον.

Η συγκεκριμένη περιοχή μελέτης έχει ευρύ φάσμα πρακτικής και χρήσιμης εφαρμογής. Αν επιλυθεί ορθά, η ίδια μέθοδος και μοντέλο μπορούν να χρησιμοποιηθούν για μελέτη άλλων προβλημάτων όπως η απώλεια πελατών, η διαχείριση των πελατειακών σχέσεων, η ανεργία των νέων και άλλα.

Λέξεις Κλειδιά: μέθοδοι στατιστικής μάθησης, επιβλεπόμενη μάθηση, αλγόριθμοι ταξινόμησης, Λογιστική Παλινδρόμηση, Δέντρα Αποφάσεων, Τυχαία Δάση, μέτρα αξιολόγησης ταξινόμησης

ABSTRACT

Employee attrition is one of the most serious issues for organizations in current global economy. Human resources can be a key factor of organizations' competitive advantage, but at the same time a costly challenge for them. For this reason, companies want to understand better the main issues behind employee attrition phenomenon. The decision of an employee to quit may be due to a variety of reasons. However, it's important for companies to find out the most significant of them and predict who has the highest possibility to leave, so that they proceed to some changes and make a better human resource management.

In this thesis, we chose to study statistical learning algorithms, in order to locate the main reasons of employee attrition, through data mining methods. Specifically, we analyze the algorithms of Logistic Regression, Decision Trees and Random Forest by using environment of R and R studio. The study is conducted on a dataset provided by kaggle.com, which is a data mining competition website. This data set has about 30.000 records and considers some of the employees' characteristics, such as satisfaction level, last evaluation rating, salary and so on.

The study firstly focuses on the theoretical depiction of algorithms and then on practical implementation and analysis of our data set. After iterative parameterization and evaluation of each algorithm, we conclude to the best possible model for each of them. By comparing their performance to unknown data, we select a final model, through which we draw our conclusions. At last we provide some suggestions for further research.

This study area has a wide range of practical and useful application. If it has been resolved properly, the same method and model can be used to study other problems such as customer churn, customer relation management, youth unemployment and more.

Key Words: statistical learning methods, supervised learning, classification algorithms, Logistic Regression, Decision Trees, Random Forest, evaluation metrics of classification

1. ΕΙΣΑΓΩΓΗ

1.1 ΣΚΟΠΟΣ ΜΕΛΕΤΗΣ

Η παρούσα εργασία μελετά την περίπτωση της εργασιακής φθοράς (employee attrition). Η εργασιακή φθορά, στον χώρο των Human Resources, αναφέρεται στο φαινόμενο κατά το οποίο οι εργαζόμενοι φεύγουν από την εταιρία στην οποία βρίσκονται. Στα πλαίσια μιας εταιρίας μετράται με τον δείκτη του attrition rate, ο οποίος καταμετρά το ποσοστό των υπαλλήλων που μετακινούνται εκτός της εταιρίας, είτε γιατί παραιτούνται εθελοντικά είτε γιατί απολύονται από αυτήν. Ορισμένοι από τους κύριους λόγους που συντελούν στην απόφαση ενός εργαζομένου να αποχωρήσει από την εργασία του είναι: η ύπαρξη καλύτερων ευκαιριών εκτός του οργανισμού, η μεγάλη ανισορροπία μεταξύ επαγγελματικής και προσωπικής ζωής, η έλλειψη αναγνώρισης ή η περιορισμένη δυνατότητα ανάπτυξης στον τρέχοντα ρόλο, η στασιμότητα στην επαγγελματική σταδιοδρομία, οι ανεπαρκείς και κακές συνθήκες εργασίας, η κακή συμπεριφορά των ανωτέρων και των συναδέλφων, όπως και άλλοι παρατίθενται στην εικόνα 1 (Carey & Ogden, 2004).

Εικόνα 1. Λόγοι Εργασιακής Φθοράς

Ο δείκτης του attrition rate (Attrition, 2018) δίνει μια εικόνα σχετικά με το ποσοστό των υπαλλήλων που έφυγαν από την εταιρία έως κάποια δεδομένη χρονική στιγμή. Αποτελεί σημαντικό παράγοντα για τις εταιρίες, καθώς απαιτείται να εκκινήσουν ξανά την διαδικασία πρόσληψης, για θέσεις οι οποίες είναι πρωταρχικής σημασίας για αυτές και δεν μπορούν να μείνουν κενές. Ταυτόχρονα αποτελεί παράγοντα ανησυχίας για αυτές, καθώς τους δημιουργεί

αρκετά προβλήματα. Ένα από τα σημαντικότερα είναι η αύξηση του κόστους τους. Σύμφωνα με μελέτες που έχουν πραγματοποιηθεί η αντικατάσταση ενός εργαζομένου, μπορεί να οδηγήσει σε αύξηση του κόστους των εταιριών, κατά μέσο όρο 16% - 20% του ετήσιου μισθού του εργαζομένου. Ενώ άλλες μελέτες αναφέρουν ότι στην περίπτωση των υψηλόμισθων εργαζομένων ή υψηλόβαθμων στελεχών, η αύξηση του κόστους μπορεί να φτάσει ακόμη και το 200% του ετήσιου μισθού τους (Merhar, 2016). Άλλα προβλήματα στα οποία μπορεί να οδηγήσει είναι η μείωση της ικανοποίησης και της δέσμευσης του πελάτη, η μείωση της παραγωγικότητας των εργαζομένων και της δυναμικότητας της ομάδας, η αύξηση του χρόνου ολοκλήρωσης των διαφόρων διαδικασιών και άλλα (Mohr & et al, 2012). Επομένως, είναι μείζονος σημασίας για αυτές, η συστηματική παρακολούθηση του attrition rate, που οδηγεί σε μείωση του εργατικού δυναμικού τους.

Έτσι, η ανάγκη παρακολούθησης αυτού του μέτρου, σε συνδυασμό με την ραγδαία ανάπτυξη της πληροφορικής και της τεχνολογίας τα τελευταία χρόνια, οδήγησαν τον τομέα του Human Resource Management, στην ολοένα και μεγαλύτερη χρήση μεθόδων εξόρυξης δεδομένων (data mining). Οι διάφορες μεθοδολογίες εξόρυξης δεδομένων δίνουν την δυνατότητα στις εταιρίες να πάρουν σωστότερες αποφάσεις, να διαχειριστούν καλύτερα τους νέους υπαλλήλους και επιπλέον να αναλύσουν την εργασιακή φθορά παλαιότερων υπαλλήλων τους (Ranjan & et al, 2008). Επομένως η χρήση τους μπορεί να καταστεί ιδιαίτερα χρήσιμη για το Human Resource Management σε τρεις περιοχές: (1) στην ανάλυση της εργασιακής φθοράς, (2) στον προγραμματισμό εργασίας και (3) στην ανάλυση προσλήψεων (Mobley, 1982).

Υπάρχουν διάφορες μέθοδοι εξόρυξης δεδομένων, που χρησιμοποιούνται στην επίλυση τέτοιου είδους προβλημάτων. Στην παρούσα εργασία, από το σύνολο των αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης που χρησιμοποιούνται στην εξόρυξη δεδομένων, θα ασχοληθούμε με αυτές που κάνουν χρήση των αρχών στατιστικής μάθησης και πιο συγκεκριμένα αυτών, που αφορούν μεθόδους ταξινόμησης (Witten, 2011). Άμεσος στόχος μας είναι, μέσω των μεθόδων **Ανάλυσης ταξινόμησης** (Classification Analysis), να αναγνωρίσουμε ποιοι είναι οι σημαντικοί παράγοντες που επηρεάζουν έναν εργαζόμενο στην απόφασή του να παραιτηθεί. Ή με άλλα λόγια, να προβλέψουμε αν δεδομένων κάποιων συνθηκών, ένας εργαζόμενος πρόκειται να φύγει από την εταιρία. Παράλληλα έμμεσος στόχος της εργασίας είναι να παρέχει χρήσιμη μεθοδολογία και συμπεράσματα στον τομέα της Διοίκησης Ανθρώπινων Πόρων, για την πρόβλεψη ή/και αντιμετώπιση του φαινομένου αυτού.

Από το σύνολο των αλγορίθμων ταξινόμησης, η μελέτη μας θα επικεντρωθεί στην ανάλυση της μεθοδολογίας των αλγορίθμων **Logistic Regression**, **Decision Trees** και **Random Forest**. Από την εφαρμογή των τριών αυτών αλγορίθμων στα δεδομένα της βάσης μας και την σύγκριση των αποτελεσμάτων που θα προκύψουν, από τις διάφορες δοκιμές που θα πραγματοποιήσουμε, θα

καταλήξουμε στην επιλογή ενός μοντέλου για κάθε αλγόριθμο, αυτού που προσαρμόζεται καλύτερα στα δεδομένα. Η σύγκριση των αποτελεσμάτων ύστερα από κάθε δοκιμή, θα πραγματοποιηθεί με την χρήση κάποιων μέτρων αξιολόγησης. Τέτοια μέτρα είναι το confusion matrix, η καμπύλη ROC, το AUC, το Recall κ.α. Τα μέτρα αυτά θα μας βοηθήσουν να κατανοήσουμε καλύτερα, αν το μοντέλο που πρόκυψε από την εκπαίδευση του αλγορίθμου, δίνει ικανοποιητικά αποτελέσματα ή αν χρήζει επιπλέον βελτιώσεων. Μετά την κατασκευή ενός μοντέλου για κάθε αλγόριθμο, θα προβούμε στην μεταξύ τους σύγκριση, κάνοντας χρήση των ίδιων μέτρων αξιολόγησης, για να αποφασίσουμε ποιο θα είναι το τελικό μοντέλο με το οποίο θα πραγματοποιηθεί η ταξινόμηση των δεδομένων. Η προσαρμογή όλων των αλγορίθμων και η δημιουργία των γραφημάτων της ανάλυσης, θα γίνει με χρήση του ανοιχτού λογισμικού της R και με τη συγγραφή του σχετικού κώδικα στο περιβάλλον του R Studio.

Στην παρούσα μελέτη θα χρησιμοποιήσουμε δεδομένα εργαζομένων πολυεθνικής εταιρίας, στα οποία περιλαμβάνονται επίσης στοιχεία πρώην υπαλλήλων της, οι οποίοι αποχώρησαν οικειοθελώς. Οι τιμές που σημειώνονται στα χαρακτηριστικά των υπαλλήλων που παραμένουν στην εταιρία, είναι αυτές που καταγράφονταν στη βάση δεδομένων της εταιρίας, την χρονική στιγμή που συγκεντρώθηκε το συγκεκριμένο σύνολο δεδομένων. Από την άλλη, οι τιμές των χαρακτηριστικών που επισημαίνονται για τους υπαλλήλους που παραιτήθηκαν, αφορούν τα τελευταία στοιχεία που ήταν καταγεγραμμένα στη βάση της εταιρίας , κατά την χρονική στιγμή της παραίτησης τους. Μέσω αυτών, θα μας δοθεί η δυνατότητα να εντοπίσουμε τους σημαντικότερους λόγους, που συμβάλλουν στην εμφάνιση του προβλήματος της εργασιακής φθοράς. Τα αποτελέσματα που θα προκύψουν από αυτή την διερεύνηση είναι αυτά, που πρέπει να ληφθούν πρώτα υπόψη από το τμήμα ανθρώπινων πόρων, για την υποβοήθηση τους στην διαδικασία λήψης αποφάσεων. Η διάρθρωση των κεφαλαίων που θα ακολουθήσουν είναι η κάτωθι:

Στο Κεφάλαιο 1 θα πραγματοποιηθεί αρχικά η περιληπτική αποτύπωση των αλγορίθμων επιβλεπόμενης μάθησης (Supervised Learning), καθώς αποτελούν το ευρύτερο πλαίσιο μέσα στο οποίο ανήκουν οι μέθοδοι του classification. Ακολούθως θα γίνει η συνοπτική παρουσίαση των μεθόδων με τις οποίες λειτουργούν οι αλγόριθμοι Logistic Regression, Decision Trees και Random Forest, όπως επίσης και των διαφόρων μέτρων που χρησιμοποιούνται για την αξιολόγησης των υποδειγμάτων.

Στο Κεφάλαιο 2 θα πραγματοποιηθεί αναλυτική περιγραφή των στοιχείων που συνθέτουν τη βάση δεδομένων μας και θα παρατεθούν στοιχεία περιγραφικής στατιστικής και γραφήματα ώστε να κατανοήσουμε καλύτερα τα δεδομένα και τα ιδιαίτερα χαρακτηριστικά τους.

Στο Κεφάλαιο 3 θα πραγματοποιηθεί η αποτύπωση της μεθοδολογία που ακολουθούμε για τα δεδομένα της βάσης μας, σε καθέναν από τους τρεις αλγορίθμους και ο υπολογισμός των

μέτρων αξιολόγησης τους, μέχρι να οδηγηθούμε τελικά στο μοντέλο που δίνει τα καλύτερα αποτελέσματα σύμφωνα με το πρόβλημα που εξετάζουμε.

Στο Κεφάλαιο 4 θα καταγραφούν τα ευρήματα που θα προκύψουν από την ανάλυση και ακολούθως τα συμπεράσματα που εξάγονται από αυτά.

1.2 ΕΠΟΠΤΕΥΟΜΕΝΗ ΜΑΘΗΣΗ

Η στατιστική μάθηση αναφέρεται σε εκείνες τις μεθόδους μελέτης δεδομένων, οι οποίες αποσκοπούν να λύσουν το πρόβλημα εξαγωγής συμπερασμάτων, βασιζόμενες στην στατιστική. Διαχωρίζονται σε μεθόδους εποπτευόμενης και μη εποπτευόμενης μάθησης (Chapmann, 2017). Στην περίπτωση της εποπτευόμενης μάθησης (Supervised Learning) η προσπάθεια εξαγωγής συμπερασμάτων καθοδηγείται από κάποια γνωστή μεταβλητή απόκρισης και γίνεται βάσει κάποιων επεξηγηματικών μεταβλητών. Αντίθετα η μη εποπτευόμενη μάθηση βασίζεται μόνο σε επεξηγηματικές μεταβλητές, καθώς δεν υπάρχει κάποια γνωστή μεταβλητή απόκρισης και προσπαθεί να εξάγει πρότυπα ή σχέσεις, ομαδοποιώντας κοινές ή μη κοινές συμπεριφορές. Ακολούθως, γίνεται διάκριση των επιβλεπόμενων μεθόδων μάθησης ανάλογα με το είδος της μεταβλητής απόκρισης. Έτσι αν η μεταβλητή στόχος είναι αριθμητική και σ υνεχής, η μέθοδος καλείται παλινδρόμηση (regression) ενώ αν είναι κατηγορική, τότε καλείται ταξινόμηση (classification) (Witten, 2011). Στην παρούσα υποενότητα θα ασχοληθούμε εκτενέστερα με τις μεθόδους επιβλεπόμενης μάθησης και τη διάκριση μεταξύ τους.

1.2.1 ΤΙ ΕΙΝΑΙ Η ΕΠΟΠΤΕΥΟΜΕΝΗ ΜΑΘΗΣΗ

Ο όρος εποπτευόμενη μάθηση, αναφέρεται σε μια από τις βασικές κατηγορίες προβλημάτων που καλείται να λύσει η στατιστική μάθηση (Vapnik, 2000). Πιο αναλυτικά, η εποπτευόμενη μάθηση (James & et al, 2013, p. 25) αφορά στη διαδικασία μάθησης από ένα σύνολο δεδομένων εκπαίδευσης (training set), τα οποία θεωρείται ότι «εποπτεύουν» τη διαδικασία αυτή. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο ζευγών εισόδου - εξόδου, γνωστών σε εμάς. Ως στοιχεία εισόδου έχουμε συνήθως έναν πίνακα με τις τιμές των διαφόρων χαρακτηριστικών, τα οποία αποτελούν τις ανεξάρτητες μεταβλητές και ως στοιχεία εξόδου ένα διάνυσμα από τιμές – μετρήσεις για την μεταβλητή απόκρισης που εξετάζουμε, δηλαδή την εξαρτημένη μεταβλητή. Οι αλγόριθμοι εποπτευόμενης μάθησης αναλύουν τα δεδομένα εκπαίδευσης και παράγουν μια σχέση, που σχετίζει το αποτέλεσμα της εξαρτημένης μεταβλητής, με τις διάφορες τιμές των ανεξάρτητων μεταβλητών. Η σχέση που προκύπτει από την ανάλυση των αποτελεσμάτων, των δεδομένων εκπαίδευσης (training set) καλείται μοντέλο. Στόχος μας είναι να σχεδιάσουμε ένα μοντέλο που να μπορεί να πραγματοποιεί ακριβή πρόβλεψη, για την τιμή της εξαρτημένης μεταβλητής, όταν εισάγουμε σε αυτό άγνωστες

παρατηρήσεις ή να κατανοεί καλύτερα την σχέση που υπάρχει μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Το ιδανικό σενάριο είναι ο αλγόριθμος να μπορεί να καθορίσει ορθά, το αποτέλεσμα της κατηγορίας που εξετάζουμε, σε άγνωστα για αυτόν δεδομένα. Για να υλοποιηθεί ωστόσο ένα τέτοιο σενάριο, θα πρέπει η γενίκευση αυτή να γίνει με ένα «λογικό» τρόπο.

1.2.2 ΠΟΙΑ Η ΔΙΑΦΟΡΟΠΟΙΗΣΗ ΜΕΤΑΞΥ ΤΩΝ ΜΕΘΟΔΩΝ

Οι αλγόριθμοι εποπτευόμενης μάθησης (Breiman L. , 2001) μπορούν να χρησιμοποιηθούν σε ένα μεγάλο εύρος προβλημάτων, τα οποία αντίστοιχα μπορούν να απαντηθούν μέσω πολλών διαφορετικών μεθόδων. Για την καλύτερη κατανόηση των αλγορίθμων αυτών, κρίνεται αναγκαίος ο διαχωρισμός τους σε υποομάδες, ανάλογα με το είδος των προβλημάτων που καλούνται να επιλύσουν. Έτσι η διάκριση των μεθόδων μπορεί να πραγματοποιηθεί είτε ανάλογα με το είδος της εξαρτημένης μεταβλητής που αυτά μελετούν, είτε σύμφωνα με το είδος της σχέσης που υπάρχει μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών. Κάθε μια από τις υποομάδες αυτές αναλύεται εκτενέστερα παρακάτω.

Ι.ΤΑΞΙΝΟΜΗΣΗ ΕΝΑΝΤΙ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Οι συγκεκριμένες δύο κατηγορίες μεθόδων απαντούν στην διάκριση ανάλογα με το είδος της εξαρτημένης μεταβλητής, που αναφέρθηκε παραπάνω. Η εξαρτημένη μεταβλητή σε ένα πρόβλημα μπορεί να είναι είτε ποσοτική είτε ποιοτική. Οι ποσοτικές μεταβλητές λαμβάνουν αριθμητικές και συνεχείς τιμές, ενώ οι ποιοτικές λαμβάνουν κατηγορικές και διακριτές τιμές. Επομένως, στην περίπτωση που το αποτέλεσμα της πρόβλεψής μας αντιπροσωπεύεται από μια συνεχή αριθμητική τιμή, που μπορεί να προσδιοριστεί κάθε φορά βάσει των τιμών των ανεξάρτητων μεταβλητών και όχι από ένα περιορισμένο σύνολο δυνατών αποτελεσμάτων, τότε η μέθοδος ονομάζεται *παλινδρόμηση* (Castle, 2018). Παραδείγματα που εμπίπτουν σε αυτή την κατηγορία είναι η πρόβλεψη της τιμής των κατοικιών, η πρόβλεψη των εσόδων που θα επιφέρει μια νέα διαφημιστική καμπάνια ή η πρόβλεψη των τιμών των μετοχών στις χρηματιστηριακές αγορές. Από την άλλη, όταν το αποτέλεσμα της πρόβλεψης μας εμπίπτει σε ένα περιορισμένο σύνολο k κατηγόριων ή κλάσεων τότε η μέθοδος καλείται *ταξινόμηση*. Παραδείγματα τέτοιων προβλημάτων είναι η πρόβλεψη του αν ένας πελάτης θα αποπληρώσει ή όχι το δάνειο του, η διάγνωση καρκίνου σε ασθενείς ή η πρόβλεψη του καιρού (βροχή, ήλιος, χιόνι).

ΙΙ.ΠΑΡΑΜΕΤΡΙΚΗ ΕΝΑΝΤΙ ΜΗ ΠΑΡΑΜΕΤΡΙΚΗΣ ΜΕΘΟΔΟΥ

Η συγκεκριμένη διάκριση μεταξύ των μεθόδων αναφέρεται στη μορφή της σχέσης f, που υπάρχει μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Μιλώντας γενικά, σκοπός της χρήσης των μεθόδων στατιστικής μάθησης είναι να εκτιμήσουν, όσο το

δυνατόν καλύτερα την άγνωστη σχέση f, που υπάρχει μεταξύ των μεταβλητών αυτών (James & et al, 2013, pp. 21-23). Δηλαδή, στοχεύουν στη σκιαγράφηση μιας συνάρτησης $\hat{f}$, τέτοιας ώστε $y \approx \hat{f}$ (X), για κάθε παρατήρηση (x_{ij}, y_i) των δεδομένων εκπαίδευσης. Όπου το $\hat{f}$ αναφέρεται στην εκτίμηση της πραγματικής σχέσης f, η οποία προσπαθεί να προσδιοριστεί όσο το δυνατόν ακριβέστερα, μέσω των διαφόρων μεθόδων. Ωστόσο, η σχέση αυτή δεν μπορεί να προσδιοριστεί με απόλυτη ακρίβεια, καθώς υπάρχει πάντα ένα μέρος που δεν μπορεί να επεξηγηθεί από την $\hat{f}$. Για τον λόγο αυτό χρησιμοποιείται ο συμβολισμός *" ≈ "* και όχι αυτός της απόλυτης ισότητας. Επίσης το x_{ij}, αναφέρεται στην τιμή της *j* ανεξάρτητης μεταβλητής για την *i* παρατήρηση και το y_i στην τιμή της εξαρτημένης μεταβλητής για την κάθε *i* παρατήρηση. Έτσι το *X* στην παραπάνω σχέση, αναφέρεται στο σύνολο των τιμών των ανεξάρτητων μεταβλητών $x_1, \ldots, x_p$, που έχει την μορφή πίνακα, ενώ το *y* αναφέρεται στο σύνολο των τιμών της εξαρτημένης μεταβλητής, που έχει τη μορφή διανύσματος. Δεδομένου ότι η μορφή της σχέσης αυτής είναι άγνωστη, στόχος μας είναι να αξιολογήσουμε διαφορετικούς αλγορίθμους στατιστικής μάθησης, για να καταλήξουμε στην καλύτερη προσέγγιση για την υποκείμενη σχέση μεταξύ των μεταβλητών. Σε κάθε αλγόριθμο ωστόσο κάνουμε διαφορετικές υποθέσεις, σχετικά με την μορφή της σχέσης αυτής.

Στην περίπτωση που το πρόβλημα προσδιορισμού της $\hat{f}$ μπορεί να απλοποιηθεί εκτιμώντας ένα περιορισμένο πλήθος παραμέτρων $b_0, b_1, \ldots, b_p$, δηλαδή υπάρχει αναλυτική συναρτησιακή σχέση μεταξύ της *y* και των $x_1, \ldots, x_p$, η μέθοδος καλείται *παραμετρική* (Sheskin, 2011). Στα μοντέλα που προκύπτουν από αυτή τη μέθοδο, το πλήθος των παραμέτρων που απαιτείται να προσδιοριστεί, είναι ανεξάρτητο του πλήθους των δεδομένων που εισάγουμε σε αυτά. Παραδείγματα τέτοιων μοντέλων είναι τα ακόλουθα:

- Polynomial Regression
- Logistic Regression
- Naive Bayes
- Simple Neural Networks

Από την άλλη, οι μέθοδοι στις οποίες δεν γίνεται κάποια αυστηρή υπόθεση σχετικά με την μορφή της $\hat{f}$, καλούνται *μη παραμετρικές*. Η μορφή της σχέσης των $x_1, \ldots, x_p$ και της *y* προσδιορίζεται κάθε φόρα, βάσει των δεδομένων εκπαίδευσης που εισάγουμε σε αυτές. Συγκεκριμένα οι μέθοδοι αυτές προσπαθούν να κάνουν μια εκτίμηση της $\hat{f}$, όσο το δυνατόν πιο κοντά στην κάθε παρατήρηση (data points) των δεδομένων εκπαίδευσης. Την ίδια στιγμή όμως, προσπαθούν να διατηρήσουν την δυνατότητα η σχέση αυτή, να μπορεί να γενικευτεί για την εκτίμηση άλλων άγνωστων δεδομένων. Αυτές οι μέθοδοι είναι ιδιαίτερα χρήσιμες, όταν έχουμε να διαχειριστούμε μεγάλο πλήθος δεδομένων χωρίς πρότερη γνώση για αυτά και όταν στο

πρόβλημα που επιλύουμε δεν μας ενδιαφέρει έντονα, η αποκλειστική επιλογή μόνο των σωστών χαρακτηριστικών. Παραδείγματα μη παραμετρικών μοντέλων είναι τα ακόλουθα:

- Decision Trees
- k-Nearest Neighbors
- Support Vector Machines
- Artificial Neural Networks

ΙΙΙ.ΓΡΑΜΜΙΚΟ ΕΝΑΝΤΙ ΜΗ ΓΡΑΜΜΙΚΟΥ ΜΟΝΤΕΛΟΥ

Οι συγκεκριμένες δύο κατηγορίες μεθόδων αναφέρονται και πάλι στην σχεσιακή μορφή $\hat{f}$ που υπάρχει μεταξύ της εξαρτημένης μεταβλητής και των επεξηγηματικών μεταβλητών. Συγκεκριμένα, όταν υπάρχει γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και των παραμέτρων των ανεξάρτητων μεταβλητών, αλλά όχι απαραίτητα και με τις ίδιες τις ανεξάρτητες μεταβλητές, τότε το μοντέλο καλείται *γραμμικό* (Frost, 2017). Εναλλακτικά, ένα μοντέλο είναι γραμμικό όταν όλοι οι όροι που το απαρτίζουν, είναι είτε μια σταθερά είτε το γινόμενο μιας παραμέτρου και μιας επεξηγηματικής μεταβλητής. Παραδείγματα γραμμικών μοντέλων είναι τα ακόλουθα:

$$y \approx b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

$$y \approx b_0 + b_1 x_1 + b_2 x_1^2$$

Όπου το y αναφέρεται στην εξαρτημένη μεταβλητή, τα x_j στις επεξηγηματικές μεταβλητές και τέλος τα b_j στις παραμέτρους των τελευταίων.

Αντίθετα τα μοντέλα που δεν ικανοποιούν τα παραπάνω κριτήρια ονομάζονται *μη γραμμικά*, και μπορούν να πάρουν οποιονδήποτε σχηματισμό εκτός του γραμμικού. Το τι μοντέλο θα επιλέξουμε τελικά για την ανάλυσή μας, εξαρτάται από το αν απώτερος σκοπός μας είναι να προβλέψουμε με ακρίβεια το αποτέλεσμα της y ή να εξάγουμε κάποιο συμπέρασμα για την σχέση που υπάρχει μεταξύ της y και των x_j ή και τα δύο (James & et al, 2013, p. 20). Τα γραμμικά μοντέλα δίνουν συνήθως πιο απλά και πιο εύκολα ερμηνεύσιμα συμπεράσματα έναντι άλλων προσεγγίσεων, υστερώντας όμως στην ακρίβεια της πρόβλεψης τους. Αντιθέτως, τα μη γραμμικά μοντέλα παρέχουν πολύ πιο ακριβείς προβλέψεις για το αποτέλεσμα της εξαρτημένης μεταβλητής, με κόστος όμως στην ερμηνευτική τους ικανότητα, καθώς καθιστούν την εξαγωγή συμπεράσματος πιο απαιτητική διαδικασία.

1.3 ΘΕΩΡΗΤΙΚΗ ΑΠΟΤΥΠΩΣΗ ΜΕΘΟΔΩΝ ΤΑΞΙΝΟΜΗΣΗΣ

Στην συγκεκριμένη υποενότητα θα ασχοληθούμε με κάποιες εκ των μεθόδων ταξινόμησης, προκειμένου να κατανοήσουμε καλύτερα τον τρόπο με τον οποίο αυτές λειτουργούν. Συγκεκριμένα επιλέγουμε να μιλήσουμε για τις μεθόδους των αλγορίθμων Logistic Regression, Decision Trees και Random Forest, καθώς είναι αυτές που θα μας απασχολήσουν στην ανάλυση των δεδομένων που θα ακολουθήσει στα επόμενα κεφάλαια. Γενικά μιλώντας, οι μέθοδοι ταξινόμησης στοχεύουν στην πρόβλεψη μιας κατηγορικής εξαρτημένης μεταβλητής. Η διαδικασία της ταξινόμησης αναφέρεται στην ανάθεση κάθε παρατήρησης στην κατάλληλη κατηγορία ή κλάση. Πολύ συχνά οι μέθοδοι αυτές, προβλέπουν πρώτα την πιθανότητα της ποιοτικής μεταβλητής, να ανήκει σε καθεμία από αυτές τις κατηγορίες, ως βάση για να προχωρήσουν στην ταξινόμηση.

1.3.1 LOGISTIC REGRESSION

Η μέθοδος της λογιστικής παλινδρόμησης (Logistic Regression) εφαρμόζεται για να αναλύσει ένα σύνολο δεδομένων (dataset), στο οποίο υπάρχουν μια ή περισσότερες επεξηγηματικές μεταβλητές, που καθορίζουν ένα αποτέλεσμα (James & et al, 2013, pp. 130-137) (Hastie & et al, 2009). Το αποτέλεσμα προσδιορίζεται από μια διχοτομική μεταβλητή, η οποία μπορεί να λάβει δύο πιθανές τιμές της μορφής “Ναι” ή “Όχι”, “0” ή “1”, “επιτυχία” ή “αποτυχία”, κλπ. Η μέθοδος αυτή αντί να μοντελοποιήσει απευθείας το αποτέλεσμα της εξαρτημένης μεταβλητής *y*, μοντελοποιεί την πιθανότητα της *y* να ανήκει σε κάθε ξεχωριστή κλάση. Δεδομένου ότι οι πιθανότητες *p(X)* θέλουμε να κυμαίνονται μεταξύ 0 και 1, για οποιαδήποτε τιμή των επεξηγηματικών μεταβλητών *X*, χωρίς να αποκλίνουν από τον κανόνα αυτό, χρησιμοποιούμε την λογιστική συνάρτηση για να μοντελοποιήσουμε την *p(X)*. Έτσι έχουμε:

$$p(X) = \frac{e^{\,b_o + b_1 x_1 + \dots + b_p x_p}}{1 + e^{\,b_o + b_1 x_1 + \dots + b_p x_p}} \qquad (I)$$

όπου τα $b_o, b_1, \dots, b_p$ αναφέρονται στις παραμέτρους των p ανεξάρτητων μεταβλητών, που περιλαμβάνονται στο dataset που επιλύεται, ενώ τα $x_1, \dots, x_p$ αναφέρονται στις ίδιες τις ανεξάρτητες μεταβλητές που εξετάζονται σε αυτό. Η λογιστική παλινδρόμηση παράγει πάντα καμπύλη σιγμοειδούς μορφής, η οποία φράσσεται μεταξύ των τιμών 0 και 1, για κάθε τιμή των ανεξάρτητων μεταβλητών *X*. Πραγματοποιώντας κάποιους μετασχηματισμούς στην παραπάνω σχέση, η λογιστική παλινδρόμηση μπορεί να λάβει μορφή παρόμοια με αυτή της γραμμικής παλινδρόμησης. Αρχικά μετασχηματίζουμε την σχέση *(I)* γράφοντας την σε όρους λόγου αποδόσεων (odds ratio):

$$\frac{p(X)}{1-p(X)} = e^{\,b_o + b_1 x_1 + \ldots + b_p x_p} \qquad (II)$$

Το αριστερό μέρος της σχέσης *(II)* καλείται λόγος αποδόσεων, υπολογίζεται βάσει της πιθανότητας να συμβεί ένα γεγονός, έναντι της πιθανότητας να μην συμβεί αυτό και αναφέρεται στην σταθερή επίδραση που έχει κάθε μια από τις ανεξάρτητες μεταβλητές x_j, στην πιθανότητα πραγματοποίησης ενός γεγονότος. Τέλος, αν πάρουμε τον φυσικό λογάριθμο και στις δύο πλευρές της *(II)* η εξίσωση λαμβάνει την ακόλουθη μορφή:

$$ln\left(\frac{p(X)}{1-p(X)}\right) = b_o + b_1 x_1 + \ldots + b_p x_p \qquad (III)$$

Το αριστερό μέρος της σχέσης *(III)* καλείται *λογαριθμική απόδοση* (log-odds ή logit). Από την τελευταία σχέση βλέπουμε ότι η συνάρτηση logit, για το μοντέλο της λογιστικής παλινδρόμησης *(I)*, είναι γραμμική συνάρτηση των επεξηγηματικών μεταβλητών. Άρα, ο συντελεστής b_1 παραδείγματος χάρη, δείχνει πόσο θα μεταβληθούν τα log-odds, αν αυξηθεί η τιμή της x_1 κατά μια μονάδα. Δεδομένου ότι στην σχέση *(I)* δεν υπάρχει σχέση ευθείας γραμμής μεταξύ *p(X)* και x_1, δεν μπορούμε να πούμε ότι το b_1 αναφέρεται στις μεταβολές της *p(X),* από μια μοναδιαία αύξηση της x_1. Καθώς το ποσό κατά το οποίο θα μεταβληθεί η *p(X),* από την μοναδιαία αύξηση της x_1, εξαρτάται από την τρέχουσα τιμή της x_1. Ωστόσο το πρόσημο του συντελεστή b_1 δείχνει την συσχέτιση που υπάρχει μεταξύ της *p(X)* και της x_1. Τέλος είναι σημαντικό να αναφέρουμε ότι η εκτίμηση των παραμέτρων του υποδείγματος της λογιστικής παλινδρόμησης, γίνεται με τη μεθόδο μέγιστης πιθανοφάνειας, σύμφωνα με την οποία κάθε παρατήρηση της μεταβλητής στόχου, ταξινομείται στην κλάση με την μεγαλύτερη πιθανότητα.

1.3.2 DECISION TREES

Μια άλλη μέθοδος ταξινόμησης είναι αυτή των δέντρων απόφασης (Decision Trees). Τα decision trees μπορούν να εφαρμοστούν τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης, ανάλογα με το είδος της εξαρτημένης μεταβλητής που εξετάζουμε (Breiman & et al, 1984). Θα επικεντρωθούμε ωστόσο στους αλγορίθμους που αφορούν προβλήματα ταξινόμησης. Οι αλγόριθμοι αυτοί προσπαθούν να λύσουν το πρόβλημα, αναπαριστώντας το, με τη μορφή ενός δέντρου. Συγκεκριμένα «σπάνε» το αρχικό σύνολο των δεδομένων εκπαίδευσης, σε ολοένα και μικρότερα υποσύνολα, ενώ ταυτόχρονα αναπτύσσουν σταδιακά την μορφή του σχετικού δέντρου. Το δέντρο αναπτύσσεται από πάνω προς τα κάτω, έχοντας τις ρίζες του στην κορυφή. Η τελική μορφή του δέντρου αποτελείται από τους κόμβους απόφασης και από τα φύλλα ή τους τελικούς κόμβους. Κάθε κόμβος απόφασης αποτελεί μια συνθήκη βάσει της οποίας γίνεται ο διαχωρισμός των δεδομένων σε δύο ή περισσότερα κλαδιά

(αποφάσεις). Τα φύλλα, δηλαδή τα κλαδιά που δεν οδηγούν σε άλλους κόμβους απόφασης, αντιπροσωπεύουν την ταξινόμηση ή το αποτέλεσμα της τιμής της εξαρτημένης μεταβλητής. Το αποτέλεσμα σε κάθε τελικό κόμβο καθορίζεται από την επικρατούσα τιμή της μεταβλητής απόκρισης, των παρατηρήσεων που καταλήγουν σε αυτή την περιοχή. Το κριτήριο σύμφωνα με το οποίο γίνεται ο διαχωρισμός των δεδομένων σε κάθε κόμβο απόφασης, αξιολογείται κάθε φορά για κάθε χαρακτηριστικό (attribute), είτε βάσει του classification error rate, είτε του δείκτη Gini ή του Cross-Entropy που αναλύονται παρακάτω (Quinlan, 1986).

Δεδομένου ότι σε κάθε παρατήρηση, των δεδομένων εκπαίδευσης, ανατίθεται μια τιμή ανάλογα με το ποια είναι η επικρατούσα τιμή της μεταβλητής στόχου, στην περιοχή στην οποία ταξινομήθηκαν, το classification error rate είναι το ποσοστό των παρατηρήσεων αυτής της περιοχής, που δεν ανήκουν στην επικρατούσα τιμή της κλάση:

$$E = 1 - max_k(\hat{p}_{mk})$$

Όπου το $\hat{p}_{mk}$ αντιπροσωπεύει το ποσοστό των παρατηρήσεων των δεδομένων εκπαίδευσης, που ανήκουν στην *m* περιοχή και αντιστοιχούν στην *k* κλάση.

Ο δείκτης Gini υπολογίζεται βάσει της παρακάτω σχέσης:

$$G = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk})$$

Αποτελεί το μέτρο βάσει του οποίου υπολογίζεται η συνολική μεταβλητότητα μεταξύ των *K* κλάσεων, δηλαδή μετρά την «καθαρότητα» κάθε κόμβου. Όσο μικρότερη η τιμή του δείκτη τόσο πιο «καθαρός» θεωρείται ο κόμβος, καθώς αποτελείται κυρίως από παρατηρήσεις μιας κλάσης.

Τέλος ο δείκτης του cross-entropy υπολογίζεται σύμφωνα με την σχέση:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \, log \, \hat{p}_{mk}$$

Ο συγκεκριμένος δείκτης αποτελεί τον βαθμό ή το ποσό της αβεβαιότητας στην τυχαιότητα των στοιχείων ή με άλλα λόγια είναι το μέτρο της ακαθαρσίας. Άρα λοιπόν, ο δείκτης παρουσιάζει χαμηλές τιμές, όταν ο κόμβος *m* είναι καθαρός (James & et al, 2013, pp. 303-314).

1.3.3 RANDOM FOREST

Η συγκριμένη μέθοδος ταξινόμησης αποτελεί έναν αλγόριθμο συναθροίσεων, καθώς βασίζεται στην διαδικασία του bootstrap aggregation ή bagging (James & et al, 2013, pp. 316-319) (Liaw & Wiener, 2002). Η γενική ιδέα της διαδικασίας του bagging αναφέρεται στο συνδυασμό πολλών ίδιων ή διαφορετικών μοντέλων μάθησης, με σκοπό την μείωση της διακύμανσης των

τιμών της μεταβλητής στόχου, που προβλέπονται μέσω μιας στατιστικής μεθόδου. Όμως, για να επιτευχθεί η μείωση της διακύμανσης των αποτελεσμάτων της πρόβλεψης, πρέπει να αυξηθεί η ακρίβεια πρόβλεψης των αποτελεσμάτων της στατιστικής μεθόδου. Ένας φυσικός τρόπος για να αυξήσουμε την ακρίβεια της πρόβλεψης, είναι με το να συγκεντρώσουμε πολλά διαφορετικά δείγματα B από τον πληθυσμό, τα οποία θα χρησιμοποιηθούν για την εκπαίδευση πολλών ξεχωριστών μοντέλων, όπου κάθε ένα από αυτά, θα μας δίνει μια ξεχωριστή πρόβλεψη $\hat{f}(X)$, των τιμών της εξαρτημένης μεταβλητής. Έτσι με τον υπολογισμό του μέσου όρου των προβλέψεων $\hat{f}^1(X)$, $\hat{f}^2(X)$, …, $\hat{f}^B(X)$, που προκύπτουν από κάθε δείγμα, καταφέρνουμε να εξασφαλίσουμε ένα μοναδικό μοντέλο χαμηλής διακύμανσης το οποίο προκύπτει βάσει της σχέσης:

$$\hat{f}avg(X) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(X)$$

όπου το $\hat{f}avg(X)$ αναφέρεται στο μέσο όρο των προβλέψεων που προκύπτουν από τα ξεχωριστά μοντέλα που εκπαιδεύονται από τα B διαφορετικά δείγματα του πληθυσμού και το $\hat{f}^b(X)$ εκφράζει τα αποτελέσματα της πρόβλεψης που προκύπτει από το καθένα από τα δείγματα αυτά.

Δεδομένου όμως, ότι η πρόσβαση σε πολλά διαφορετικά δείγματα του πληθυσμού είναι περιορισμένη, για το σκοπό αυτό λαμβάνουμε πολλά επαναλαμβανόμενα τυχαία δείγματα από τα αρχικά δεδομένα εκπαίδευσης, δημιουργώντας έτσι B διαφορετικά σύνολα δεδομένων εκπαίδευσης. Έπειτα, εκπαιδεύουμε την μέθοδο μας σε καθένα από τα υποσύνολα δεδομένων εκπαίδευσης b, ώστε να λάβουμε την πρόβλεψη $\hat{f}^{*b}(X)$ καθενός από αυτά και τελικά να υπολογίσουμε τον μέσο όρο όλων των προβλέψεων σύμφωνα με την παρακάτω σχέση:

$$\hat{f}bag(X) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(X)$$

όπου το $\hat{f}bag(X)$ αναφέρεται στο μέσο όρο των προβλέψεων που προκύπτουν από τα ξεχωριστά μοντέλα που εκπαιδεύονται από τα δείγματα B, που εξασφαλίζονται από το αρχικό σύνολο δεδομένων εκπαίδευσης. Η συγκεκριμένη διαδικασία καλείται *bagging*. Βάσει αυτής, η πρόβλεψη του αποτελέσματος της εξαρτημένης μεταβλητής για μια παρατήρηση του test set, προκύπτει από την καταγραφή της κλάσης που προβλέπεται από καθένα από τα B δέντρα που εκπαιδεύονται και από τον υπολογισμό του αποτελέσματος της πλειοψηφίας των ψήφων (majority vote) (Breiman L. , 2001). Αυτό σημαίνει ότι η τελική πρόβλεψη του αποτελέσματος είναι η συνηθέστερη κλάση που προκύπτει από τις B προβλέψεις που πραγματοποιούνται. Το πλήθος των δέντρων B δεν αποτελεί κρίσιμη παράμετρο στο bagging, καθώς η χρήση πολύ

μεγάλου αριθμού δέντρων, δεν θα οδηγήσει σε overfitting. Στην πράξη ωστόσο επιθυμούμε την χρήση ενός ικανοποιητικού αριθμού B δέντρων, ώστε να εξασφαλίσουμε την μείωση του σφάλματος της πρόβλεψης μας.

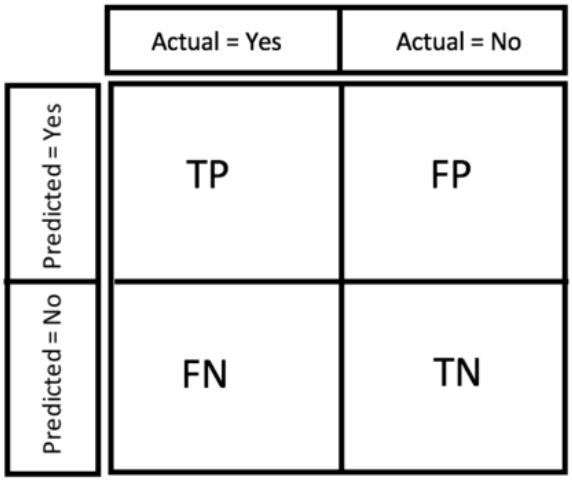
Έτσι η Random Forest που βασίζεται στην μέθοδο του bagging, δημιουργεί πολλαπλά, παράλληλα δέντρα απόφασης, τα οποία προκύπτουν αντίστοιχα από την τυχαία επιλογή, υποσυνόλων δεδομένων του training set. Στην συνέχεια συνενώνει τα δέντρα αυτά, συναθροίζοντας τις ψήφους που προκύπτουν από την πρόβλεψη της κλάσης, του κάθε ξεχωριστού δέντρου, προκειμένου να αποφασίσει για την τελική κλάση του αντικειμένου που εξετάζεται. Ταυτόχρονα στο συγκεκριμένο μοντέλο, κατά την δημιουργία των δέντρων, υπάρχει ένας ακόμη παράγοντας τυχαιότητα. Αντί να αναζητά το πιο σημαντικό χαρακτηριστικό βάσει του οποίου θα γίνει ο διαχωρισμός των δεδομένων (split) σε κάθε κόμβο, εξαναγκάζει την επιλογή του καλύτερου χαρακτηριστικού για τον διαχωρισμό των δεδομένων, μέσα από ένα περιορισμένο υποσύνολο των ανεξάρτητων μεταβλητών. Το υποσύνολο αυτό αλλάζει σε κάθε στάδιο που επιλέγεται το κριτήριο βάσει του οποίο θα γίνεται ο διαχωρισμό των δεδομένων κάθε κόμβου απόφασης. Συνήθως το πλήθος των ανεξάρτητων μεταβλητών που ελέγχεται σε κάθε split ισούται με $m= \sqrt{p}$, όπου p είναι το σύνολο των ανεξάρτητων μεταβλητών που υπάρχουν στο σύνολο των δεδομένων που εξετάζεται. Με αυτόν τον τρόπο, επιτυγχάνεται μεγαλύτερη ανεξαρτησία και διαφορετικότητα κατά την επιλογή της ανεξάρτητης μεταβλητής, που θα χρησιμοποιηθεί σε κάθε split, καθώς οι ισχυρές μεταβλητές μπορεί να μην ληφθούν καν υπόψη. Έτσι οδηγούμαστε στην δημιουργία ενός καλύτερου μοντέλου, ικανού να δώσει πιο ακριβή και ευσταθή πρόβλεψη.

1.4 ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ ΥΠΟΔΕΙΓΜΑΤΟΣ

Μείζονος σημασίας για κάθε πρόβλημα, είναι η αξιολόγηση του τελικού υποδείγματος (μοντέλου) ταξινόμησης, που προέκυψε από τα δεδομένα εκπαίδευσης (Mishra, 2018). Συγκεκριμένα το επόμενο βήμα μετά την εκπαίδευση του μοντέλου, είναι να εξετάσουμε πόσο αποτελεσματικό είναι κατά την χρήση του σε ένα άγνωστο σύνολο δεδομένων (test set). Οι αλγόριθμοι ταξινόμησης μπορούν να χρησιμοποιηθούν για την επίλυση πολλών διαφορετικών προβλημάτων. Για τον λόγο αυτό, υπάρχουν πολλά διαφορετικά μέτρα, βάσει των οποίων αξιολογούμε την επίδοση του κάθε υποδείγματος, τα οποία παρατίθενται παρακάτω. Το ποιο από αυτά είναι καταλληλότερο, εξαρτάται κάθε φορά από την φύση του προβλήματος που εξετάζεται.

Μπορούμε να πάρουμε μια πρώτη εικόνα για την επίδοση του μοντέλου μας, στα άγνωστα δεδομένα, μέσω του πίνακα του **Confusion Matrix.** Κάθε γραμμή του πίνακα αντιπροσωπεύει το πλήθος των παρατηρήσεων που προβλέφθηκαν από το μοντέλο σε κάθε κλάση, ενώ κάθε

στήλη αντιπροσωπεύει το πλήθος των παρατηρήσεων που προβλέφθηκαν σε κάθε πραγματική κλάση, όπως φαίνεται και στην απεικόνιση παρακάτω.



Εικόνα 2. Confusion Matrix

Ο πίνακας του Confusion Matrix (Nisbet & et al, 2017) από μόνος του, δεν αποτελεί μέτρο αξιολόγησης της απόδοσης ενός μοντέλου. Ωστόσο, μέσα σε αυτόν περικλείονται έννοιες, ιδιαίτερα χρήσιμες για τον προσδιορισμό της πλειοψηφίας των μέτρων αξιολόγησης. Έτσι, πριν προχωρήσουμε στον προσδιορισμό των μέτρων αυτών, είναι αναγκαίος ο προσδιορισμός των συμβολισμών της εικόνας 2 και ο ορισμός των εννοιών που αφορούν τους συμβολισμούς αυτούς, οι οποίοι παρατίθενται παρακάτω.

True Positive (TP): αναφέρεται στο σύνολο των παρατηρήσεων όπου η πραγματική κλάση τους ήταν "Ναι" ή "1" και το υπόδειγμα της προέβλεψε επίσης ως "Ναι" ή "1".

True Negatives (TN): αναφέρεται στο σύνολο των παρατηρήσεων όπου η πραγματική κλάση τους ήταν "Όχι" ή "0" και το υπόδειγμα της προέβλεψε επίσης ως "Όχι" ή "0".

False Positives (FP): αναφέρεται στο σύνολο των παρατηρήσεων όπου η πραγματική κλάση τους ήταν "Όχι" ή "0" και το υπόδειγμα της προέβλεψε ως "Ναι" ή "1".

False Negatives (FN): αναφέρεται στο σύνολο των παρατηρήσεων όπου η πραγματική κλάση τους ήταν "Ναι" ή "1" και το υπόδειγμα της προέβλεψε ως "Όχι" ή "0".

Τα στοιχεία της κύριας διαγωνίου του πίνακα (TP και TN) αντιπροσωπεύουν το σύνολο των σωστών ταξινομήσεων του υποδείγματος. Η περίπτωση FN που αναφέρεται στην εσφαλμένη ταξινόμηση μιας θετικής παρατήρησης ως αρνητική, αποτελεί το Σφάλμα Τύπου Ι ενώ η αντίθετη περίπτωση FP αποτελεί το Σφάλμα Τύπου ΙΙ.

Ακρίβεια (accuracy): είναι το ποσοστό των σωστών προβλέψεων που πραγματοποίησε το μοντέλο στο σύνολο των προβλέψεων που διενενέργησε, το οποίο υπολογίζεται βάσει της ακόλουθης σχέσης.

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

Ορθότητα (precision): αναφέρεται στην ακρίβεια του μοντέλου στην θετική πρόβλεψη. Είναι το ποσοστό των παρατηρήσεων που πράγματι ανήκουν στην κλάση "Ναι" ή "1" έναντι του συνόλου αυτών που προβλέφθηκαν από το μοντέλο ως "Ναι" ή "1". Δείχνει δηλαδή από το σύνολο των παρατηρήσεων που ταξινόμησε το μοντέλο στην κλάση 1, το ποσοστό των περιπτώσεων που αξιολόγησε σωστά.

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

Ανάκληση ή ευαισθησία (recall ή sensitivity): αναφέρεται στην ικανότητα του μοντέλου να εντοπίσει τις παρατηρήσεις της θετικής κλάσης. Δηλαδή μετρά από το σύνολο των δεδομένων της θετική κλάσης, πόσες προβλέφθηκαν σωστά.

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

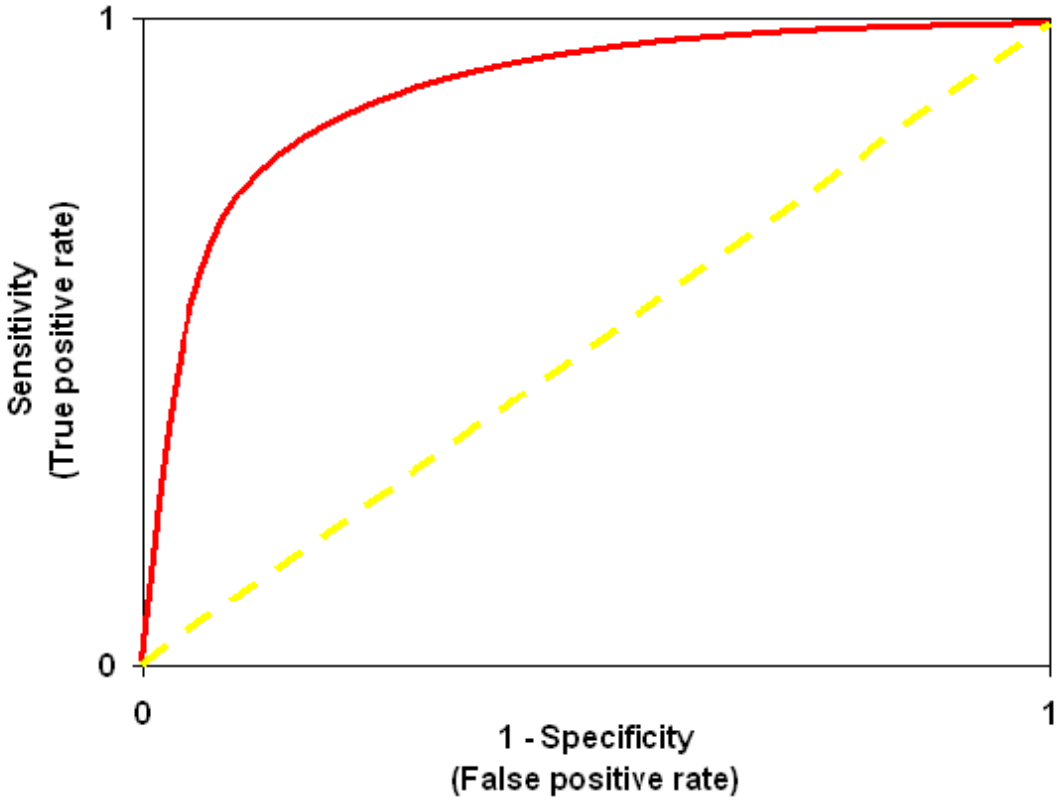
Εξειδίκευση (specificity): αποτελεί το αντίθετο της ανάκλησης, καθώς αναφέρεται στην ικανότητα του μοντέλου να εντοπίσει τις παρατηρήσεις της αρνητικής κλάσης. Δηλαδή μετρά τι ποσοστό των παρατηρήσεων που πράγματι ανήκουν στην αρνητική κλάση, προβλέπεται σωστά από το μοντέλο.

$$Specificity = \frac{TN}{TN+FP}$$

F1 – Score: αποτελεί τον σταθμισμένο μέσο (αρμονικό μέσο) των μέτρων της ορθότητας και της ανάκλησης. Τα δύο αυτά μέτρα παρέχουν μια αλληλοσυμπληρούμενη εικόνα της αποτελεσματικότητας του μοντέλου, τα οποία με το *F1 – Score* συνδυάζονται. Συνήθως είναι πιο χρήσιμο μέτρο από αυτό της ακρίβειας, ειδικά αν υπάρχει άνιση εκπροσώπηση κάποιας κλάσης.

$$F1\text{-}Score = \frac{2 * precision * recall}{precision + recall}$$

Το εμβαδό κάτω από την καμπύλη ROC (AUC - Area under the ROC Curve): αποτελεί ένα από τα πιο ευρέως χρησιμοποιούμενα μέτρα αξιολόγησης. Συγκεκριμένα η καμπύλη ROC (Receiver Operating Characteristic curve) αποτελεί την γραφική αναπαράσταση της επίδοση ενός μοντέλου ταξινόμησης, για όλες τις πιθανές τιμές των συνόρων απόφασης (Fawcett, 2006) (Huang & Ling, 2005). Όπου, σύνορο απόφασης είναι το «κατώφλι» που θέτουμε στο μοντέλο, προκειμένου να ταξινομήσει την κάθε παρατήρηση στην εκάστοτε κλάση. Συγκεκριμένα αν η πιθανότητα της παρατήρησης είναι πάνω από το «κατώφλι» που έχουμε θέσει, τότε το μοντέλο ταξινομεί την παρατήρηση στην κλάση 1 αλλιώς στην κλάση 0. Ταυτόχρονα, μέσω της καμπύλης ROC εκφράζεται η ευαισθησία (sensitivity) του ταξινομητή, καθώς δείχνει πως μεταβάλλεται η αναλογία των True Positives, σε σχέση με την αναλογία των False Positives. Τα μέτρα αυτά αναπαρίστανται στους άξονες y και x αντίστοιχα, του γραφήματος της εικόνας 3. Μετά την εκπαίδευση του μοντέλου, επιθυμούμε η κόκκινη γραμμή του γραφήματος να βρίσκεται, όσο το δυνατόν πιο κοντά στην πάνω δεξιά γωνία του σχήματος, καθώς αυτό υποδηλώνει καλύτερη και πιο αξιόπιστη πρόβλεψη των αποτελεσμάτων που δίνει ο ταξινομητής μας. Αντίθετα, όσο πιο κοντά βρίσκεται στην διακεκομμένη γραμμή, τόσο μεγαλύτερη η τυχαιότητα στην οποία στηρίζεται ο ταξινομητής μας. Τα σημεία που βρίσκονται πάνω στην κόκκινη καμπύλη, αντιπροσωπεύουν τον συμβιβασμό (trade-off) μεταξύ του ποσοστού των True Positives και των False Positives. Έτσι μέσω του AUC μετράται το εμβαδό της δυσδιάστατης περιοχής, που βρίσκεται κάτω από την καμπύλη ROC. Σε όρους ολοκληρωματικού λογισμού, το AUC αποτελεί το ολοκλήρωμα της καμπύλης ROC για τιμές από (0, 0) έως (1, 1). Επομένως, και οι τιμές που μπορεί να λάβει κυμαίνονται μεταξύ 0 και 1. Στόχος μας ωστόσο είναι, το μοντέλο που εκπαιδεύεται να δίνει τιμές για το AUC όσο το δυνατόν πιο κοντά στο 1, καθώς αυτό υποδηλώνει καλύτερες επιδόσεις για το μοντέλο μας (Hossin & Sulaiman, 2015).



Εικόνα 3. Καμπύλη ROC

2. ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

Στο παρόν κεφάλαιο θα προχωρήσουμε στην αναλυτική περιγραφή του προβλήματος, που στοχεύουμε να μελετήσουμε και στην περιγραφή των μεταβλητών που περιλαμβάνονται στην βάση δεδομένων μας. Ακόμη θα παρατεθούν στοιχεία περιγραφικής στατιστικής, ώστε να λάβουμε μια πιο συγκεκριμένη εικόνα σχετικά με το πρόβλημα και τα ιδιαίτερα χαρακτηριστικά των μεταβλητών ενδιαφέροντος.

2.1 ΠΕΡΙΓΡΑΦΗ ΠΡΟΒΛΗΜΑΤΟΣ

Το πρόβλημα που μελετάμε αφορά σε μια μεγάλη πολυεθνική εταιρία, η οποία επιθυμεί να κατανοήσει τους λόγους για τους οποίους φεύγουν πρόωρα οι καλύτεροι ή οι πιο έμπειροι εργαζόμενοι της. Επιπλέον, η εταιρία έχει σαν στόχο να προβλέψει ποιοι από τους πιο ικανούς εργαζομένους της, είναι αυτοί που έχουν μεγαλύτερη πιθανότητα να αποχωρήσουν από την εταιρία. Για τον σκοπό αυτό, έχουν συγκεντρωθεί δεδομένα, που αφορούν σε διάφορα χαρακτηριστικά των υπαλλήλων και τα οποία θα μπορούσε να έχει στην διάθεση του και να αναλύσει, οποιοδήποτε τμήμα Ανθρώπινου Δυναμικού, για τους υπαλλήλους του. Συγκεκριμένα, τα δεδομένα που μελετούμε στην παρούσα εργασία συλλέχθηκαν από την ανοιχτή βάση δεδομένων, που παρέχει ο ιστότοπος του kaggle.com. Στο συγκεκριμένο data set περιλαμβάνονται δεδομένα περίπου 30.000 εργαζομένων, για τους οποίους έχουμε πληροφορίες που σχετίζονται με 10 χαρακτηριστικά τους, μέσα στα οποία παρέχεται και η πληροφορία, σχετικά με το αν έχει αποχωρήσει ο υπάλληλος από την εταιρία ή όχι. Τα χαρακτηριστικά αυτά έχουν επεξεργαστεί και ομαδοποιηθεί, προκειμένου να απομακρυνθούν στοιχεία που αφορούν προσωπικά δεδομένα των εργαζομένων, όπως ο μισθός τους. Έτσι λοιπόν, το πρόβλημα που έχουμε να αντιμετωπίσουμε είναι ένα κλασσικό πρόβλημα ταξινόμησης (classification), η μελέτη του οποίου, θα μας δώσει την δυνατότητα να προσδιορίσουμε τους καθοριστικούς παράγοντες της εργασιακής φθοράς.

2.2 ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ

Το σύνολο των δεδομένων που χρησιμοποιείται, προκειμένου να προβλέψουμε τους λόγους, που επηρεάζουν έναν εργαζόμενο στην απόφαση του να αποχωρήσει από την εταιρία, αποτελείται από 29.998 στιγμιότυπα (instances) και 10 μεταβλητές χαρακτηριστικών (attributes). Στην εικόνα 4 βλέπουμε τις 10 πρώτες εγγραφές του dataset «employees» που κατασκευάσαμε, για να αποκτήσουμε μια πρώτη αντίληψη, σχετικά με το είδος των δεδομένων που περιλαμβάνει η βάση μας.

	satisfaction_ level	last_evaluati on_rating	projects_ worked_on	average_ montly_ hours	time_spend _company	work_ accident	promotion_ last_5years	department	salary	Attrition
1	3.8	5.3	3	167	3	0	0	sales	low	1
2	8.0	8.6	6	272	6	0	0	sales	medium	1
3	1.1	8.8	8	282	4	0	0	sales	medium	1
4	3.7	5.2	3	169	3	0	0	sales	low	1
5	4.1	5.0	3	163	3	0	0	sales	low	1
6	1.0	7.7	7	257	4	0	0	sales	low	1
7	9.2	8.5	6	269	5	0	0	sales	low	1
8	8.9	10.0	6	234	5	0	0	sales	low	1
9	4.2	5.3	3	152	3	0	0	sales	low	1
10	1.1	8.1	7	315	4	0	0	sales	low	1

Εικόνα 4. Σύνολο Δεδομένων

Όπως φαίνεται στην εικόνα 4 και στον πίνακα 1 που ακολουθεί, οι μεταβλητές που περιλαμβάνονται στο dataset αφορούν σε χαρακτηριστικά των εργαζομένων όπως: το επίπεδο ικανοποίησης τους, η επίδοση τους όπως καταγράφηκε βάσει της τελευταίας αξιολόγησης, ο αριθμός των έργων στα οποία απασχολούνται, οι μέσες μηνιαίες ώρες εργασίας τους, τα έτη απασχόλησης τους στην εταιρία, το αν υπήρξε ή όχι κάποιο εργατικό ατύχημα κατά την απασχόλησή τους στην εταιρία, το αν έλαβαν ή όχι κάποια προαγωγή τα τελευταία 5 χρόνια, το τμήμα στο οποίο απασχολούνται, το μισθολογικό τους επίπεδο και τέλος το αν υπήρξε εργασιακή φθορά ή όχι.

Μεταβλητές εισόδου	Περιγραφή
satisfaction_level	επίπεδο ικανοποίηση (κλίμακα τιμών 0-10)
last_evaluation_rating	βαθμός τελευταίας αξιολόγησης (κλίμακα τιμών 0-10)
projects_worked_on	πλήθος έργων ενασχόλησης
average_montly_hours	μέσες μηνιαίες ώρες εργασίας
time_spend_company	χρόνια απασχόλησης στην εταιρία
work_accident	υπήρξε κάποιο εργατικό ατύχημα; ("0", "1")
promotion_last_5years	υπήρξε προαγωγή τα τελευταία 5 χρόνια; ("0", "1")
department	τμήμα απασχόλησης ("accounting", "hr", "sales", "technical", "support", "IT", "management", "product_mng", "marketing", "RandD")
salary	κλίμακα μισθού ("low", "medium", "high")
attrition	υπήρξε εργασιακή φθορά; ("0", "1")

Πίνακας 1. Περιγραφή Δεδομένων

Η μεταβλητή satisfaction_level που αναφέρεται στο επίπεδο ικανοποίησης των εργαζομένων και η μεταβλητή last_evaluation_rating που αναφέρεται στο βαθμό της τελευταίας αξιολόγησης τους, λαμβάνουν πραγματικές τιμές στην κλίμακα 0 έως 10. Όπου το 0 αντιπροσωπεύει πολύ χαμηλά επίπεδα του δείκτη που εξετάζεται κάθε φορά, ενώ το 10 αντιπροσωπεύει το υψηλότερο επίπεδο που μπορεί να λάβει ο κάθε δείκτης. Έπειτα, οι μεταβλητές projects_worked_on, average_monthly_hours και time_spend_company περιλαμβάνουν πληροφορίες σχετικά με τον

αριθμό των project που αναλαμβάνει ο εργαζόμενος, τις μέσες μηνιαίες ώρες εργασίας του και τα χρόνια απασχόλησης του στην εταιρία, αντίστοιχα. Αυτές λαμβάνουν ακέραιες τιμές, για να μας δώσουν στοιχεία σχετικά με τα χαρακτηριστικά που καλύπτουν. Βλέπουμε λοιπόν από τον πίνακα 1, ότι οι πρώτες πέντε μεταβλητές, του συνόλου δεδομένων μας είναι αριθμητικές.

Αντίθετα, οι υπόλοιπες πέντε που περιλαμβάνονται σε αυτό, αφορούν σε κατηγορικές μεταβλητές. Συγκεκριμένα, οι μεταβλητές work_accident, promotion_last_5years και attrition λαμβάνουν τιμές της μορφής "1" που σημαίνει "Ναι" ή "0" που σημαίνει "Όχι", προκειμένου να απαντήσουν στα ερωτήματα, σχετικά με το αν υπήρξε εργατικό ατύχημα, ή προαγωγή τα τελευταία 5 χρόνια ή εργασιακή φθορά αντίστοιχα. Επίσης, στον πίνακα 1 βλέπουμε ότι η μεταβλητή department παίρνει 10 διαφορετικές τιμές, καθώς αποτελείται από 10 επίπεδα (levels), όσα και τα διαφορετικά τμήματα που υπάρχουν στην εταιρία. Αντίστοιχα, η μεταβλητή salary μπορεί να λάβει τρείς διαφορετικές τιμές, καθώς αποτελείται από 3 επίπεδα, που εκφράζουν τα διαφορετικά μισθολογικά επίπεδα των εργαζομένων.

2.3 ΣΤΟΙΧΕΙΑ ΠΕΡΙΓΡΑΦΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Πριν προχωρήσουμε στην περιγραφική στατιστική και στην ανάλυση των δεδομένων μας, απαιτείται να προηγηθεί η προ-επεξεργασία των δεδομένων της βάσης μας. Σε αυτό το στάδιο απαιτείται να εξεταστεί αν λείπουν ή όχι τιμές (missing values) από τις μεταβλητές του data set και να εξασφαλιστεί η ύπαρξη ομοιόμορφης δομής στις τιμές των μεταβλητών, δηλαδή οι τιμές που λαμβάνει κάθε μεταβλητή, να απαντούν στον ίδιο τύπο δεδομένων. Ύστερα από έλεγχο που πραγματοποιήθηκε στο data set των employees, διαπιστώθηκε ότι δεν υπάρχουν missing values ή μη ενδεδειγμένες τιμές στα στοιχεία των μεταβλητών μας, οπότε μπορούμε να προχωρήσουμε χωρίς κάποια περαιτέρω ενέργεια στην ανάλυση τους.

Στην εικόνα 5 για τις αριθμητικές μεταβλητές, καταγράφονται συνοπτικά στοιχεία περιγραφικής στατιστικής. Συγκεκριμένα επισημαίνεται η μέγιστη (max) και η ελάχιστη τιμή τους (min), ο μέσος (mean) και η διάμεσος (median) τους καθώς και οι τιμές του πρώτου (1st Qu.) και τρίτου (3rd Qu.) τεταρτημορίου τους. Όσον αφορά τις κατηγορικές μεταβλητές η σύνοψη των δεδομένων, μας δείχνει το πλήθος των παρατηρήσεων που κατανέμονται, σε καθένα από τα levels που λαμβάνουν οι μεταβλητές αυτές. Βάσει της εικόνας 5, διαπιστώνουμε ότι το μέσο επίπεδο ικανοποίησης αντιστοιχεί σε 6,1 ενώ μόλις στο 25% των εργαζομένων σημειώνονται επίπεδα ικανοποίησης άνω του 8,2. Ο μέσος βαθμός αξιολόγησης των υπαλλήλων αντιστοιχεί στο 7,2 ενώ η ελάχιστη τιμή που καταγράφεται στο συγκεκριμένο δείκτη είναι η 3,6. Επιπλέον, μπορεί να σημειωθεί ότι το σύνολο των υπαλλήλων συμμετέχει τουλάχιστον σε 2 project, ενώ ο μέσος όρος συμμετοχής τους είναι τα 4 με 5. Όσον αφορά τον διάμεσο των ωρών εργασίας, βλέπουμε ότι ισούται με 205 ώρες, αριθμός που αντικατοπτρίζει

τουλάχιστον 45 ώρες υπερωριών μηνιαίως, για το 50% των εργαζομένων. Ακόμη παρατηρούμε ότι το σύνολο των εργαζομένων παραμένει τουλάχιστον 2 χρόνια στην εταιρία, με το 50% αυτών να παραμένει περίπου 3 χρόνια, ενώ μόλις το 25% του συνόλου εργάζεται σε αυτή πάνω από 4 χρόνια. Στην συγκεκριμένη εταιρία παρατηρούμε επίσης ότι το 14,5% είχε κάποιο εργατικό ατύχημα, ενώ μόλις το 2% έλαβε κάποια προαγωγή τα τελευταία 5 χρόνια, ποσοστό που είναι ιδιαίτερα χαμηλό. Τέλος, από τα δεδομένα διαφαίνεται ότι το ποσοστό των εργαζομένων που παραιτήθηκαν από την εταιρία αγγίζει περίπου το 24%.

```
satisfaction_level last_evaluation_rating projects_worked_on average_montly_hours time_spend_company
Min.   : 0.900     Min.   : 3.600         Min.   :2.000      Min.   : 96.0        Min.   : 2.000
1st Qu.: 4.400     1st Qu.: 5.600         1st Qu.:3.000      1st Qu.:161.0        1st Qu.: 3.000
Median : 6.400     Median : 7.200         Median :4.000      Median :205.0        Median : 3.000
Mean   : 6.128     Mean   : 7.161         Mean   :4.303      Mean   :206.1        Mean   : 3.498
3rd Qu.: 8.200     3rd Qu.: 8.700         3rd Qu.:5.000      3rd Qu.:250.0        3rd Qu.: 4.000
Max.   :10.000     Max.   :10.000         Max.   :8.000      Max.   :320.0        Max.   :10.000

Work_accident promotion_last_5years       Department      salary      Attrition
0:25660       0:29360               sales      :8280   high  : 2474   0:22856
1: 4338       1:  638               technical  :5440   low   :14632   1: 7142
                                    support    :4458   medium:12892
                                    IT         :2454
                                    product_mng:1804
                                    marketing  :1716
                                    (Other)    :5846
```

Εικόνα 5. Σύνοψη Περιγραφικων Στοιχείων Μεταβλητών

Επιπλέον, για τις αριθμητικές μεταβλητές παρατηρούμε ότι η μέση τιμή και η διάμεσος παρουσιάζουν μικρές αποκλίσεις μεταξύ τους. Αυτό φανερώνει ότι στο δείγμα δεν παρουσιάζονται ακραίες τιμές ή έντονη ασυμμετρία. Αυτό επιβεβαιώνεται και από τις χαμηλές τυπικές αποκλίσεις, σχεδόν όλων των μεταβλητών, που παρατίθενται στην εικόνα 6, πράγμα που σημαίνει ότι οι τιμές τείνουν να βρίσκονται κοντά στον μέσο. Μόνη εξαίρεση αποτελεί η τυπική απόκλιση των μέσων μηνιαίων ωρών εργασίας, η οποία ισούται με 50,1 ώρες, γεγονός που υποδηλώνει ότι οι τιμές είναι περισσότερο διεσπαρμένες γύρω από τον μέσο.

```
satisfaction_level last_evaluation_rating     projects_worked_on
          2.486265                1.711663               1.330128
average_montly_hours     time_spend_company
          50.191940                1.460112
```

Εικόνα 6. Τυπική Απόκλιση Μεταβλητών

Στο επόμενο μέρος της ανάλυσης μας, προχωρούμε στον έλεγχο της κατανομής συχνοτήτων, των τιμών που λαμβάνουν οι συνεχείς μεταβλητές του dataset, καθώς και στην γραφική απεικόνιση της κατανομής συχνοτήτων τους, μέσω της κατασκευής των ιστογραμμάτων τους. Με τον τρόπο αυτό στοχεύουμε στο να αποκτήσουμε μια καλύτερη εικόνα, σχετικά με το πως διασπείρονται οι τιμές σε κάθε κλάση, δηλαδή στα υπό-διαστήματα που κατασκευάζουμε για να ομαδοποιήσουμε τις τιμές των παρατηρήσεων, κάθε αριθμητικής μεταβλητής. Από την άλλη, για τις κατηγορικές μεταβλητές προχωρούμε στην κατασκευή των ραβδογραμμάτων τους, πάνω

στα οποία επισημαίνεται και το πλήθος των παρατηρήσεων, που ανήκουν σε κάθε μια από τις κατηγορίες αυτών.

Παρακάτω στην εικόνα 7 παρουσιάζεται το πλήθος των εργαζομένων που αντιστοιχεί σε κάθε επίπεδο ικανοποίησης. Αντίστοιχα, στον πίνακα 2 παρατίθεται η κατανομή των παρατηρήσεων σε καθένα από τα επίπεδα αυτά. Παρατηρούμε ότι σημειώνονται παρατηρήσεις σε όλη την έκταση της κλίμακας τιμών, με τις περισσότερες να καταγράφονται μεταξύ των επιπέδων ικανοποίησης του διαστήματος (7, 8] και (8, 9].

(0,1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]	(9, 10]
1106	1850	926	2366	3376	3834	3944	4592	4436	3568

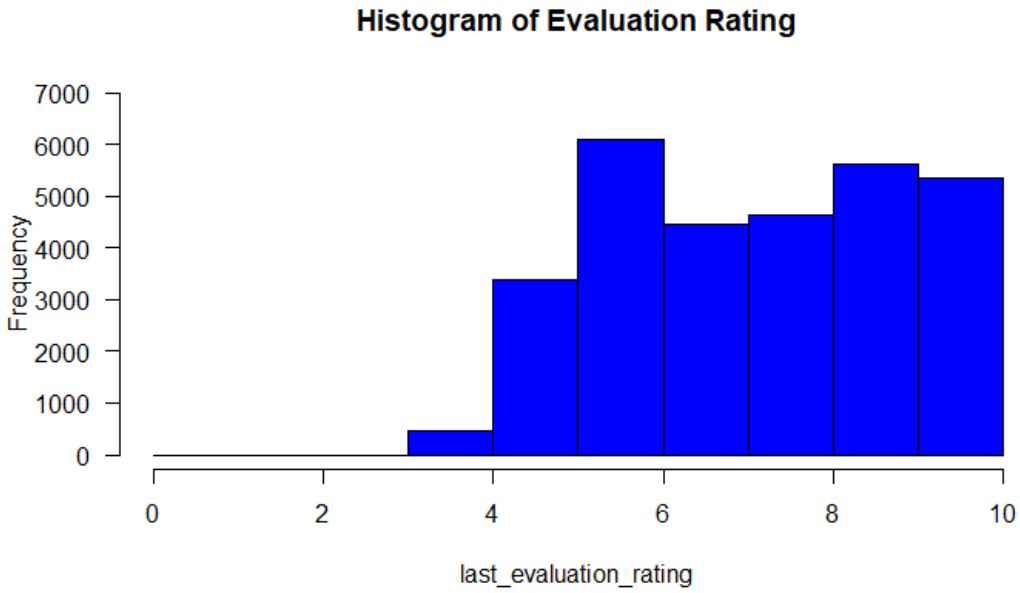
Πίνακας 2. Κατανομή Επιπέδων Ικανοποίησης

Εικόνα 7. Ιστόγραμμα Επιπέδου Ικανοποίησης

Ακολούθως, στον πίνακα 3 και στην εικόνα 8 παρατίθενται αντίστοιχα ο πίνακας και το ιστόγραμμα συχνοτήτων, του βαθμού της τελευταίας αξιολόγησης των εργαζομένων. Βάσει αυτών, διαπιστώνουμε ότι η μεταβλητή αυτή λαμβάνει τιμές μεγαλύτερες του 3, ενώ το μεγαλύτερο πλήθος παρατηρήσεων είναι συγκεντρωμένο στο διάστημα (5, 6].

(0,1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]	(9, 10]
0	0	0	472	3370	6084	4468	4626	5628	5350

Πίνακας 3. Κατανομή Βαθμών Αξιολόγησης

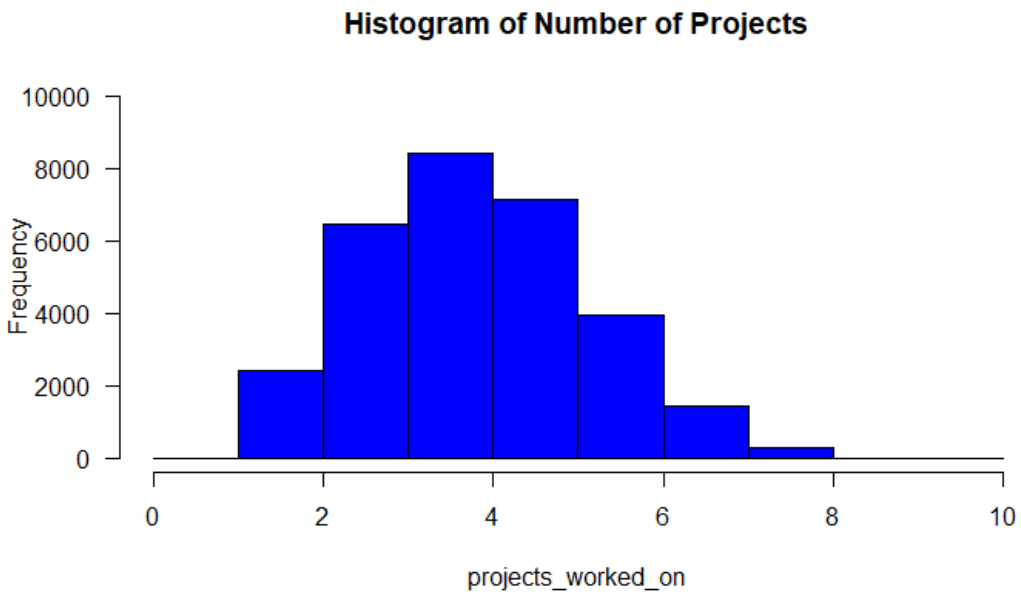


Εικόνα 8. Ιστόγραμμα Βαθμού Αξιολόγησης

Στην συνέχεια απεικονίζονται στοιχεία που αφορούν στην κατανομή της μεταβλητής που σχετίζεται με τον αριθμό των έργων, στα οποία απασχολείται ένας εργαζόμενος. Βλέπουμε λοιπόν στον πίνακα 4 ότι οι τιμές είναι συγκεντρωμένες μεταξύ των αριθμών 2 και 8, με το μεγαλύτερο πλήθος παρατηρήσεων να καταγράφεται στα 4 projects, όπως διαφαίνεται και στην εικόνα 9.

1	2	3	4	5	6	7	8	9	10
0	2388	6443	8420	7126	3935	1430	256	0	0

Πίνακας 4. Κατανομή Αριθμού Project



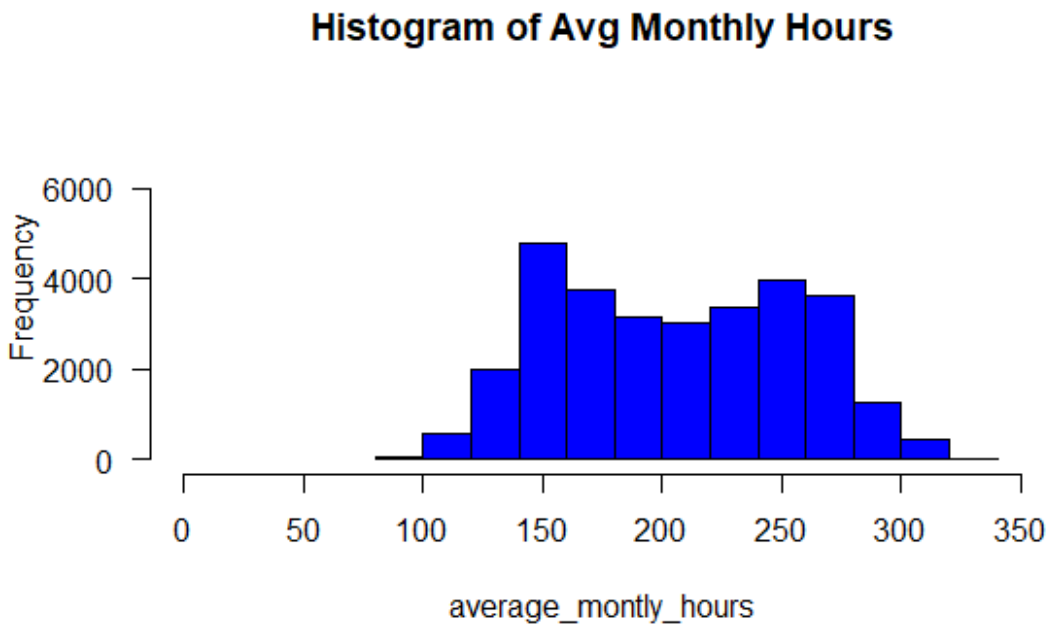
Εικόνα 9. Ιστόγραμμα Αριθμού Project

Στον πίνακα 5 και στην εικόνα 10 απεικονίζονται στοιχεία σχετικά με τις συχνότητες των μέσων μηνιαίων ωρών εργασίας. Παρατηρούμε ότι το μεγαλύτερο πλήθος παρατηρήσεων καταγράφεται στο διάστημα μεταξύ 140 και 160 ωρών εργασίας, ενώ η αμέσως μεγαλύτερη συχνότητα σημειώνεται μεταξύ των 240 και 260 ωρών. Επιπλέον, από το σύνολο των παρατηρήσεων που καταγράφονται στις κλάσεις που βρίσκονται άνω του 160, είναι εμφανές ότι η πλειοψηφία των εργαζομένων απασχολείται υπερωριακά, δεδομένου ότι οι 160 ώρες είναι αυτές που αποτελούν το κανονικό μηνιαίο ωράριο εργασίας.

(80, 100]	(100, 120]	(120, 140]	(140, 160]	(160, 180]	(180, 200]
73	571	1982	4767	3764	3138

(200, 220]	(220, 240]	(240, 260]	(260, 280]	(280, 300]	(300, 320]
3010	3381	3966	3621	1276	449

Πίνακας 5. Κατανομή Μέσων Μηνιαίων Ωρών Εργασίας

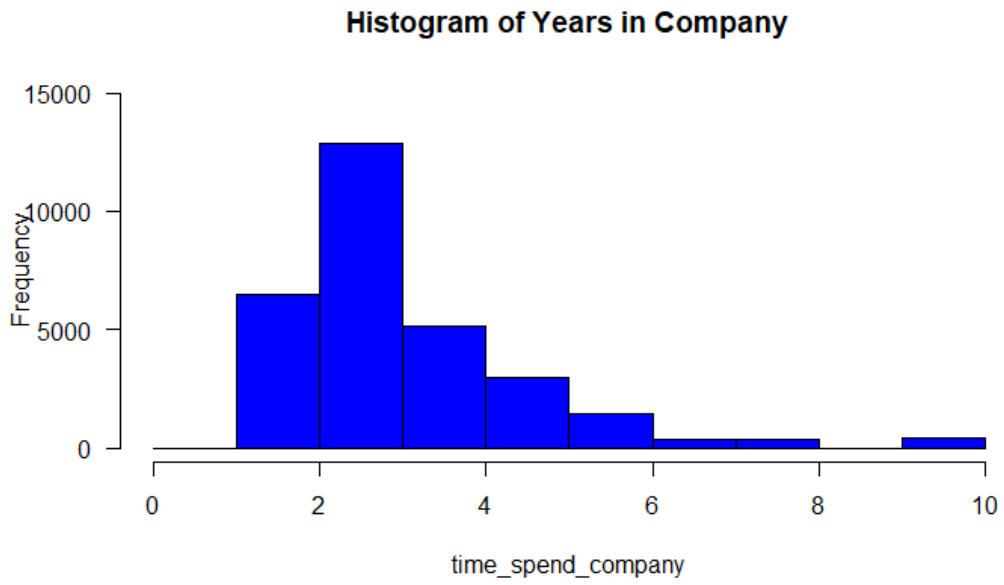


Εικόνα 10. Ιστόγραμμα Μηνιαίων Ωρών Εργασίας

Ο πίνακας 6 που ακολουθεί δείχνει την κατανομή των ετών απασχόλησης στην εταιρία. Σε συνδυασμό με την εικόνα 11 διαπιστώνουμε ότι οι περισσότεροι εργαζόμενοι απασχολούνται 3 χρόνια στην εταιρία, ενώ η κλάση με την αμέσως υψηλότερη συχνότητα είναι αυτή που αναφέρεται στα 2 χρόνια απασχόλησης.

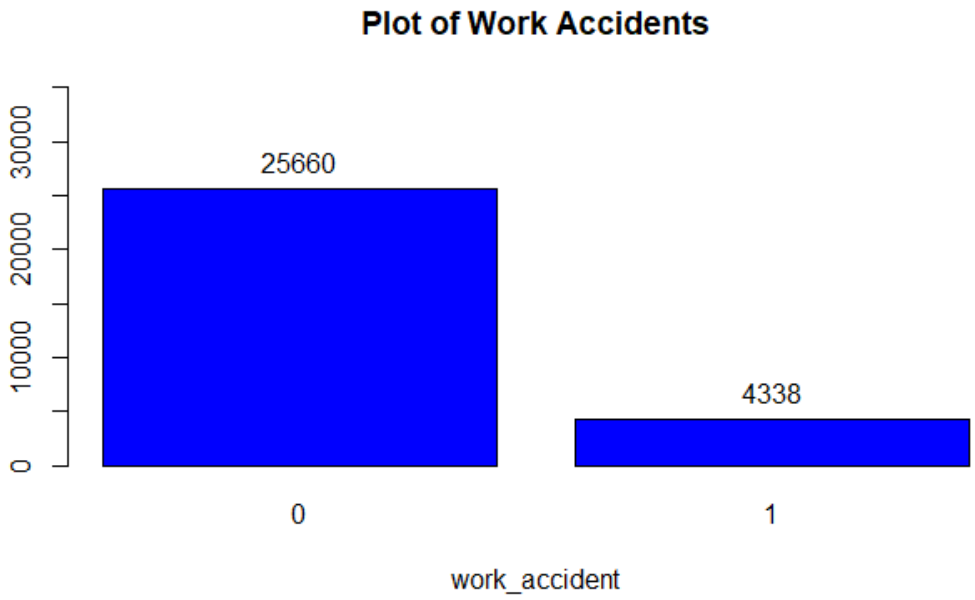
1	2	3	4	5	6	7	8	9	10
0	6488	12886	5114	2946	1436	376	324	0	428

Πίνακας 6. Κατανομή Ετών Απασχόλησης

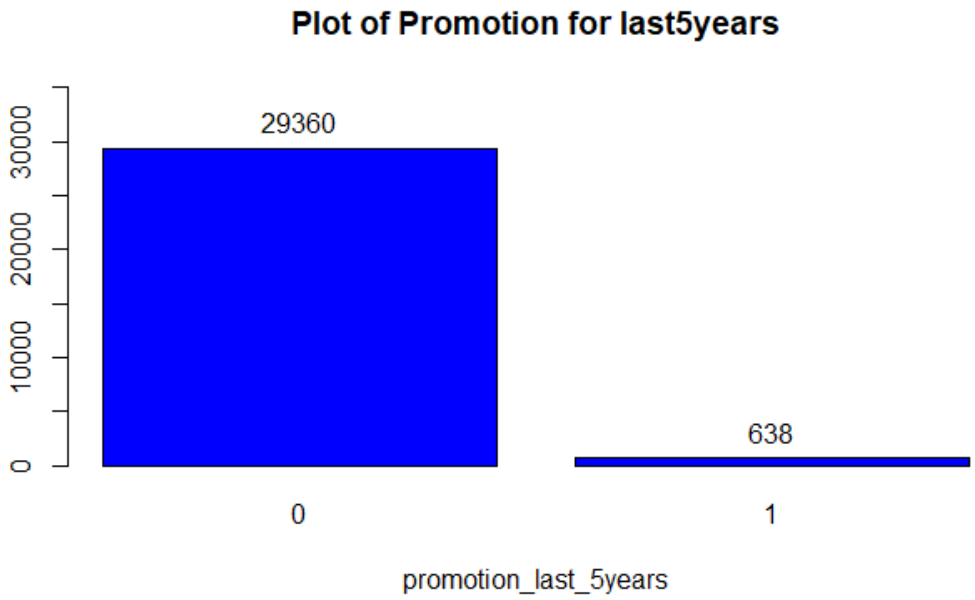


Εικόνα 11. Ιστόγραμμα Ετών Απασχόλησης

Στις εικόνες που ακολουθούν παρατίθενται τα ραβδογράμματα των κατηγορικών μεταβλητών, του data set των εργαζομένων. Στην εικόνα 12 βλέπουμε ότι η πλειοψηφία των παρατηρήσεων καταγράφεται στην περίπτωση της μη ύπαρξης κάποιου ατυχήματος στον εργασιακό χώρο. Ενώ στην εικόνα 13 βλέπουμε ότι οι συντριπτική πλειοψηφία των εργαζομένων, δεν έλαβαν κάποια προαγωγή τα τελευταία 5 χρόνια .

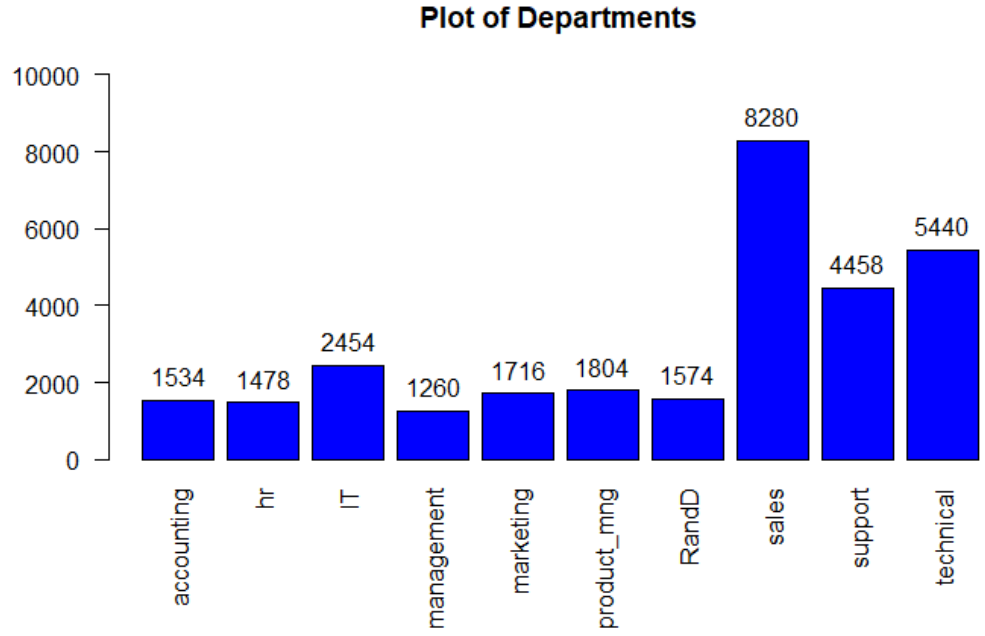


Εικόνα 12. Ραβδόγραμμα Εργατικών Ατυχημάτων

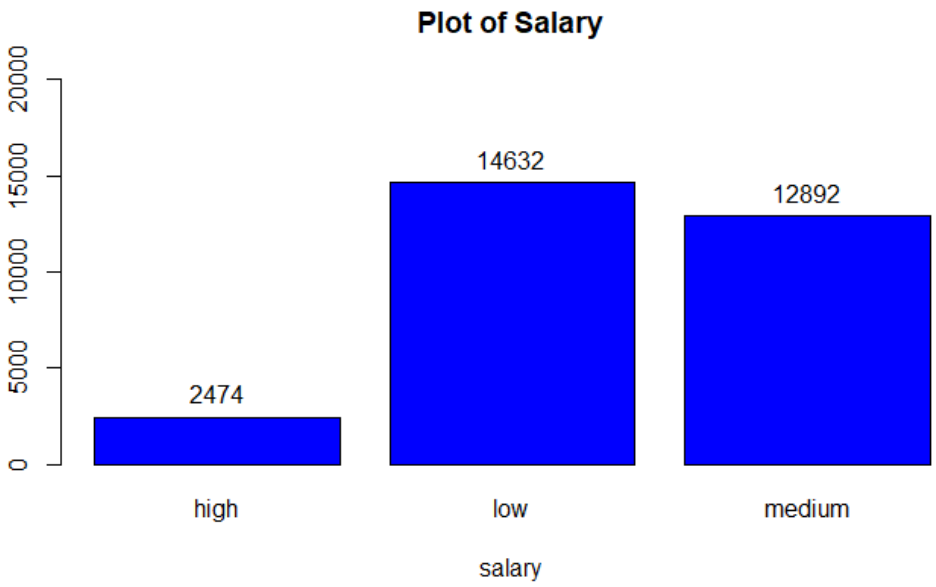


Εικόνα 13. Ραβδόγραμμα Προαγωγών

Στην εικόνα 14 βλέπουμε το πλήθος των υπαλλήλων που απασχολούνται στα διάφορα τμήματα της εταιρίας, με το τμήμα των sales και ακολούθως το technical, να συγκεντρώνουν τους περισσότερους εργαζομένους. Ακολούθως, βάσει του ραβδογράμματος της εικόνας 15 λαμβάνουμε την πληροφορία, ότι οι περισσότεροι εργαζόμενοι στην συγκεκριμένη εταιρία είναι κυρίως χαμηλόμισθοι ή μεσαίας μισθολογικής κλίμακας.

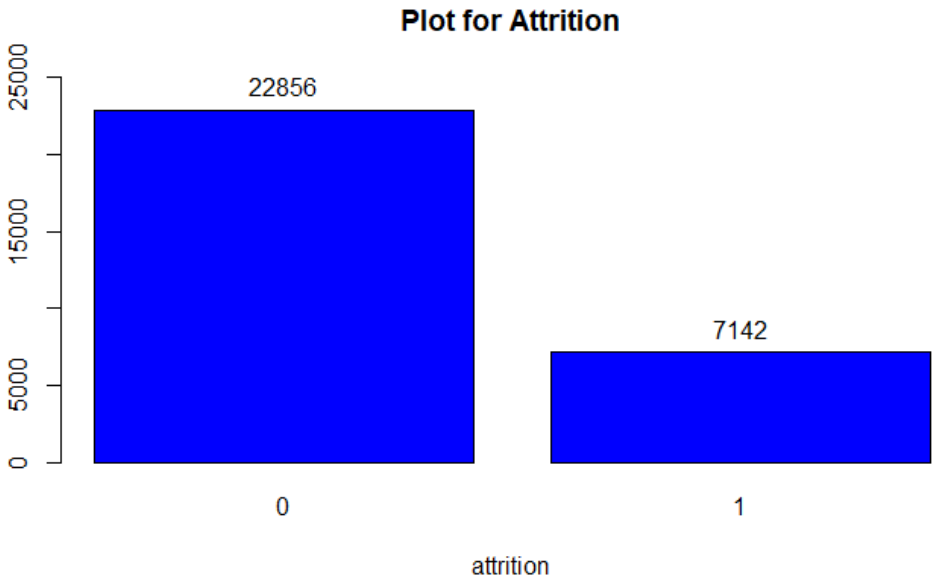


Εικόνα 14. Ραβδόγραμμα Εργαζομένων ανά Τμήμα



Εικόνα 15. Ραβδόγραμμα Μισθολογικού Επιπέδου

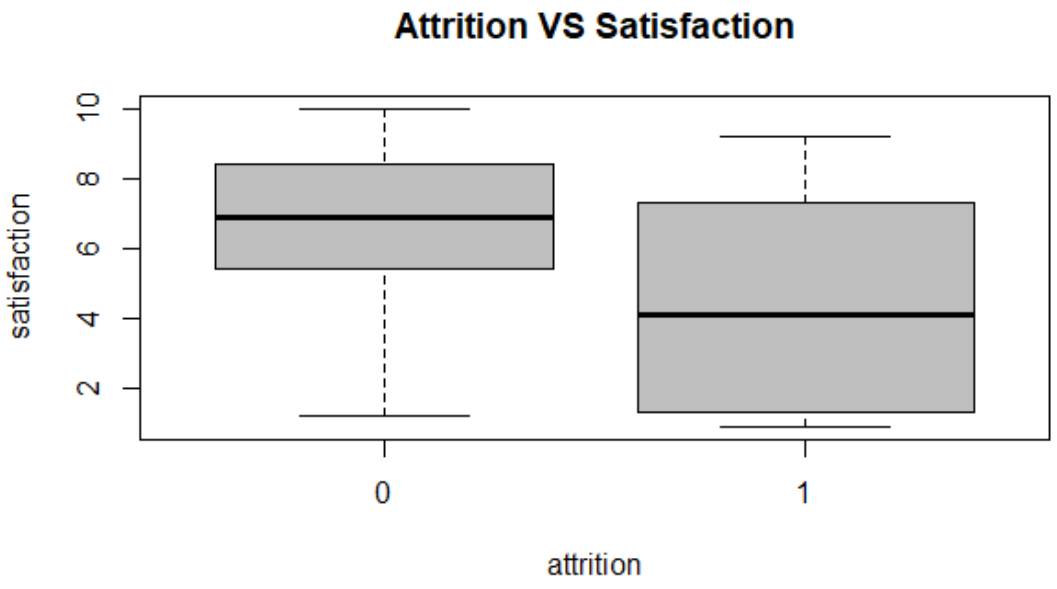
Τέλος στην εικόνα 16 παρατίθεται η κατανομή των εργαζομένων, σε σχέση με το αν έχουν αποχωρήσει εθελοντικά ή όχι από την εταιρία. Βάσει αυτής, βλέπουμε ότι το πλήθος αυτών που αποχωρούν από την εταιρία είναι πολύ μικρότερο αυτών που παραμένουν.



Εικόνα 16. Ραβδόγραμμα Εργασιακής Φθοράς

Ένα ακόμη χρήσιμο εργαλείο για να κατανοήσουμε καλύτερα τα δεδομένα μας και την σχέση τους με την εργασιακή φθορά, είναι τα boxplots. Συγκεκριμένα, μέσω αυτών στοχεύουμε στο να κατανοήσουμε καλύτερα, πως κατανέμονται οι τιμές των χαρακτηριστικών των εργαζομένων που παραμένουν στην εταιρία, σε αντιδιαστολή με την κατανομή των τιμών αυτών που παραιτούνται. Η χρήση λοιπόν των boxplots, θα μας βοηθήσει στο να διαφανούν πιο άμεσα τυχόν ανομοιότητες, που υπάρχουν στα χαρακτηριστικά των δυο κατηγοριών εργαζομένων που εξετάζουμε. Κατά συνέπεια, η ύπαρξη διαφοροποιήσεων στις κατανομές αυτών, θα μας δώσει μια πρώτη εικόνα, σχετικά με τα αποτελέσματα που αναμένουμε να προκύψουν από την πολυμεταβλητή ανάλυση, που θα ακολουθήσει στο επόμενο κεφάλαιο, σχετικά με τους πιθανούς παράγοντες που συμβάλλουν στην εργασιακή φθορά.

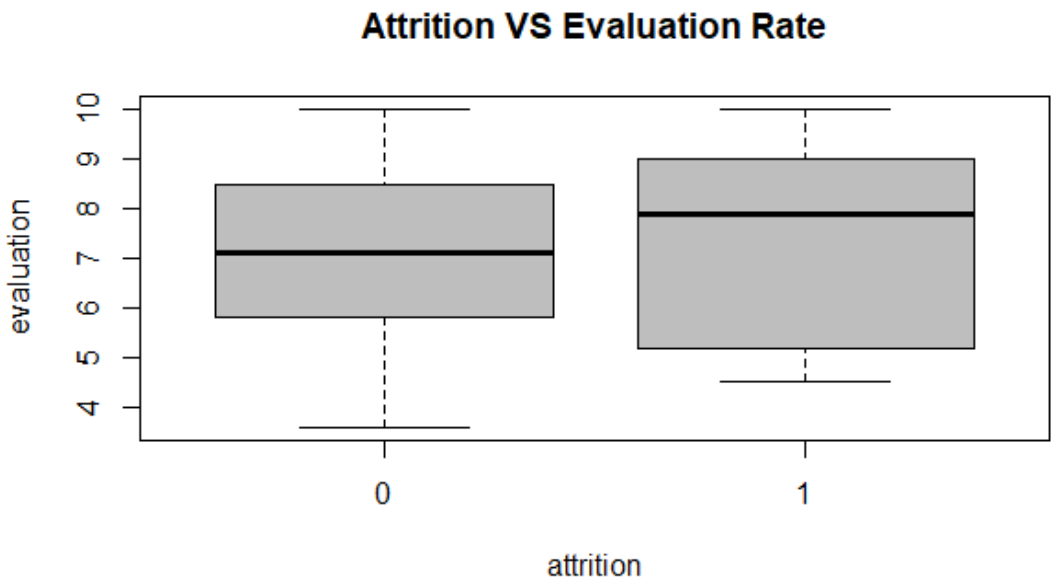
Το boxplot αποτελεί στην ουσία την γραφική απεικόνιση των κυριότερων μέτρων θέσης των συνεχών μεταβλητών, η οποία διευκολύνει σε μεγάλο βαθμό την εξαγωγή συμπερασμάτων για την κατανομή των παρατηρήσεων. Είναι γνωστό και ως το διάγραμμα των πέντε αριθμών, καθώς σε αυτό απεικονίζεται η μέγιστη και η ελάχιστη τιμή των παρατηρήσεων, η διάμεσος, το πρώτο τεταρτημόριο (Q1) και το τρίτο τεταρτημόριο (Q3). Προχωρούμε λοιπόν, στην κατασκευή τους για κάθε μια από τις συνεχείς μεταβλητές μας, σε σχέση πάντα με τις δύο τιμές που λαμβάνει η μεταβλητή attrition, η οποία αποτελεί τη μεταβλητή ενδιαφέροντος μας (Benjamini, 1988).



Εικόνα 17. Boxplot Satisfaction Level έναντι Attrition

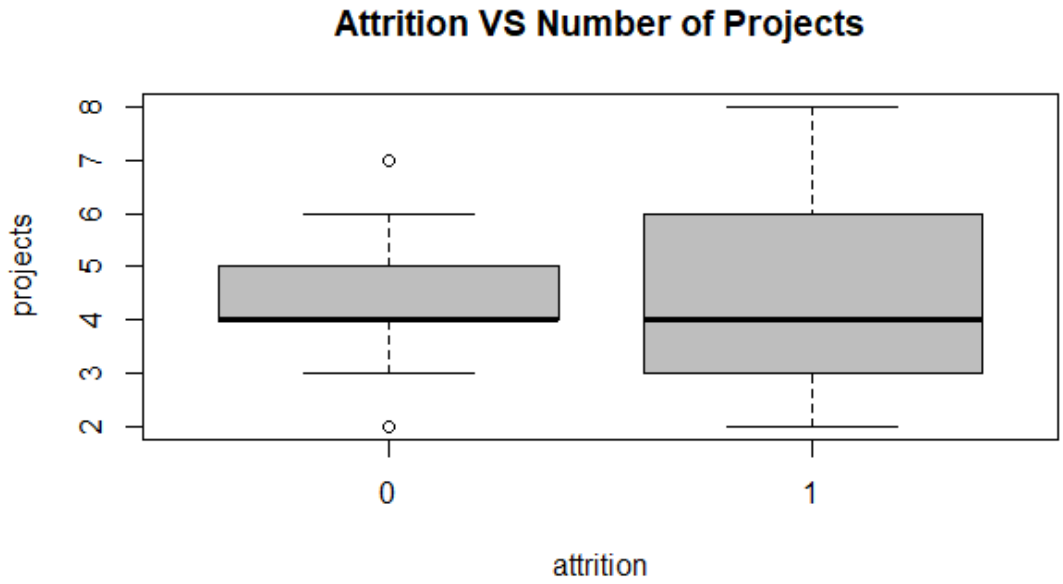
Στην εικόνα 17 φαίνεται ο διαχωρισμός των δεδομένων, μεταξύ αυτών που παραμένουν στην εταιρία και αυτών που έφυγαν, σε συνάρτηση με το επίπεδο ικανοποίησης τους. Παρατηρούμε ότι τα επίπεδα ικανοποίησης των εργαζομένων, κατανέμονται συμμετρικά γύρω από την κάθε

διάμεσο. Ταυτόχρονα είναι εμφανές ότι τα επίπεδα ικανοποίησης των ατόμων που αποχωρούν από την εργασία τους, είναι χαμηλότερα αυτών που παραμένουν. Ενώ την ίδια στιγμή, η διασπορά (ενδοτεταρτημοριακό εύρος) των επιπέδων ικανοποίησης, αυτών που παραιτήθηκαν είναι σημαντικά μεγαλύτερη αυτών που παραμένουν, καθώς τα τελευταία περιορίζονται μεταξύ των τιμών 5,5 και 8,5.



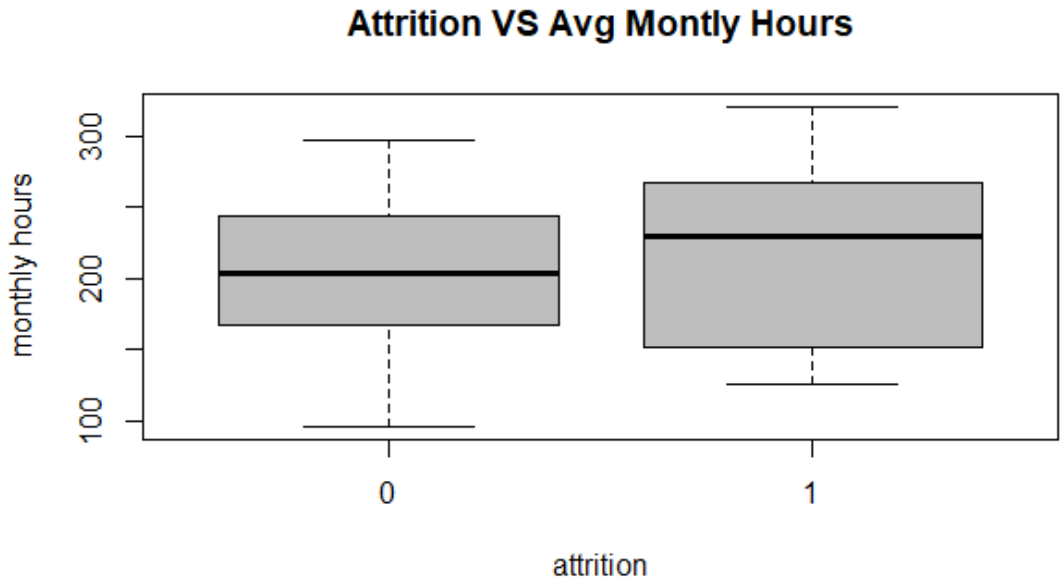
Εικόνα 18. Boxplot Last Evaluation Rate έναντι Attrition

Από τον διαχωρισμό των δεδομένων, βάσει του δείκτη τελευταίας αξιολόγησης των εργαζομένων, που παρουσιάζεται στην εικόνα 18, βλέπουμε ότι υπάρχει συμμετρία στις αξιολογήσεις αυτών που παραμένουν στην εταιρία και αρνητική ασυμμερτρία στις αξιολογήσεις αυτών που αποχώρησαν. Η αρνητική ασυμμετρία υποδηλώνει ότι υπάρχει μεγάλη μεταβλητότητα στους βαθμούς αξιολόγησης, που είναι μικρότεροι της τιμής της διαμέσου. Ταυτόχρονα όμως, όπως διαφαίνεται από τις τιμές των δύο διαμέσων, το 50% των ατόμων που αποχωρούν από την εταιρία είναι περισσότερα ικανά από το 50% αυτών που παραμένουν.



Εικόνα 19. Boxplot Projects έναντι Attrition

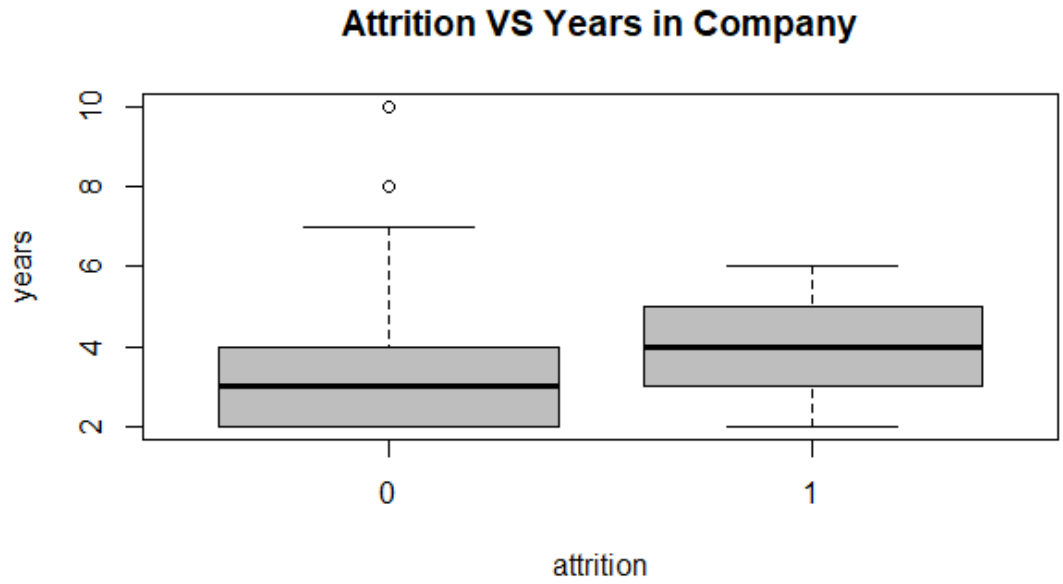
Μελετώντας τα boxplots της εικόνας 19 που σχετίζονται με τον μέσο αριθμό των projects, που αναλαμβάνουν οι εργαζόμενοι, βλέπουμε καταρχάς ότι υπάρχουν δύο έκτοπα σημεία στην κατανομή αυτών που παραμένουν στην εταιρία, στις τιμές 2 και 7. Επιπλέον είναι εμφανές, ότι η πλειοψηφία αυτών που συνεχίζουν και εργάζονται στην εταιρία, αναλαμβάνουν μικρότερο αριθμό projects (4-5), έναντι της πλειοψηφίας αυτών που αποχωρούν, που αναλαμβάνουν κυρίως 4 με 6 projects. Επίσης, ένα 25% αυτών που παραιτούνται, φαίνεται ότι αναλαμβάνει ακόμη και 6 με 8 projects, την ίδια στιγμή που το μέγιστο πλήθος που αναλαμβάνουν όσοι παραμένουν στην εταιρία είναι τα 6 projects. Ωστόσο, οι διαφορές στο πλήθος των projects μπορεί να συσχετίζεται με τα περισσότερα χρόνια στην εταιρία αυτών που παραιτούνται έναντι αυτών που παραμένουν σε αυτή.



Εικόνα 20. Boxplot Average Monthly Hours έναντι Attrition

Από την κατανομή των μέσων ωρών εργασίας της εικόνας 20, παρατηρούμε ότι υπάρχει συμμετρία στις ώρες εργασίας αυτών που παραμένουν και αρνητική ασυμμετρία στις ώρες αυτών που παραιτούνται. Δηλαδή, για αυτούς που παραιτούνται, σημειώνεται μεγαλύτερη διασπορά στα επίπεδα των ωρών που είναι μικρότερα της διαμέσου. Επίσης είναι εμφανές ότι οι τιμές της διαμέσου, της μέγιστης και ελάχιστης τιμής αυτών που παραιτούνται βρίσκονται σε υψηλότερα επίπεδα έναντι αυτών που μένουν στην εταιρία. Γενικότερα λοιπόν, όσοι αποχωρούν από την εταιρία είναι άτομα που εργάζονται πολύ περισσότερες ώρες από αυτούς που βρίσκονται ακόμα σε αυτήν.

Τέλος, η εικόνα 21 μας δίνει την πληροφορία σχετικά με το πως κατανέμεται ο αριθμός των ετών απασχόλησης στην εταιρία, σε σχέση με την ύπαρξη ή όχι εργασιακής φθοράς. Μια σημαντική επισήμανση που μπορούμε να κάνουμε είναι ότι το μεσαίο 50% των εργαζομένων που παραμένουν στην εταιρία εργάζονται σε αυτήν 2 με 4 χρόνια. Ενώ το μεσαίο 50% αυτών που αποχωρούν, σημειώνουν διάστημα προϋπηρεσίας στην εταιρία, που κυμαίνεται μεταξύ τριών και πέντε χρόνων.

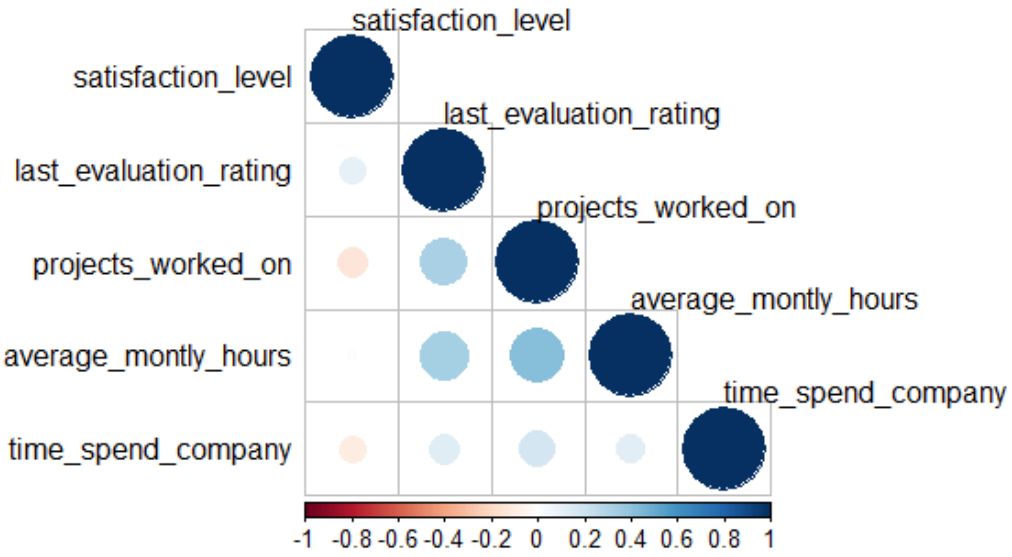


Εικόνα 21. Boxplot Years in Company έναντι Attrition

Βάσει των παραπάνω μπορούμε να καταλήξουμε σε ένα πρώιμο συμπέρασμα, ότι ενδεχόμενες αιτίες για να αποχωρήσουν κάποιοι εργαζόμενοι από την εταιρία, είναι το χαμηλό επίπεδο ικανοποίησης τους, οι πολλές ώρες εργασίας και ο υπερβολικός όγκος καθηκόντων. Ενώ επίσης οι υψηλές δεξιότητες και η μακρόχρονη εμπειρία, μπορεί να είναι άλλοι λόγοι, που πιθανόν τους καλλιεργούν την επιθυμία για αλλαγή του εργασιακού τους περιβάλλοντος ή για αναζήτηση νέων προκλήσεων.

Ένα άλλο σημαντικό μέτρο που μπορεί να μας βοηθήσει να κατανοήσουμε καλύτερα τα δεδομένα μας είναι η συσχέτιση μεταξύ των μεταβλητών (Bonett & Wright, 2000). Ταυτόχρονα, το συγκεκριμένο μέτρο μας επισημαίνει αν κάποιες από τις μεταβλητές του προβλήματος πρέπει να αφαιρεθούν, σε περίπτωση που συσχετίζονται υψηλά μεταξύ τους, προκειμένου να εξασφαλίσουμε ότι όλες οι μεταβλητές έχουν το ίδιο βάρος στην επίδραση που ασκούν στο attrition. Κάνοντας χρήση της παραμετρικής μεθόδου συσχετίσεων Pearson, η οποία εξετάζει αν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών, λαμβάνουμε τα αποτέλεσμα που εμφανίζονται στην εικόνα 22.

Βάσει αυτού, διαπιστώνουμε ότι οι μεταβλητές που αφορούν στον αριθμό των projects, στις μέσες μηνιαίες ώρες εργασίας και στα χρόνια απασχόλησης στην εταιρία, συσχετίζονται αρνητικά με το επίπεδο ικανοποίησης των εργαζομένων, χωρίς ωστόσο να υπάρχει σε καμία από αυτές κάποια ισχυρή γραμμική σχέση. Μεταξύ των υπολοίπων αριθμητικών μεταβλητών, διαπιστώνεται ότι υπάρχει θετική συσχέτιση μεταξύ τους, ωστόσο η γραμμική σχέση μεταξύ τους είναι επίσης ασθενής, γεγονός που επιβεβαιώνεται από τις τιμές των συσχετίσεων που παρατίθενται στην εικόνα 23.

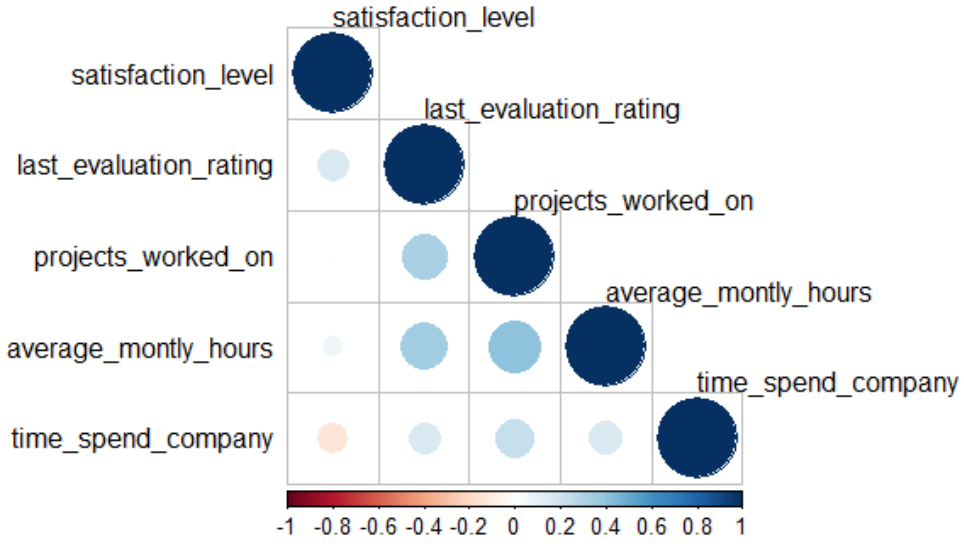


Εικόνα 22. Correlation Plot Pearson

	satisfaction_level	last_evaluation_rating	projects_worked_on	average_montly_hours	time_spend_company
satisfaction_level	1				
last_evaluation_rating	0,105	1			
projects_worked_on	-0,132	0,324	1		
average_montly_hours	-0,020	0,338	0,422	1	
time_spend_company	-0,101	0,132	0,182	0,127	1

Εικόνα 23. Πίνακας συσχετίσεων Pearson

Αντίστοιχα αποτελέσματα λαμβάνουμε κάνοντας χρήση της μη- παραμετρικής μεθόδου Spearman, η οποία εξετάζει την ύπαρξη μονοτονικής σχέσης μεταξύ των μεταβλητών. Από την εικόνα 24 και τις τιμές της εικόνας 25, μπορούμε να εξάγουμε το συμπέρασμα ότι η μονοτονική σχέση μεταξύ των μεταβλητών είναι επίσης ασθενής. Τώρα παρατηρούμε ότι το επίπεδο ικανοποίησης έχει αρνητική αλλά ασθενή μονοτονική σχέση με τον αριθμό των projects και με τα χρόνια εργασίας στην εταιρία, όπως φαίνεται στην εικόνα 25.



Εικόνα 24. Correlation Plot Spearman

	satisfaction_level	last_evaluation_rating	projects_worked_on	average_montly_hours	time_spend_company
satisfaction_level	1				
last_evaluation_rating	0,162	1			
projects_worked_on	-0,004	0,323	1		
average_montly_hours	0,062	0,340	0,405	1	
time_spend_company	-0,139	0,161	0,236	0,168	1

Εικόνα 25. Πίνακας συσχετίσεων Spearman

3. ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ

3.1 ΜΕΘΟΔΟΛΟΓΙΑ ΑΝΑΛΥΣΗΣ

Στο παρόν κεφάλαιο θα ακολουθήσει η πρακτική εφαρμογή των τριών μεθόδων στατιστικής μάθησης, που αποτυπώθηκαν θεωρητικά στο πρώτο κεφάλαιο. Στόχος μας είναι η εκπαίδευση των αλγορίθμων και η εξαγωγή τριών ξεχωριστών μοντέλων ταξινόμησης, μεταξύ των οποίων θα αποφανθούμε τελικά, ποιο είναι το καλύτερο για την μελλοντική ταξινόμηση άγνωστων δεδομένων. Για τον σκοπό αυτό, σε κάθε αλγόριθμο κρατάμε το 70% του συνόλου των παρατηρήσεων, για την εκπαίδευση του αλγορίθμου και το υπόλοιπο 30% για τον έλεγχο των αποτελεσμάτων του. Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων, καλούνται δεδομένα εκπαίδευσης (training set), ενώ αυτά που χρησιμοποιούνται για την αξιολόγηση των μοντέλων που προκύπτουν από αυτούς, καλούνται δεδομένα ελέγχου (test set). Τα δεδομένα εκπαίδευσης αποτελούν στην ουσία παραδείγματα, τα οποία καθοδηγούν την διαδικασία «μάθησης» των αλγορίθμων, ενώ τα δεδομένα ελέγχου βοηθούν στο να αξιολογήσουμε πόσο καλά πραγματοποιήθηκε αυτή η διαδικασία.

Η διαδικασία που ακολουθείται σε όλους τους αλγορίθμους είναι η ακόλουθή. Αρχικά, πραγματοποιείται η εκπαίδευση τους, κάνοντας χρήση των δεδομένων εκπαίδευσης. Ύστερα από αυτή την πρώτη διαδικασία εκπαίδευσής τους, προχωρούμε στην αξιολόγηση των αποτελεσμάτων τους, κάνοντας χρήση των μέτρων αξιολόγησης, που αναφέρθηκαν στο Κεφάλαιο 1, τόσο στα γνωστά για αυτούς δεδομένα του training set, όσο και στα άγνωστα δεδομένα του test set. Βάσει της επίδοσης τους και στα δύο σετ δεδομένων, κρίνουμε αν το μοντέλο μας είναι αξιόπιστο, για την γενίκευση του σε άλλα άγνωστα δεδομένα του πληθυσμού ή αν χρήζει βελτιώσεων. Οι βελτιώσεις, όπου κρίνεται αναγκαίο, επιτυγχάνονται με την σταδιακή μεταβολή διαφόρων παραμέτρων στον κάθε αλγόριθμο. Έπειτα από κάθε μεταβολή, επαναξιολογούνται τα αποτελέσματα των αλγορίθμων, εκ νέου στο training set και στο test set. Τέλος, επαναλαμβάνουμε την διαδικασία της παραμετροποίησης και της αξιολόγησης, έως ότου λάβουμε κάποιο ικανοποιητικό αποτέλεσμα, σχετικά με το πρόβλημα που εξετάζουμε.

3.2 ΜΟΝΤΕΛΟ LOGISTIC REGRESSION

Πριν προχωρήσουμε στην εφαρμογή της λογιστικής παλινδρόμηση (Logistic Regression) στα δεδομένα της βάσης «employees» που χρησιμοποιούμε, απαιτείται να προβούμε στον μετασχηματισμό των κατηγορικών μεταβλητών σε ψευδομεταβλητές (dummies) (Suits, 1957) (Πανάρετος, 2003). Δεδομένου ότι οι κατηγορικές μεταβλητές αποτελούνται από δύο ή περισσότερα διακεκριμένα επίπεδα (levels), με τον τρόπο αυτό εξασφαλίζεται η ποσοτικοποίηση των μεταβλητών αυτών, προκειμένου να καταστεί εφικτή η συμμετοχή και

ανάλυση των διαφορετικών τους επιπέδων, από το στατιστικό μοντέλο. Στο παρόν πρόβλημα, οι κατηγορικές μεταβλητές department και salary είναι αυτές που μετασχηματίζουμε σε dummies. Αυτό σημαίνει ότι κάθε level των κατηγορικών μεταβλητών, θα αποτελεί μια ξεχωριστή δυαδική μεταβλητή στο μοντέλο μας, η οποία θα παίρνει τιμές "0" ή "1". Όπου η τιμή 1 επισημαίνει, πότε η αρχική μεταβλητή λαμβάνει την τιμή κάποιας εκ των ξεχωριστών κατηγοριών, που περιλαμβάνονται σε αυτή.

Μετά την δημιουργία μιας dummy για κάθε level των κατηγορικών μεταβλητών, απαιτείται μια εξ' αυτών να μείνει εκτός της εκπαίδευσης του μοντέλου, για την αποφυγή του προβλήματος τέλειας πολυσυγγραμικότητας. Το συγκεκριμένο πρόβλημα αναφέρεται στην περίπτωση όπου μεταξύ δύο ή περισσότερων μεταβλητών υπάρχει σχέση γραμμικής εξάρτησης. Με άλλα λόγια η συσχέτιση που υπάρχει μεταξύ τους ισούται με 1 ή -1. Συνεπώς μια ή περισσότερες εξ' αυτών δεν θα πρέπει να συμπεριληφθούν στο μοντέλο, για να είναι δυνατός ο προσδιορισμός των ατομικών επιδράσεων των ανεξάρτητων μεταβλητών στην εξαρτημένη μεταβλητή. Έτσι από τον μετασχηματισμό της μεταβλητής department προκύπτουν δέκα dummies, όσα και τα levels από τα οποία αποτελείται, ενώ αντίστοιχα από την μεταβλητή salary προκύπτουν τρείς dummies. Η ψευδομεταβλητή που μένει εκτός της εκπαίδευσης του αλγορίθμου, είναι αυτή που θα αποτελέσει το επίπεδο αναφοράς μας, κατά την ερμηνεία του μοντέλου (Gujarati & Porter, 2009).

Το τελικό dataset που θα χρησιμοποιήσουμε για την εκπαίδευση του αλγορίθμου, αποτελείται συνολικά από 19 μεταβλητές έναντι των 10 που είχαμε αρχικά. Στην εικόνα 26 βλέπουμε τις 10 πρώτες εγγραφές του αναδιαμορφωμένου πλέον dataset των employees.

	satisfaction_level	last_evaluation_rating	projects worked_on	average_montly_hours	time_spend_company	work_accident	promotion_last_5years	Attrition	Dep_hr	Dep_IT	Dep_management	Dep_marketing	Dep_product_mng	Dep_RandD	Dep_sales	Dep_support	Dep_technical	salary_low	salary_medium
1	3.8	5.3	3	167	3	0	0	1	0	0	0	0	0	0	1	0	0	1	0
2	8.0	8.6	6	272	6	0	0	1	0	0	0	0	0	0	1	0	0	0	1
3	1.1	8.8	8	282	4	0	0	1	0	0	0	0	0	0	1	0	0	0	1
4	3.7	5.2	3	169	3	0	0	1	0	0	0	0	0	0	1	0	0	1	0
5	4.1	5.0	3	163	3	0	0	1	0	0	0	0	0	0	1	0	0	1	0
6	1.0	7.7	7	257	4	0	0	1	0	0	0	0	0	0	1	0	0	1	0
7	9.2	8.5	6	269	5	0	0	1	0	0	0	0	0	0	1	0	0	1	0
8	8.9	10.0	6	234	5	0	0	1	0	0	0	0	0	0	1	0	0	1	0
9	4.2	5.3	3	152	3	0	0	1	0	0	0	0	0	0	1	0	0	1	0
10	1.1	8.1	7	315	4	0	0	1	0	0	0	0	0	0	1	0	0	1	0

Εικόνα 26. Dataset Employees for Logistic Regression

Σε αυτό συμπεριλαμβάνονται οι μεταβλητές του επιπέδου ικανοποίησης και του βαθμού τελευταίας αξιολόγησης των εργαζομένων, οι οποίες λαμβάνουν πραγματικές τιμές, όπως αναφέρθηκε και στο Κεφάλαιο 2. Επιπλέον συμμετέχουν οι μεταβλητές του αριθμού των projects, των μέσων μηνιαίων ωρών εργασίας και των ετών απασχόλησης στην εταιρία, όπου όλες λαμβάνουν ακέραιες τιμές, όπως επίσης αναφέρθηκε στο Κεφάλαιο 2. Ακολούθως

περιλαμβάνονται οι κατηγορικές μεταβλητές που αναφέρονται στην ύπαρξη ή μη εργατικού ατυχήματος και στην λήψη ή μη προαγωγής εντός πενταετίας, οι οποίες λαμβάνουν τιμές της μορφής “0” ή “1”. Τέλος, στο νέο dataset περιλαμβάνονται οι εννέα από τις δέκα ψευδομεταβλητές που αφορούν τα διάφορα τμήματα της εταιρίας, εξαιρουμένου του τμήματος accounting και οι δύο από τις τρείς ψευδομεταβλητές που αφορούν τα επίπεδα των μισθών, εξαιρουμένου του επιπέδου high, οι οποίες όπως αναφέρθηκε παραπάνω είναι επίσης δίτιμες της μορφής “0” ή “1”.

Ύστερα από την προ-επεξεργασία των δεδομένων και μετά τον διαχωρισμό τους σε υποσύνολα, του training set και του test set σε ποσοστά 70% και 30% αντίστοιχα, όπως προαναφέρθηκε στην αρχή του παρόντος κεφαλαίου, είμαστε πλέον σε θέση να προχωρήσουμε στην εκπαίδευση του αλγορίθμου μας. Η διαδικασία που θα ακολουθήσουμε για να καταλήξουμε στην κατασκευή του τελικού υποδείγματος, είναι αυτή της προοδευτικής ή σταδιακής απόρριψης των μεταβλητών (backward stepwise elimination) (Hastie & et al, 2009). Ξεκινούμε λοιπόν, με την εφαρμογή του αλγορίθμου στο πλήρες μοντέλο, στο οποίο συμπεριλαμβάνονται όλες οι μεταβλητές που περιγράφηκαν παραπάνω. Έπειτα, αφαιρούμε σταδιακά μεταβλητές, μια κάθε φορά, αυτής που είναι λιγότερο σημαντική για την ανάλυση μας. Σταματούμε την διαδικασία αφαίρεσης μεταβλητών από το μοντέλο, όταν απομείνουν σε αυτό μόνο οι στατιστικά σημαντικές μεταβλητές, θεωρώντας επίπεδο σημαντικότητας το 0,05.

Έτσι, στο περιβάλλον της R εφαρμόζουμε τον αλγόριθμο της λογιστικής παλινδρόμησης στο πλήρες μοντέλο, καλώντας την συνάρτηση glm και ορίζοντας family=binomial. Η σύνοψη των αποτελεσμάτων που λαμβάνουμε παρατίθεται στην εικόνα 27.

```
Call:
glm(formula = Attrition ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1622  -0.6690  -0.4116  -0.1237   3.0847

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.3392483  0.1618115  -8.277  < 2e-16 ***
satisfaction_level      -0.3998073  0.0081434 -49.096  < 2e-16 ***
last_evaluation_rating   0.0591654  0.0123553   4.789 1.68e-06 ***
projects_worked_on      -0.2663297  0.0163143 -16.325  < 2e-16 ***
average_montly_hours     0.0042351  0.0004309   9.829  < 2e-16 ***
time_spend_company       0.2552213  0.0129494  19.709  < 2e-16 ***
Work_accident           -1.5444684  0.0764043 -20.214  < 2e-16 ***
promotion_last_5years   -1.3422829  0.2126545  -6.312 2.75e-10 ***
Dep_hr                   0.2425442  0.1096115   2.213   0.0269 *
Dep_IT                  -0.2201836  0.1023857  -2.151   0.0315 *
Dep_management          -0.4026022  0.1326162  -3.036   0.0024 **
Dep_marketing           -0.1079769  0.1118341  -0.966   0.3343
Dep_product_mng         -0.1120151  0.1078780  -1.038   0.2991
Dep_RandD               -0.7100806  0.1242241  -5.716 1.09e-08 ***
Dep_sales               -0.0730182  0.0857941  -0.851   0.3947
Dep_support              0.0127687  0.0918272   0.139   0.8894
Dep_technical            0.0607131  0.0894700   0.679   0.4974
salary_low               1.8777441  0.1059425  17.724  < 2e-16 ***
salary_medium            1.3572344  0.1065481  12.738  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23049  on 20997  degrees of freedom
Residual deviance: 18145  on 20979  degrees of freedom
AIC: 18183

Number of Fisher Scoring iterations: 5
```

Εικόνα 27. Αρχικό μοντέλο Logistic Regression

Έχοντας πλέον μια πρώτη εικόνα των αποτελεσμάτων που προέκυψαν από τον αλγόριθμό, μπορούμε να ξεκινήσουμε την προσαρμογή του μοντέλου, επιλέγοντας την μεταβλητή που θα πρέπει να αφαιρέσουμε πρώτη από αυτό. Γενικά μιλώντας, το μέτρο που μας υποδεικνύει την σημαντικότητα κάθε μεταβλητής και μας καθοδηγεί σε αυτή την διαδικασία είναι το p-value. Έτσι αφαιρούμε τις μεταβλητές που είναι μη στατιστικά σημαντικές, ξεκινώντας από αυτή που είναι λιγότερο σημαντική, από τις υπόλοιπες μη σημαντικές μεταβλητές. Μη στατιστικά σημαντικές είναι οι μεταβλητές εκείνες, για τις οποίες το p-value τους είναι μεγαλύτερο του επιπέδου σημαντικότητας που έχουμε ορίσει. Αυτό σημαίνει ότι για τις μεταβλητές αυτές, δεν υπάρχει συσχέτιση με την εξαρτημένη μεταβλητή attrition, δηλαδή δεν συνεισφέρουν στην ανάλυση του μοντέλου μας.

Βάσει των αποτελέσματα της εικόνας 27, βλέπουμε ότι το μεγαλύτερο p-value = 0,8894 το έχει η μεταβλητή Dep_support. Οπότε προχωρούμε στην αφαίρεση της συγκεκριμένης μεταβλητής και ξανατρέχουμε τον αλγόριθμο, προκειμένου να ελέγξουμε ξανά τα αποτελέσματα του νέου μοντέλου με τις νέες τιμές των p-values.

```
Call:
glm(formula = Attrition ~ . - Dep_support, family = binomial,
    data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1660  -0.6691  -0.4117  -0.1237   3.0847

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.3301798  0.1480662  -8.984  < 2e-16 ***
satisfaction_level     -0.3997759  0.0081402 -49.111  < 2e-16 ***
last_evaluation_rating  0.0591795  0.0123550   4.790 1.67e-06 ***
projects_worked_on     -0.2662990  0.0163127 -16.325  < 2e-16 ***
average_montly_hours    0.0042344  0.0004309   9.828  < 2e-16 ***
time_spend_company      0.2551788  0.0129457  19.711  < 2e-16 ***
Work_accident          -1.5443182  0.0763944 -20.215  < 2e-16 ***
promotion_last_5years  -1.3426628  0.2126626  -6.314 2.73e-10 ***
Dep_hr                  0.2332246  0.0867067   2.690 0.007149 **
Dep_IT                 -0.2295033  0.0773714  -2.966 0.003015 **
Dep_management         -0.4118418  0.1147483  -3.589 0.000332 ***
Dep_marketing          -0.1172865  0.0895541  -1.310 0.190307
Dep_product_mng        -0.1213397  0.0844768  -1.436 0.150898
Dep_RandD              -0.7193922  0.1046084  -6.877 6.11e-12 ***
Dep_sales              -0.0823376  0.0535211  -1.538 0.123948
Dep_technical           0.0513811  0.0591304   0.869 0.384877
salary_low              1.8779294  0.1059330  17.728  < 2e-16 ***
salary_medium           1.3573549  0.1065434  12.740  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23049  on 20997  degrees of freedom
Residual deviance: 18145  on 20980  degrees of freedom
AIC: 18181


Number of Fisher Scoring iterations: 5
```

Εικόνα 28. Μοντέλο Logistic Regression μετά την πρώτη προσαρμογή

Βάσει των νέων αναδιαμορφωμένων αποτελεσμάτων της εικόνας 28 καταλήγουμε στο συμπέρασμα ότι η μεταβλητή που πρέπει να αφαιρεθεί αμέσως μετά είναι Dep_technical, καθώς έχει την υψηλότερη τιμή του p-value η οποία ισούται με 0,3848. Επαναλαμβάνουμε το συγκεκριμένο βήμα, αφαιρώντας διαδοχικά τις Dep_marketing, Dep_product_mng, Dep_sales καθώς διαπιστώθηκε ότι το p-value τους είναι υψηλότερο του επιπέδου σημαντικότητας. Καταλήγουμε τελικά στο μοντέλο της εικόνας 29, όπου όλες οι μεταβλητές είναι στατιστικά σημαντικές, επομένως με αυτό μπορούμε να προχωρήσουμε στην πρόβλεψη των κλάσεων για την μεταβλητή attrition.

```
Call:
glm(formula = Attrition ~ . - Dep_support - Dep_technical - Dep_marketing -
    Dep_product_mng - Dep_sales, family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1482  -0.6682  -0.4120  -0.1244   3.0705

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.3647478  0.1439748  -9.479  < 2e-16 ***
satisfaction_level     -0.3997713  0.0081362 -49.135  < 2e-16 ***
last_evaluation_rating  0.0594460  0.0123497   4.814 1.48e-06 ***
projects_worked_on     -0.2648340  0.0162958 -16.252  < 2e-16 ***
average_montly_hours    0.0042351  0.0004306   9.835  < 2e-16 ***
time_spend_company      0.2530509  0.0129079  19.604  < 2e-16 ***
Work_accident          -1.5424579  0.0763445 -20.204  < 2e-16 ***
promotion_last_5years  -1.3457767  0.2119824  -6.349 2.17e-10 ***
Dep_hr                  0.2688946  0.0793284   3.390 0.000700 ***
Dep_IT                 -0.1936908  0.0689788  -2.808 0.004985 **
Dep_management         -0.3746626  0.1091394  -3.433 0.000597 ***
Dep_RandD              -0.6838606  0.0985718  -6.938 3.99e-12 ***
salary_low              1.8765707  0.1058868  17.722  < 2e-16 ***
salary_medium           1.3554463  0.1065141  12.726  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23049  on 20997  degrees of freedom
Residual deviance: 18154  on 20984  degrees of freedom
AIC: 18182

Number of Fisher Scoring iterations: 5
```

Εικόνα 29. Τελικό μοντέλο Logistic Regression

Βλέπουμε λοιπόν ότι το τελικό μοντέλο που προέκυψε από τον αλγόριθμό της λογιστικής παλινδρόμησης, αποτελείται από τις ανεξάρτητες μεταβλητές satisfaction level, last evaluation rating, projects worked on, average monthly hours, time spend company, work accident και promotion last 5years. Επίσης, συμπεριλαμβάνονται σε αυτό οι ψευδομεταβλητές salary_low και salary_medium που προέκυψαν από την μετατροπή της αρχικής μεταβλητής salary και οι οποίες αναφέρονται στα επίπεδα μισθών low και medium. Τέλος, στο μοντέλο συμμετέχουν ορισμένες από τις ψευδομεταβλητές που κατασκευάσαμε από την αρχική μεταβλητή του department, οι οποίες αφορούν τα τμήματα HR, IT, Management και R&D.

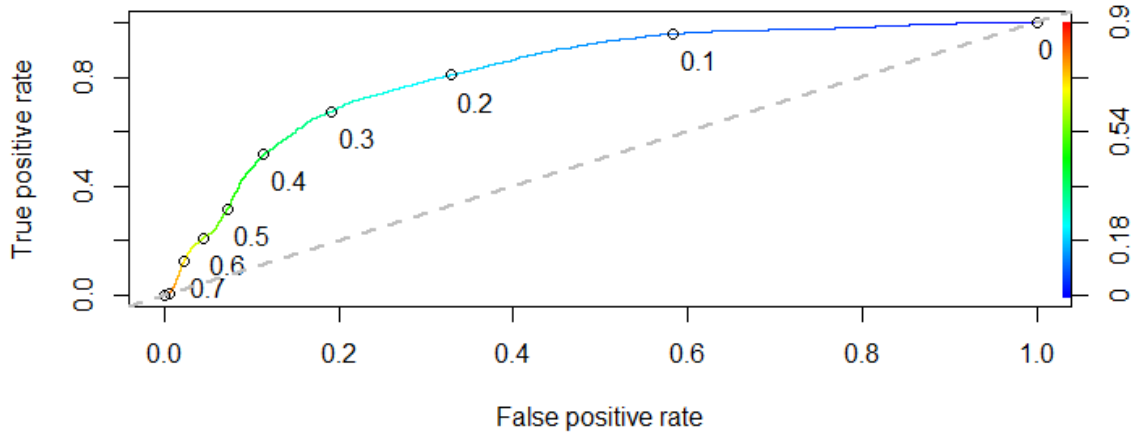
Από τις τιμές των συντελεστών παλινδρόμησης, μπορούμε να κρίνουμε ποιο είναι το μέγεθος της επίδρασης που ασκείται, από την κάθε μεταβλητή στην εμφάνιση του προβλήματος της εργασιακής φθοράς, καθώς και την κατεύθυνση της επίδρασης που ασκείται από αυτές, δηλαδή αν επιδρούν θετικά ή αρνητικά σε αυτό. Ξεκινούμε λοιπόν, με την αποτύπωση των αποτελεσμάτων που προέκυψαν βάσει της εικόνας 29, από την μεταβλητή που ασκεί την μεγαλύτερη επίδραση στην εμφάνιση του φαινομένου, προς αυτή με την μικρότερη επίδραση. Βλέπουμε λοιπόν, ότι για την ψευδομεταβλητή salary_low ο συντελεστής παλινδρόμησης ισούται με 1,88, γεγονός που υποδηλώνει ότι αν μεταβληθεί ο μισθός των εργαζομένων από υψηλός (high) σε χαμηλός (low), θα έχει σαν συνέπεια να αυξηθούν οι λογαριθμικές αποδόσεις (log-odds) της εμφάνισης εργασιακής φθοράς κατά 1,88. Την αμέσως μεγαλύτερη επίδραση στην εργασιακή φθορά βλέπουμε ότι ασκεί η μεταβλητή work accident, της οποίας όμως ο

συντελεστής έχει αρνητικό πρόσημο. Αυτό σημαίνει ότι αν κάποιος εργαζόμενος υποστεί κάποιο εργατικό ατύχημα, οι λογαριθμικές αποδόσεις της εργασιακής φοράς μειώνονται κατά 1,54. Ακολουθεί η ψευδομεταβλητή salary_medium, βάσει της οποία βλέπουμε ότι αν ο μισθός μεταβληθεί από υψηλός σε μεσαίας κλίμακας, τότε οι λογαριθμικές αποδόσεις της εργασιακής φθοράς αυξάνονται κατά 1,36. Αντίστοιχου μεγέθους επίδραση, αλλά προς την αντίθετη κατεύθυνση, ασκεί η μεταβλητή promotion_last_5years, σύμφωνα με την οποία αν κάποιος έχει λάβει προαγωγή τα τελευταία 5 χρόνια, τότε οι λογαριθμικές αποδόσεις του attrition μειώνονται κατά 1,35. Επίσης αρνητική επίδραση στις λογαριθμικές αποδόσεις του attrition, ασκεί η μετακίνηση ενός εργαζομένου από το τμήμα του accounting σε αυτό του R&D, καθώς όπως φαίνεται από την ψευδομεταβλητή dep_RandD, μειώνονται κατά 0,68. Βάσει της μεταβλητής του επιπέδου ικανοποίησης, βλέπουμε ότι αν αυτή αυξηθεί κατά μια μονάδα, τότε οι λογαριθμικές αποδόσεις της εμφάνισης εργασιακής φθοράς μειώνονται κατά 0,40. Προς την ίδια κατεύθυνση επιδρά και η μετακίνηση ενός εργαζομένου από το τμήμα του accounting σε αυτό του management, αφού όπως βλέπουμε οδηγεί σε μείωση των λογαριθμικών αποδόσεων κατά 0,37, ενώ αντίθετα η μετακίνηση από το τμήμα του accounting στο hr, οδηγεί σε αύξηση αυτών κατά 0,26. Έπειτα, παρατηρούμε ότι η αύξηση του αριθμού των project κατά μια μονάδα, οδηγεί σε μείωση των λογαριθμικών αποδόσεων κατά 0,26, ενώ αντίθετα η αύξηση κατά μια μονάδα του αριθμού των ετών απασχόλησης στην εταιρία, οδηγεί σε αύξηση αυτών κατά 0,25. Τέλος, η μοναδιαία αύξηση του βαθμού της τελευταίας αξιολόγησης βλέπουμε ότι συμβάλλει στην αύξηση των λογαριθμικών αποδόσεων του attrition κατά 0,05 ενώ η αύξηση των μέσων μηνιαίων ωρών εργασίας σε αύξηση αυτών κατά 0,004.

Έχοντας πλέον διαμορφώσει το τελικό μοντέλο του αλγορίθμου, είμαστε σε θέση να προχωρήσουμε στην πρόβλεψη της κλάσης, που ανήκει η κάθε παρατήρηση. Βάσει της θεωρητικής αποτύπωσης που προηγήθηκε στο Κεφάλαιο 1, γνωρίζουμε ότι ο αλγόριθμος της λογιστικής παλινδρόμησης εξάγει ως αποτέλεσμα την πιθανότητα της εξαρτημένης μεταβλητής, να ανήκει σε κάθε ξεχωριστή κλάση. Έτσι, δεδομένου ότι ο αλγόριθμος προβλέπει την πιθανότητα κάθε υπαλλήλου να παραιτηθεί, αυτό που απαιτείται από πλευράς μας να πράξουμε, είναι να επιλέξουμε το κατάλληλο σύνορο, βάσει του οποίου θα γίνεται ο διαχωρισμός των παρατηρήσεων στην κάθε κλάση, το οποίο καλείται *σύνορο απόφασης (threshold).*

Γενικά, σε κάθε πρόβλημα ταξινόμησης είναι ιδιαίτερα σημαντική η επιλογή του συνόρου απόφασης, καθώς θα πρέπει μέσω αυτού να δίνεται η απαραίτητη βαρύτητα στην κλάση που μας ενδιαφέρει περισσότερο. Στο παρόν πρόβλημα, μεγαλύτερη σημασία για εμάς έχει η σωστή πρόβλεψη της κλάσης “1”, βάσει της οποίας προβλέπεται πότε ένας εργαζόμενος θα παραιτηθεί από την εταιρία. Άρα, αυτό που επιθυμούμε από το μοντέλο μας είναι να μας δίνει όσο το δυνατόν υψηλότερο true positive rate. Η καμπύλη ROC είναι αυτή που μας δίνει μια ταυτόχρονη εικόνα, της μεταβολής του true positive rate και του false positive rate, σε

συνδυασμό με τις μεταβολές των τιμών του threshold. Η απεικόνιση αυτής, μας βοηθά να αποκτήσουμε μια πρώτη αντίληψη σχετικά με το ποιο είναι το ιδανικό διάστημα, μέσα στο οποίο πρέπει να ορίσουμε το σύνορο απόφασης, για το πρόβλημα που εξετάζουμε. Στην προκειμένη περίπτωση, σύμφωνα με την εικόνα 30, βλέπουμε ότι αν ορίσουμε threshold μεταξύ 0,2 και 0,3 εξασφαλίζουμε ένα ικανοποιητικό true positive rate, διατηρώντας ταυτόχρονα το false positive rate σε σχετικά χαμηλά επίπεδα. Παράλληλα, γνωρίζουμε ότι μια συνετή επιλογή είναι να ορίσουμε το threshold ίσο με την συχνότητα εμφάνισης του φαινομένου (event rate). Από τον υπολογισμό λοιπόν, του event rate τόσο στα δεδομένα του training set, όσο και σε αυτά του test set, προκύπτει ότι το event rate ισούται με 0,24, τιμή που εμπίπτει στο διάστημα (0,2, 0,3) που μας υπέδειξε η καμπύλη ROC.



Εικόνα 30. Καμπύλη ROC μοντέλου Logistic Regression

Ορίζουμε λοιπόν, σύνορο απόφασης ίσο με 0,24 και προχωρούμε στην ταξινόμηση των δεδομένων σε κάθε κλάση, ενώ έπειτα αξιολογούμε τα αποτελέσματα αυτής της ταξινόμησης. Σε πρώτο στάδιο η πρόβλεψη της κλάσης και ο υπολογισμός των μέτρων αξιολόγησης του αλγορίθμου γίνεται για τα δεδομένα του training set και έπειτα για τα δεδομένα του test set. Ανάλογα με το πόσο μεγάλες ή μικρές είναι οι αποκλίσεις στα μέτρα αξιολόγησης, μεταξύ του training set και του test set, διαπιστώνουμε αν αντιμετωπίζουμε πρόβλημα overfitting ή όχι αντίστοιχα.

Τα αποτελέσματα των μέτρων αξιολόγησης που προέκυψαν από τα δύο σύνολα δεδομένων παρατίθενται στον πίνακα 9. Με μια πρώτη παρατήρηση των αποτελεσμάτων αυτών, διαπιστώνουμε ότι οι αποστάσεις μεταξύ των μέτρων του training set και του test set είναι μικρές, επομένως καταλήγουμε στο συμπέρασμα ότι δεν αντιμετωπίζουμε πρόβλημα overfitting, οπότε μπορούμε να προχωρήσουμε στην περαιτέρω αξιολόγηση του μοντέλου.

Evaluation Metric	Training set	Test set
Accuracy	0,742	0,744
Precision	0,473	0,476
Recall	0,752	0,768
Specificity	0,739	0,736
F1 Score	0,581	0,588
AUC	0,817	0,826

Πίνακας 7. Μέτρα Αξιολόγησης μοντέλου Logistic Regression

Αξιολογώντας την προβλεπτική ικανότητα του αλγορίθμου στα άγνωστα δεδομένα του test set διαπιστώνουμε ότι το μοντέλο μας ταξινόμησε σωστά το 74,4% των παρατηρήσεων, συνεπώς το ποσοστό λάθους σε αυτό (error rate) ισούται με 100-74,4 = 25,6%, ποσοστό που φανερώνει την μέτρια επίδοση του μοντέλου στα άγνωστα δεδομένα. Ακολούθως, σύμφωνα με το precision διαπιστώνουμε ότι βάσει του συνόλου αυτών που προέβλεψε ότι θα αποχωρήσουν από την εργασία τους, αξιολόγησε ορθά μόλις το 47,6% των υπαλλήλων. Δηλαδή το μοντέλο παρουσιάζει χαμηλή ακρίβεια στην θετική πρόβλεψη. Η συγκεκριμένη κακή επίδοση του μοντέλου υποδηλώνει ταυτόχρονα την ύπαρξη υψηλού false positive rate. Ακολούθως, από το recall και το specificity βλέπουμε ότι προέβλεψε σωστά μόλις το 76,8% του συνόλου των εργαζομένων που πράγματι αποχώρησαν από την εργασία τους και το 73,6% αυτών που πράγματι παραμένουν στην εταιρία, ποσοστά τα οποία δεν είναι ιδιαίτερα ικανοποιητικά. Πολύ κακή επίδοση σημειώνεται και στον δείκτη του F1 Score που ισούται με 58,8%, γεγονός που δικαιολογείται από τις κακές επιδόσεις του precision και του recall, καθώς αποτελεί τον αρμονικό μέσο τους. Τέλος, το AUC που αποτελεί μέτρο της προβλεπτικής ικανότητας του μοντέλου είναι μετρίως ικανοποιητικό, σε σχέση με το άριστο που θα είχαμε αν ισούταν με 1, καθώς στο παρόν μοντέλο ισούται με 0,82. Συνολικά λοιπόν, διαπιστώνουμε ότι το μοντέλο που προκύπτει από τον αλγόριθμο της λογιστικής παλινδρόμησης δεν δίνει ιδιαίτερα αξιόπιστα αποτελέσματα, προκειμένου να γενικευτεί για την πρόβλεψη της κλάσης άλλων άγνωστων δεδομένων.

3.3 ΜΟΝΤΕΛΟ ΤΑΞΙΝΟΜΗΣΗΣ DECISION TREES

Ο επόμενος αλγόριθμος που θα χρησιμοποιήσουμε για την ανάλυση των δεδομένων της βάσης μας είναι αυτός των Δέντρων Απόφασης (Decision Trees). Στο συγκεκριμένο αλγόριθμο δεν απαιτείται να προβούμε σε κάποια μετατροπή των κατηγορικών μεταβλητών σε dummies, ώστε να μπορέσουν να συμπεριληφθούν στην ανάλυσή μας όπως πράξαμε πριν, καθώς ο αλγόριθμος δουλεύει είτε για κατηγορικές είτε για συνεχείς μεταβλητές. Η μορφή της βάσης δεδομένων που χρησιμοποιούμε για την ανάλυση του αλγορίθμου παρατίθεται στην εικόνα 31 και είναι ίδιας μορφής με αυτή που περιγράφηκε στην αρχή του Κεφαλαίου 2. Απαρτίζεται δηλαδή από τις δέκα μεταβλητές που συμπεριλαμβάνονται στη βάση δεδομένων «employees» που

χρησιμοποιούμε. Προς υπενθύμιση όσων αναφέρθηκαν στο Κεφάλαιο 2, το σύνολο δεδομένων των εργαζομένων αποτελείται από τις πέντε αριθμητικές μεταβλητές που αφορούν το επίπεδο ικανοποίησης, τον βαθμό τελευταίας αξιολόγησης, τον αριθμό των projects, τις μέσες μηνιαίες ώρες εργασίας, τα έτη απασχόλησης στην εταιρία, καθώς και από τις πέντε κατηγορικές μεταβλητές που σχετίζονται με την ύπαρξη εργατικού ατυχήματος, την λήψη κάποιας προαγωγής τα τελευταία πέντε χρόνια, τον τομέα απασχόλησης, το μισθολογικό επίπεδο και την ύπαρξη εργασιακής φθοράς.

	satisfaction_level	last_evaluation_rating	projects_worked_on	average_montly_hours	time_spend_company	work_accident	promotion_last_5years	department	salary	Attrition
1	3.8	5.3	3	167	3	0	0	sales	low	1
2	8.0	8.6	6	272	6	0	0	sales	medium	1
3	1.1	8.8	8	282	4	0	0	sales	medium	1
4	3.7	5.2	3	169	3	0	0	sales	low	1
5	4.1	5.0	3	163	3	0	0	sales	low	1
6	1.0	7.7	7	257	4	0	0	sales	low	1
7	9.2	8.5	6	269	5	0	0	sales	low	1
8	8.9	10.0	6	234	5	0	0	sales	low	1
9	4.2	5.3	3	152	3	0	0	sales	low	1
10	1.1	8.1	7	315	4	0	0	sales	low	1

Εικόνα 31. Dataset Employees for Decision Trees

Από την εφαρμογή του αλγορίθμου Decision Trees στα δεδομένα του training set, λαμβάνουμε μια σύνοψη των αποτελεσμάτων του, όπως εμφανίζονται στην εικόνα 32.

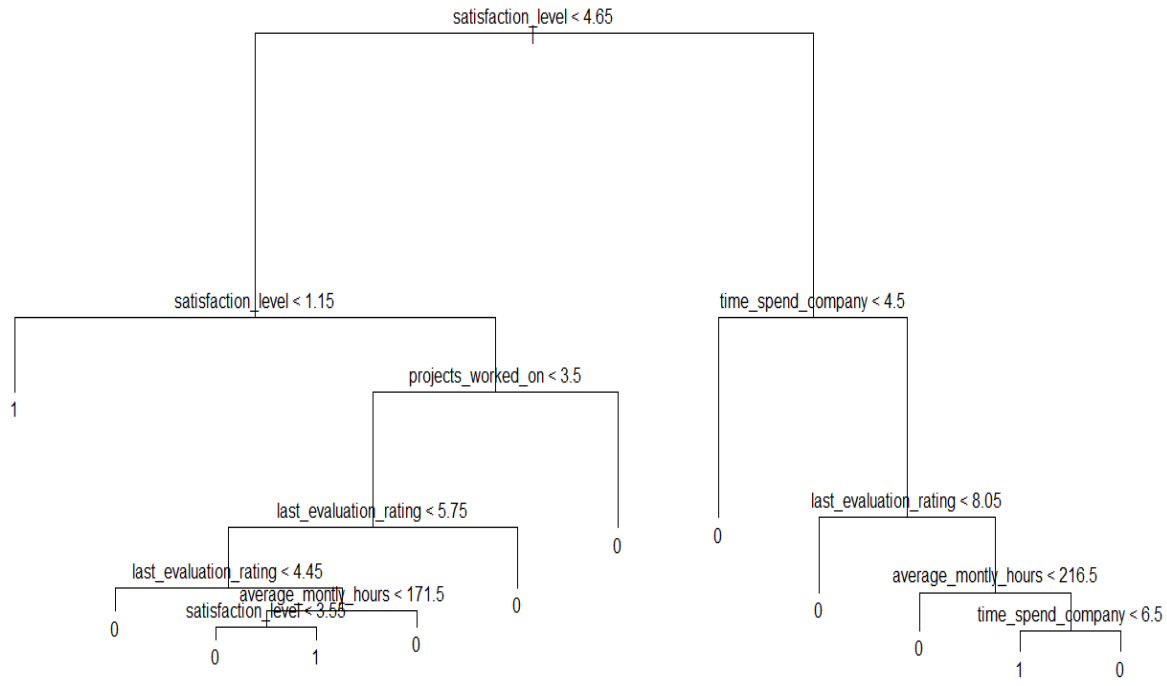
```
Classification tree:
tree(formula = as.factor(Attrition) ~ ., data = training_set)
Variables actually used in tree construction:
[1] "satisfaction_level"     "projects_worked_on"     "last_evaluation_rating"
[4] "average_montly_hours"   "time_spend_company"
Number of terminal nodes:  12
Residual mean deviance:  0.2318 = 4864 / 20990
Misclassification error rate: 0.02762 = 580 / 20998
```

Εικόνα 32. Συνοπτικά αποτελέσματα Decision Tree

Από τα αποτελέσματα αυτά, βλέπουμε ότι στο μοντέλο που προέκυψε από την εκπαίδευση του αλγορίθμου, χρησιμοποιήθηκαν ως κόμβοι απόφασης κατά την προσαρμογή του δέντρου απόφασης, πέντε από τις ανεξάρτητες μεταβλητές. Συγκεκριμένα συμπεριλήφθηκε το επίπεδο ικανοποίησης των εργαζομένων, ο αριθμός των project που αναλαμβάνουν, τα αποτελέσματα της τελευταίας αξιολόγησής τους, οι μέσες μηνιαίες ώρες εργασίας τους και τα έτη απασχόλησής τους στην εταιρία. Βλέπουμε επιπλέον ότι προέκυψαν συνολικά 12 τερματικοί κόμβοι κατά την εκπαίδευση του δέντρου και τέλος διαπιστώνουμε ότι το misclassification error rate, που δείχνει το μέσο ποσοστό ανομοιογένειας που υπάρχει στους τελικούς κόμβους,

δηλαδή το ποσοστό των παρατηρήσεων που ταξινομήθηκαν σε άλλη κλάση από αυτή που πράγματι ανήκουν, ισούται με μόλις 2,76%, ποσοστό που είναι ιδιαίτερα χαμηλό.

Η μορφή του δέντρου που προέκυψε από την εκπαίδευση του αλγορίθμου, φαίνεται στην εικόνα 33. Από την θεωρία, γνωρίζουμε ότι όσο πιο ψηλά βρίσκεται μια ανεξάρτητη μεταβλητή στο δέντρο απόφασης, τόσο μεγαλύτερη η επίδραση που ασκεί, στην διαμόρφωση του αποτελέσματος της εξαρτημένη μεταβλητής (James & et al, 2013). Έτσι, στην προκειμένη περίπτωση το επίπεδο ικανοποίησης των εργαζομένων είναι ο πιο σημαντικός παράγοντας για το attrition, καθώς βρίσκεται στην αρχή του δέντρου, ενώ ταυτόχρονα βρίσκεται και στον ένα εκ των δύο επόμενων κόμβων. Η αμέσως επόμενη πιο σημαντική μεταβλητή βλέπουμε ότι είναι τα έτη απασχόλησης στην εταιρία, ακολούθως το πλήθος των projects που αναλαμβάνει κάθε εργαζόμενος, έπειτα ο βαθμός αξιολόγησης τους και τέλος οι μέσες μηνιαίες ώρες εργασίας τους.



Εικόνα 33. Διαγραμματική Απεικόνιση Δέντρου 12 φύλλων

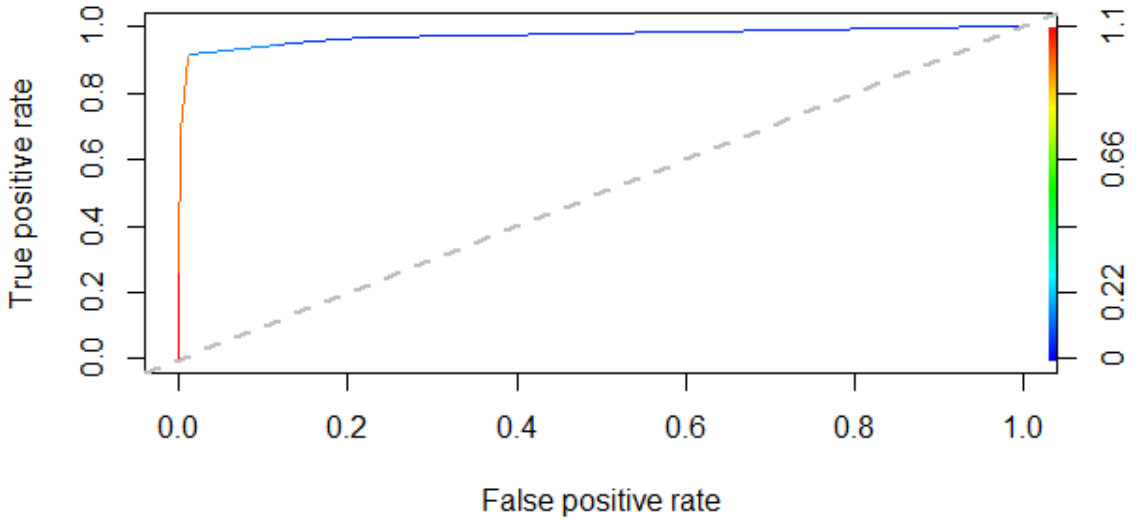
Στην συνέχεια χρησιμοποιώντας το δέντρο που εκπαιδεύτηκε, προχωρούμε στην διαδικασία πρόβλεψης της κλάσης των δεδομένων του training set και έπειτα αυτών του test set. Ύστερα από την πρόβλεψη των κλάσεων καθενός σετ δεδομένων, προχωρούμε στην αξιολόγηση των αποτελεσμάτων αυτών, προκειμένου να εκτιμήσουμε αν η εκπαίδευση του αλγορίθμου δίνει αξιόπιστα αποτελέσματα ή αν χρήζει παραμετροποίησης και επανεκπαίδευσης. Στον πίνακα 10 εμφανίζονται συνοπτικά τα αποτελέσματα των μέτρων αξιολόγησης του κάθε σετ δεδομένων. Με μια πρώτη ματιά βλέπουμε ότι τα αποτελέσματα του training set είναι πολύ υψηλά, ωστόσο το ίδιο υψηλά είναι και τα αποτελέσματα του test set. Οι διαφορές των τιμών τους εντοπίζονται

κυρίως σε επίπεδο ακρίβειας τέταρτου δεκαδικού ψηφίου, οπότε μπορούμε να συμπεράνουμε ότι δεν αντιμετωπίζουμε πρόβλημα overfitting.

Evaluation Metric	Training set	Test set
Accuracy	0,9723	0,9721
Precision	0,9676	0,9679
Recall	0,9146	0,9132
Specificity	0,9904	0,9905
F1 Score	0,9404	0,9397
AUC	0,9704	0,9730

Πίνακας 8. Μέτρα Αξιολόγησης Δέντρου 12 φύλλων

Αξιολογώντας τα αποτελέσματα που προέκυψαν από τον αλγόριθμο στα άγνωστα δεδομένα, βλέπουμε ότι ταξινόμησε στην σωστή κλάση το 97,2% των παρατηρήσεων, ενώ σημείωσε ποσοστό λάθους (error rate) που αντιστοιχεί μόλις στο 2,8%. Επιπλέον, βάσει της τιμής του precision, βλέπουμε ότι το μοντέλο σύμφωνα με το σύνολο αυτών που πρόβλεψε ότι θα αποχωρήσουν, αξιολόγησε σωστά το 96,8% των περιπτώσεων. Δηλαδή παρουσιάζει υψηλή ακρίβεια στην θετική πρόβλεψη. Από το recall και το specificity βλέπουμε ότι το μοντέλο εντοπίζει σωστά το 91,3% των περιπτώσεων που πράγματι παραιτούνται και το 99,1% αυτών που παραμένουν στην εταιρία αντίστοιχα. Το F1 score, που όπως ήδη έχουμε αναφέρει, αποτελεί τον αρμονικό μέσο του precision και του recall, ισούται με 94,0%, ποσοστό που είναι ιδιαίτερα υψηλό, γεγονός που δικαιολογείται από τα υψηλά ποσοστά που σημειώνονται και στα δύο αυτά μέτρα. Τέλος το AUC που αποτελεί το μέτρο που εκφράζει την προβλεπτική ικανότητα του μοντέλου, στην προκειμένη περίπτωση είναι ίσο με 97,3%, ποσοστό ιδιαίτερα υψηλό σε σχέση με το άριστο που θα είχαμε αν ισούταν με 100%. Καθώς όπως φαίνεται και από την καμπύλη ROC στην εικόνα 34, το ποσοστό των true positive που προκύπτουν από το μοντέλο είναι πολύ υψηλό ενώ αντίστοιχα αυτό των false positive είναι αρκετά χαμηλό.



Εικόνα 34. Καμπύλη ROC μοντέλου Decision Trees

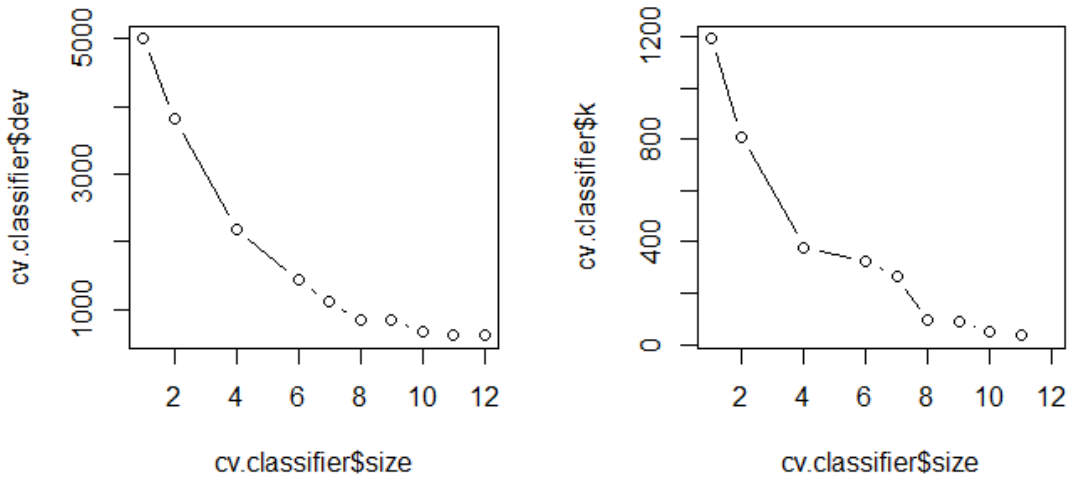
Παρά τα υψηλά ποσοστά προβλεπτικής ικανότητας του συγκεκριμένου μοντέλου, εξετάζουμε το ενδεχόμενο λήψης ακόμη καλύτερων αποτελεσμάτων, μέσω της διαδικασίας «κλαδέματος» του δέντρου. Για τον σκοπό αυτό, χρησιμοποιούμε την μέθοδο του cross-validation, προσδιορίζοντας ταυτόχρονα σε αυτή ότι θέλουμε να καθοδηγείται από το επίπεδο του classification error rate (αντί της μεταβλητότητας), προκειμένου να εξετάσουμε ποιος είναι ο βέλτιστος αριθμός των καταληκτικών κόμβων του δέντρου (Kohavi, 1995).

```
$`size`
 [1] 12 11 10  9  8  7  6  4  2  1

$dev
 [1]  605  629  677  831  840 1105 1430 2185 3807 4999

$k
 [1]   -Inf   33.0   49.0   86.0   92.0  265.0  325.0  377.5  811.0 1192.0

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Εικόνα 35. Αποτελέσματα μεθόδου Cross-Validation

Η εικόνα 35, μας δείχνει πως διαμορφώνονται οι τιμές των διαφόρων παραμέτρων, καθώς μεταβάλλεται ο αριθμός των φύλλων του δέντρου, όπως αυτές προέκυψαν από την χρήση της μέθοδο cross-validation. Συγκεκριμένα, η παράμετρος size αναφέρεται στο σύνολο των φύλλων του δέντρου, η dev στο cross-validation error rate και η *k* στον δείκτη κόστους-πολυπλοκότητας, δηλαδή στην «ποινή» που θα έχουμε, αν επιλέξουμε μεγαλύτερο δέντρο. Για

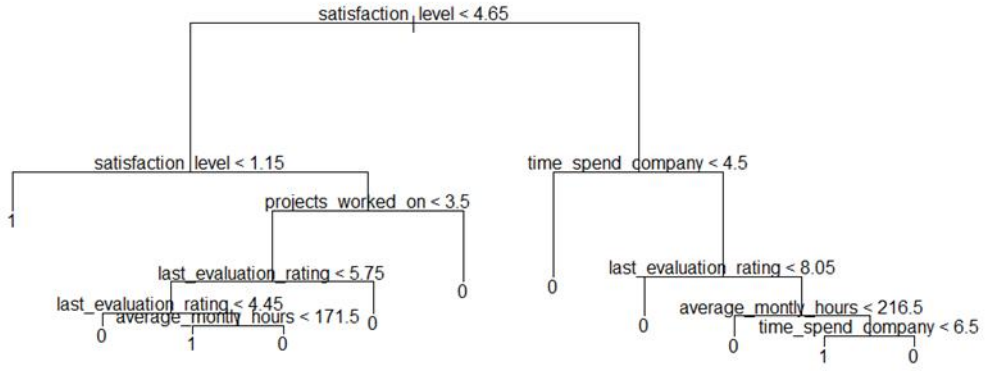
να αποκτήσουμε καλύτερη αντίληψη για τα αποτελέσματα της εικόνας 35 και να κρίνουμε αν το δέντρο μας χρειάζεται «κλάδεμα», προχωράμε στην διαγραμματική απεικόνιση των error rates σε συνάρτηση με το πλήθος των καταληκτικών κόμβων και του δείκτη κόστους-πολυπλοκότητας σε συνάρτηση με το πλήθος των καταληκτικών κόμβων.



Εικόνα 36. Διαγραμματική απεικόνιση αποτελεσμάτων Cross-Validation

Από τα αποτελέσματα των εικόνων 35 και 36 βλέπουμε ότι η χαμηλότερη τιμή στο cross-validation error rate και στον δείκτη κόστους-πολυπλοκότητας, σημειώνεται στην περίπτωση των 12 κόμβων. Δεδομένου ότι το αρχικό μας δέντρο «αναπτύχθηκε» και κατέληξε σε 12 τερματικούς κόμβους, συνεπάγεται ότι δεν είναι απαραίτητο να προβούμε σε κάποιο «κλάδεμα», καθώς οι 12 καταληκτικοί κόμβοι είναι αυτοί που μας εξασφαλίζουν το χαμηλότερο επίπεδο error rate. Προχωρούμε ωστόσο, στον έλεγχο των αποτελεσμάτων που προκύπτουν από την επιλογή 11 και 10 φύλλων, προκειμένου να επιβεβαιώσουμε την ορθότητα του συγκεκριμένου συμπεράσματος. Αν τα αποτελέσματα που θα προκύψουν από το κλάδεμα του δέντρου χειροτερέψουν, τότε θα καταλήξουμε στο αρχικό δέντρο με τα 12 φύλλα, ενώ αν βελτιωθούν θα επιλέξουμε αυτό που δίνει τα καλύτερα αποτελέσματα.

Το δέντρο που προκύπτει από την επιλογή 11 τερματικών κόμβων παρατίθεται στην εικόνα 37 και τα αποτελέσματα των μέτρων αξιολόγησής του, σημειώνονται συνοπτικά στον πίνακα 11.



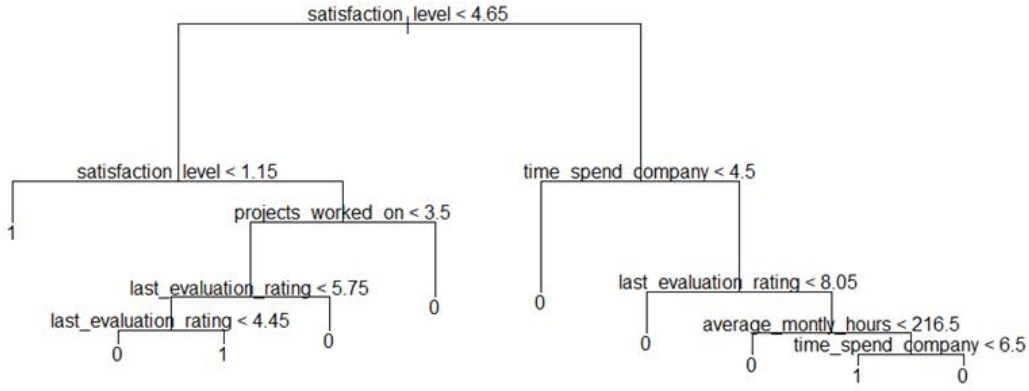
Εικόνα 37. Διαγραμματική Απεικόνιση δέντρου 11 φύλλων

Evaluation Metric	Training set	Test set
Accuracy	0,971	0,971
Precision	0,961	0,960
Recall	0,915	0,915
Specificity	0,988	0,988
F1 Score	0,937	0,937
AUC	0,969	0,972

Πίνακας 9. Μέτρα Αξιολόγησης Δέντρου 11 φύλλων

Συγκρίνοντας τα αποτελέσματα των μέτρων αξιολόγησης των δύο δέντρων, διαπιστώνουμε ότι οι διαφορές μεταξύ τους είναι μηδαμινές, με τα αποτελέσματα του δέντρου με τα 12 φύλλα να είναι στην πλειοψηφία τους, υψηλότερα από αυτά του δέντρου με τα 11 φύλλα. Μόνη εξαίρεση αποτελεί το μέτρο του recall, το οποίο αυξάνεται κατ' ελάχιστο στο δεύτερο δέντρο, γεγονός που υποδηλώνει ότι το δέντρο με τα 11 φύλλα, εντοπίζει οριακά περισσότερούς υπαλλήλους, που αποχωρούν από την εταιρία.

Ακολούθως το δέντρο απόφασης που προκύπτει από την επιλογή 10 φύλλων έχει την μορφή που παρατίθεται στην εικόνα 38, ενώ τα μέτρα αξιολόγησής του, καταγράφονται στον πίνακα 12 παρακάτω.



Εικόνα 38. Διαγραμματική Απεικόνιση δέντρου 10 φύλλων

Evaluation Metric	Training set	Test set
Accuracy	0,968	0,969
Precision	0,949	0,953
Recall	0,917	0,916
Specificity	0,988	0,985
F1 Score	0,933	0,934
AUC	0,968	0,971

Πίνακας 10. Μέτρα Αξιολόγησης Δέντρου 10 φύλλων

Παρατηρούμε ότι αντίστοιχα αποτελέσματα με πριν, σημειώνονται και στην περίπτωση του δέντρου με τα 10 φύλλα. Στην πλειοψηφία των μέτρων σημειώνεται και πάλι, ελαφρώς χαμηλότερη επίδοση από αυτή που είχαμε στο αρχικό δέντρο, με εξαίρεση το recall το οποίο αυξάνεται οριακά.

Στον πίνακα 13 παρατίθενται συγκεντρωτικά ορισμένα από τα αποτελέσματα των μέτρων αξιολόγησης που προέκυψαν από τις παραπάνω δοκιμές. Συγκεκριμένα, καταγράφονται τα μέτρα του AUC, του accuracy και του recall, όπως αυτά καταγράφηκαν από την χρήση τους στα δεδομένα του training set και του test set.

Nodes	Set	AUC	Accuracy	Recall
12	**Training**	**0,970**	**0,972**	**0,915**
	Test	0,973	0,972	0,913
11	**Training**	**0,969**	**0,971**	**0,915**
	Test	0,972	0,971	0,915
10	**Training**	**0,968**	**0,968**	**0,917**
	Test	0,971	0,969	0,916

Πίνακας 11. Συνοπτικά αποτελέσματα δοκιμών Decision Tree

Συμπερασματικά λοιπόν, μπορούμε να επισημάνουμε ότι την υψηλότερη τιμή στο μέτρο AUC σημειώνει το δέντρο με τους 12 τερματικούς κόμβους, ενώ όσο μειώνουμε τον αριθμό των φύλλων, μειώνεται και αυτό κατ' ελάχιστο. Το ίδιο ακριβώς συμβαίνει και με τα αποτελέσματα της συνολικής ακρίβειας (accuracy) του μοντέλου, η οποία επίσης μειώνεται σταδιακά με την μείωση των φύλλων. Εξαίρεση όπως είδαμε και παραπάνω αποτελεί το μέτρο του recall, το οποίο αυξάνεται σε μικρό βαθμό, καθώς μειώνουμε τα φύλλα του δέντρου. Αυτό υποδηλώνει ότι το μοντέλο, δίνει καλύτερη πρόβλεψη για το σύνολο των εργαζομένων που αποχωρούν από την εταιρία, καθώς μειώνεται ο αριθμός των τερματικών κόμβων, δηλαδή ο αριθμός των False Negative συστηματικά μειώνεται. Σύμφωνα με όσα προαναφέρθηκαν καταλήγουμε στο συμπέρασμα ότι το αρχικό δέντρο που εκπαιδεύτηκε, με τους 12 τερματικούς κόμβους είναι αυτό που δίνει συνολικά τα καλύτερα αποτελέσματα πρόβλεψης, οπότε είναι αυτό που επιλέγουμε για την μελλοντική πρόβλεψη άλλων άγνωστων δεδομένων.

Στην εικόνα 39 παρουσιάζονται τα αναλυτικά στοιχεία του δέντρου των 12 φύλλων, όπως αυτά προέκυψαν από την αρχική εκπαίδευση του αλγορίθμου. Σε αυτήν βλέπουμε το σύνολο των κλαδιών του δέντρου που διαμορφώθηκαν, τα κριτήρια βάσει των οποίων γίνεται ο διαχωρισμός των δεδομένων σε κάθε κόμβο απόφασης, το σύνολο των δεδομένων που κατέληξαν σε κάθε κόμβο, την απόκλιση των δεδομένων που υπάρχει σε καθένα από αυτούς, την κυρίαρχη κλάση που προβλέπεται σε κάθε κλαδί ("0" ή "1") και το ποσοστό των παρατηρήσεων που λαμβάνουν την τιμή "0" και "1" αντίστοιχα. Τέλος, οι αστερίσκοι είναι αυτοί που υποδηλώνουν ποιοι είναι οι τερματικοί κόμβοι του δέντρου.

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

  1) root 20998 23050.00 0 ( 0.76193 0.23807 )
    2) satisfaction_level < 4.65 5856  7874.00 1 ( 0.39822 0.60178 )
      4) satisfaction_level < 1.15 1213     0.00 1 ( 0.00000 1.00000 ) *
      5) satisfaction_level > 1.15 4643  6436.00 0 ( 0.50226 0.49774 )
       10) projects_worked_on < 3.5 2725  2774.00 1 ( 0.20624 0.79376 )
         20) last_evaluation_rating < 5.75 2362  1454.00 1 ( 0.09229 0.90771 )
           40) last_evaluation_rating < 4.45 86     0.00 0 ( 1.00000 0.00000 ) *
           41) last_evaluation_rating > 4.45 2276  1008.00 1 ( 0.05800 0.94200 )
             82) average_montly_hours < 171.5 2207   641.30 1 ( 0.03308 0.96692 )
              164) satisfaction_level < 3.55 37    15.56 0 ( 0.94595 0.05405 ) *
              165) satisfaction_level > 3.55 2170   382.70 1 ( 0.01751 0.98249 ) *
             83) average_montly_hours > 171.5 69    57.11 0 ( 0.85507 0.14493 ) *
         21) last_evaluation_rating > 5.75 363   149.10 0 ( 0.94766 0.05234 ) *
       11) projects_worked_on > 3.5 1918  1043.00 0 ( 0.92284 0.07716 ) *
    3) satisfaction_level > 4.65 15142  9671.00 0 ( 0.90259 0.09741 )
      6) time_spend_company < 4.5 12340  1938.00 0 ( 0.98485 0.01515 ) *
      7) time_spend_company > 4.5 2802  3866.00 0 ( 0.54033 0.45967 )
       14) last_evaluation_rating < 8.05 1061   340.70 0 ( 0.96230 0.03770 ) *
       15) last_evaluation_rating > 8.05 1741  2075.00 1 ( 0.28317 0.71683 )
         30) average_montly_hours < 216.5 307   153.20 0 ( 0.93160 0.06840 ) *
         31) average_montly_hours > 216.5 1434  1184.00 1 ( 0.14435 0.85565 )
           62) time_spend_company < 6.5 1342   785.00 1 ( 0.08569 0.91431 ) *
           63) time_spend_company > 6.5 92     0.00 0 ( 1.00000 0.00000 ) *
```

Εικόνα 39. Αποτελέσματα Εκτύπωσης Ταξινομητή Decision Tree

Συνδυαστικά με την εικόνα 33 στην οποία παρατέθηκε η διαγραμματική απεικόνιση του δέντρου, διαπιστώνουμε ότι στον πρώτο κόμβο απόφασης, που επισημαίνεται με τον αριθμό 2 στην εικόνα 39, το κριτήριο με το οποίο γίνεται ο διαχωρισμός των δεδομένων είναι το satisfaction_level < 4,65 και η κυρίαρχη κλάση σε αυτόν είναι το 1, καθώς όπως επισημαίνεται το 39,8% των παρατηρήσεων λαμβάνει την τιμή 0 ενώ το 60,2% την τιμή 1. Ωστόσο, στην περίπτωση μη ικανοποίησης του συγκεκριμένου κριτηρίου, εξασφαλίζεται καλύτερος διαχωρισμός των δεδομένων, καθώς όπως φαίνεται στο στοιχείο 3 της είκόνας 39, το 90,3% των παρατηρήσεων λαμβάνει την τιμή 0, ενώ το υπόλοιπο 9,7% λαμβάνει την τιμή 1. Άρα η κυρίαρχη κλάση στον συγκεκριμένο κόμβο είναι το 0. Ο αμέσως επόμενος διαχωρισμός των δεδομένων, γίνεται επίσης βάσει του satisfaction_level, αλλά για τιμές μικρότερες του 1,15. Βάσει αυτού, η κυρίαρχη κλάση είναι το 1, καθώς βλέπουμε ότι το 100% των παρατηρήσεων που καταλήγουν σε αυτόν, λαμβάνει την συγκεκριμένη τιμή. Ταυτόχρονα ο συγκεκριμένος κόμβος βλέπουμε ότι είναι τερματικός. Αντίθετα, στην περίπτωση μη ικανοποίησης της παραπάνω ανισότητας, τα ποσοστά διαμορφώνονται στο 50,2% και 49,8% για την κλάση 0 και 1 αντίστοιχα. Στο ίδιο επίπεδο, ο άλλος κόμβος απόφασης έχει κριτήριο για τον διαχωρισμό των δεδομένων, την μεταβλητή time_spend_company για τιμές μικρότερες του 4,5. Σε αυτόν, η κυρίαρχη κλάση είναι το 0, καθώς το 98,4% των παρατηρήσεων που εισήλθαν σε αυτόν λαμβάνουν την συγκεκριμένη τιμή, ενώ μόλις το 1,6% λαμβάνει την τιμή 1. Ο συγκεκριμένος κόμβος είναι επίσης τερματικός, όπως επισημαίνεται στο αποτέλεσμα με τον αριθμό 6 της εικόνας 39. Αντιθετα στην περίπτωση που δεν ικανονοποιείται το παραπάνω κριτήριο η κυρίαρχη κλάση είναι και πάλι το 0, ενώ τώρα ο διαχωρισμός των παρατηρήσεων γινεται σε ποσοστά 54,0% και 46,0% υπέρ της κλάσης 0. Ομοίως επεξηγούνται και τα αποτελέσματα των υπόλοιπων κόμβων απόφασης που ακολουθούν. Συνοπτικά αναφέρουμε ότι στο αμέσως επόμενο επίπεδο, το κριτήριο διαχωρισμού των δεδομένων είναι το projects_worked_on < 3,5 με κυρίαρχη κλάση το 1. Ακολούθως στα δύο επόμενα κλαδιά, το κριτήριο διαχωρισμού είναι το last_evaluation_rating από την μια για τιμές μικρότερες του 5,75 και από την άλλη για τιμές μικρότερες του 8,05. Σε αυτά η επικρατούσα τιμή της κλάσης είναι το 1 και το 0 αντίστοιχα και οι κόμβοι είναι τερματικοί. Έπειτα, ακολουθούν οι κόμβοι απόφασης με κριτήρια διαχωρισμού το last_evaluation_rating < 4,45 και το average_monthly_hours < 216,5 οι οποίοι έχουν κυρίαρχη κλάση το 0 και είναι πίσης τερματικοί. Στο προτελευταίο επίπεδο έχουμε κριτήριο διαχωρίσιμου το average_monthly_hours < 171,5 με κυρίαρχη κλάση το 1. Στο τελευταίο επίπεδο του δέντρου, οι κόμβοι απόφασης είναι επίσης τερματικοί και έχουν κριτήρια διαχωρισμού το satisfaction_level < 3,55 στο οποίο η κυρίαρχη κλάση είναι το 0 και το time_spend_company < 6,5 με κυρίαρχη κλάση το 1.

3.4 ΜΟΝΤΕΛΟ ΤΑΞΙΝΟΜΗΣΗΣ RANDOM FOREST

Ο τρίτος αλγόριθμος που θα χρησιμοποιήσουμε για την ανάλυση των δεδομένων μας, είναι αυτός των Τυχαίων Δασών (Random Forest), ο οποίος είναι παρόμοιας λογικής αυτής των δέντρων απόφασης, καθώς όπως προαναφέρθηκε στο Κεφάλαιο 1, προκύπτει από την ανάπτυξη πολλαπλών παράλληλων δέντρων. Αντίστοιχα λοιπόν, η δομή των δεδομένων μας παραμένει ίδια με αυτή της εικόνας 31 που παρατέθηκε στην προηγούμενη υποενότητα, καθώς δεν απαιτείται να πραγματοποιηθεί κάποια προ-επεξεργασία των δεδομένων. Εφαρμόζουμε λοιπόν, τον αλγόριθμο Random Forest στα δεδομένα εκπαίδευση και λαμβάνουμε τα αποτελέσματα της εικόνας 40.

```
Call:
 randomForest(formula = Attrition ~ ., data = training_set, type = classification,
 importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.97%
Confusion matrix:
      0    1 class.error
0 15960   39 0.002437652
1   164 4835 0.032806561
```

Εικόνα 40. Εκτύπωση αρχικού μοντέλου Random Forest

Βάσει αυτής, διαπιστώνουμε ότι ο αριθμός των δέντρων που αναπτύχθηκε, κατά την εφαρμογή του αλγορίθμου στα δεδομένα εκπαίδευσης ισούται με 500. Επίσης, βλέπουμε ότι στα δέντρα αυτά κάθε κόμβος απόφασης, προέκυψε από την αξιολόγηση τριών ανεξάρτητων μεταβλητών κάθε φορά, ενώ το "Out of Bag error rate" ή OOB error rate ισούται με μόλις 0,97%, γεγονός που υποδηλώνει την πολύ καλή προσαρμογή του αλγορίθμου στα δεδομένα εκπαίδευσης. Τέλος, μπορούμε να επισημάνουμε ότι το μοντέλο μας, πραγματοποιεί καλύτερη πρόβλεψη για την κλάση 0, παρά για την κλάση 1, όπως υποδηλώνεται από τα classification error rates, που σημειώνονται επίσης στην εικόνα 40.

Στο σημείο αυτό, είναι χρήσιμο να επισημάνουμε ότι το OOB error rate αναφέρεται στην εκτίμηση του μέσου όρου των σφαλμάτων, που προκύπτει από το σύνολο των 500 δέντρων που εκπαιδεύτηκαν (James & et al, 2013). Αναλυτικότερα, σε συνέχεια όσων αναφέρθηκαν στην θεωρία του αλγορίθμου που προηγήθηκε στο Κεφάλαιο 1, η Random Forest χρησιμοποιεί επαναληπτικά ένα διαφορετικό υποσύνολο, του συνόλου δεδομένων που εισάγονται σε αυτή, προκειμένου να κατασκευάσει τα πολλαπλά δέντρα αποφάσεων. Έπειτα, μετά την δημιουργία του κάθε δέντρου, προχωρά στην ταξινόμηση των δεδομένων που δε συμμετείχαν ή ήταν "out of bag" κατά τη δημιουργία του δέντρου. Τέλος, από το σύνολο αυτών των ταξινομήσεων υπολογίζει τον μέσο όρο των σφαλμάτων που προέκυψαν, δηλαδή το "Out of Bag error rate" ή

OOB error rate, που αναφέρεται στο ποσοστό των δεδομένων που ταξινομήθηκαν σε άλλη κλάση από αυτή που πράγματι ανήκαν.

Τα συγκεκριμένα αποτελέσματα ωστόσο, προέκυψαν από τις default τιμές των παραμέτρων του αλγορίθμου. Για τον λόγο αυτό προχωρούμε στην επεξεργασία των διαφόρων παραμέτρων του, προκειμένου να εξασφαλίσουμε, αν υπάρχει, ένα ακόμη καλύτερο μοντέλο, το οποίο θα απαντά καλύτερα και πιο αξιόπιστα στο πρόβλημα που εξετάζουμε. Κατά την διαδικασία αυτή θα χρησιμοποιήσουμε ορισμένα μόνο από τα μέτρα αξιολόγησης των μοντέλων, προκειμένου να είναι πιο άμεση και εύκολη η σύγκριση των αποτελεσμάτων κάθε δοκιμής. Συγκεκριμένα θα εστιάσουμε στο OOB error rate, που προκύπτει από την διαδικασία εκπαίδευσης του κάθε μοντέλου, το AUC για το training set και το AUC για το test set. Επιλέγουμε να εστιάσουμε στο AUC, διότι αποτελεί ένα από τα σημαντικότερα μέτρα αξιολόγησης του μοντέλου μας, καθώς μετρά το trade-off μεταξύ του true positive rate (recall) και του false positive rate (1-specificity). Έτσι όσο πιο κοντά στο 1 είναι η τιμή του, τόσο υψηλότερη η προβλεπτική ικανότητα του μοντέλου (Huang & Ling, 2005).

Στο πρόβλημα που επιλύουμε έχει μεγαλύτερη σημασία για εμάς να εντοπίσουμε όσο το δυνατόν ακριβέστερα την κλάση 1, δηλαδή τους υπαλλήλους που παραιτούνται. Για τον λόγο αυτό ξεκινάμε με τον ορισμό του συνόρου απόφασης ίσο με 0,24. Προς υπενθύμιση όσων αναφέρθηκαν στην υποενότητα 3.1, το 0,24 αντιστοιχεί στο event rate, δηλαδή στο ποσοστό των παραιτήσεων που έχουν σημειωθεί ιστορικά στο σύνολο των δεδομένων μας.

```
Call:
 randomForest(formula = as.factor(Attrition) ~ ., data = training_set,
 type = classification, importance = TRUE, cutoff = c(0.76,          0.24))

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 1.08%
Confusion matrix:
      0    1 class.error
0 15868  131 0.008188012
1    96 4903 0.019203841
```

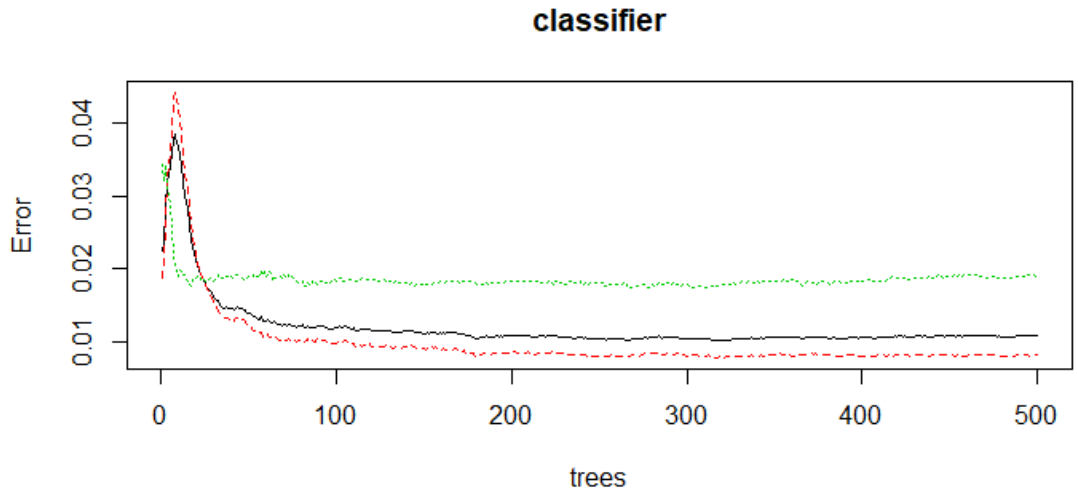
Εικόνα 41. Εκτύπωση μοντέλου Random Forest για threshold=0.24

Εφαρμόζουμε εκ νέου τον αλγόριθμο στα δεδομένα του training set από τον οποίο λαμβάνουμε τα αποτελέσματα της εικόνας 41. Σε σχέση με αυτά της εικόνας 40, βλέπουμε ότι το OOB error rate αυξάνεται κατά ένα μικρό ποσοστό από 0,97% σε 1,08%. Την ίδια στιγμή όμως, τα αποτελέσματα της πρόβλεψης για την κλάση 1 βελτιώνονται, καθώς μειώνεται όπως βλέπουμε το classification error rate της συγκεκριμένης κλάσης από 0,033 που ήταν αρχικά σε 0,019, ενώ παράλληλα τα AUC τόσο του training set όσο και του test set, σημειώνουν πολύ υψηλά νούμερα 0,9999 και 0,9985 αντίστοιχα, όπως βλέπουμε στο στοιχείο 2 του πίνακα 15. Ωστόσο,

ανεξαρτήτως των καλύτερων αποτελεσμάτων που προέκυψαν για threshold = 0,24, σε καμία περίπτωση δεν θα επιλέγαμε την ανάλυση του μοντέλου με threshold = 0,5 και αυτό γιατί θέλουμε να αποβάλλουμε από το μοντέλο μας τον παράγοντα της τυχαιότητας.

Δεδομένου ότι από προεπιλογή ο αλγόριθμος προχωρά στην εκπαίδευση 500 δέντρων, πλήθος το οποίο είναι πολύ μεγάλο, στο στάδιο αυτό θα προσπαθήσουμε να περιορίσουμε το πλήθος αυτό, σε κάποιο μικρότερο αριθμό, προσπαθώντας παράλληλα να διατηρήσουμε την προβλεπτική ικανότητα του μοντέλου στα ίδια επίπεδα ή σε ελαφρώς χαμηλότερα. Για τον σκοπό αυτό, εξετάζουμε το πως εξελίσσεται το classification error rate του ταξινομητή, που κατασκευάσαμε προηγουμένως για threshold = 0,24, σε συνάρτηση με το πλήθος των δέντρων που χρησιμοποιούνται κάθε φορά στην εκπαίδευση της Random Forest.

Η διαγραμματική απεικόνιση των δύο αυτών μεγεθών παρατίθεται στην εικόνα 42. Η μαύρη γραμμή στη συγκεκριμένη εικόνα, δείχνει πως μειώνεται το Out-Of-Bag error rate σύμφωνα με την αύξηση του αριθμού των δέντρων, ενώ αντίστοιχα οι χρωματιστές γραμμές δείχνουν πως μειώνεται το classification error rate της κάθε κλάσης. Συγκεκριμένα η πράσινη γραμμή δείχνει το classification error rate της κλάσης 1, ενώ η κόκκινη της κλάσης 0. Παρατηρούμε ότι τα errors μειώνονται σημαντικά έως τα 70-80 δέντρα, ενώ από τα 150 και έπειτα βλέπουμε ότι δεν μειώνονται περαιτέρω και αρχίζουν να αποκτούν μια πιο σταθερή τάση. Δεδομένου ότι από ένα πλήθος δέντρων και έπειτα, η επιπλέον πληροφορία που λαμβάνουμε από το μοντέλο, δεν αυξάνεται περαιτέρω ή αυξάνεται κατ' ελάχιστο, προχωρούμε στη διαδοχική μείωση του πλήθους αυτών, ώστε να εντοπίσουμε το πλήθος που μας εξασφαλίζει καλύτερα αποτελέσματα πρόβλεψης ή παρόμοια με αυτά που είχαμε από την default επιλογή των 500 δέντρων.



Εικόνα 42. Διαγραμματικη απεικόνιση Error rates συναρτήσει πλήθους δέντρων

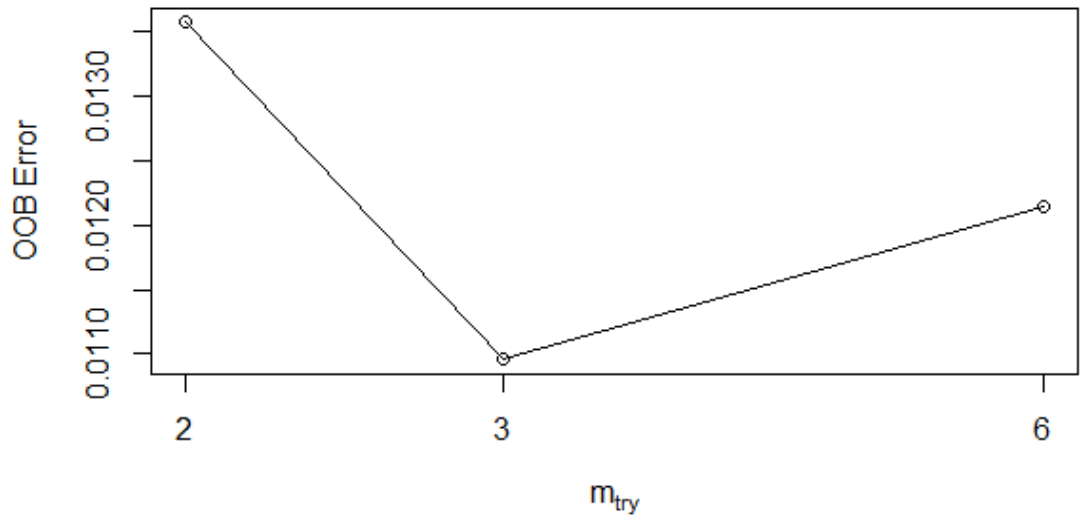
Για τον σκοπό αυτό ορίζουμε διαδοχικά τις τιμές 300, 280, 250, 200, 180 και 150 μέσω της επιπρόσθετης παραμέτρου που καθορίζουμε κατά την εκπαίδευση του αλγορίθμου, της ntree. Σε κάθε μια από αυτές τις δοκιμές υπολογίζουμε ταυτόχρονα το AUC των δυο ξεχωριστών set. Τα αποτελέσματα αυτών των δοκιμών συγκεντρώνονται στα στοιχεία 3 έως 8 του πίνακα 15. Από την μελέτη αυτών καταλήγουμε στο συμπέρασμα ότι η επιλογή 300 δέντρων μας δίνει το χαμηλότερο OOB error rate = 1.04% σε σχέση με τις άλλες δοκιμές και μια από τις υψηλότερες τιμές για το AUC στα άγνωστα δεδομένα του test set. Ταυτόχρονα, από την επιλογή των 300 δέντρα εξασφαλίζουμε χαμηλότερο OOB error rate, σε σχέση με αυτό του προηγούμενου μοντέλου των 500 δέντρων, ενώ το AUC του test set παραμένει στα ίδια υψηλά επίπεδα με αυτά του προηγούμενου μοντέλου.

Επιπλέον, από την εκτύπωση των αποτελεσμάτων του ταξινομητή, που προέκυψε από τα 300 δέντρα, διαπιστώνουμε ότι πετυχαίνουμε περαιτέρω μείωση των False Negative προβλέψεων του μοντέλου, όπως φαίνεται στην εικόνα 43. Δηλαδή, εξασφαλίζουμε περαιτέρω αύξηση της προβλεπτικής ικανότητας του μοντέλου για την κλάση 1, που αποτελεί και την κλάση μεγαλύτερου ενδιαφέροντος μας.

```
Call:
 randomForest(formula = as.factor(Attrition) ~ ., data = training_set,
 type = classification, importance = TRUE, cutoff = c(0.76,          0.24),
 ntree = 300)
               Type of random forest: classification
                     Number of trees: 300
No. of variables tried at each split: 3

        OOB estimate of  error rate: 1.04%
Confusion matrix:
      0    1 class.error
0 15869  130 0.008125508
1    89 4910 0.017803561
```

Εικόνα 43. Εκτύπωση μοντέλου Random Forest για ntree=300

Ακολούθως, θεωρώντας δεδομένη την επιλογή των 300 δέντρων, προχωρούμε στον έλεγχο του υποσυνόλου των ανεξάρτητων μεταβλητών mtry, που πρέπει να χρησιμοποιεί ο αλγόριθμος, για την διενέργεια κάθε split. Από προεπιλογή ο αλγόριθμος καθορίζει mtry = $\sqrt{9}$ = 3, όπως αναφέρθηκε στο Κεφάλαιο 1, όπου το 9 αποτελεί το σύνολο των ανεξάρτητων μεταβλητών του dataset «employees». Για να εντοπίσουμε λοιπόν, τον ιδανικό αριθμό της παραμέτρου mtry στο πρόβλημα που επιλύουμε, προχωρούμε και πάλι στην διαγραμματική απεικόνιση του OOB error συναρτήσει του mtry αυτή τη φορά, που παρατίθεται στην εικόνα 44.



Εικόνα 44. Διαγραμματική απεικόνιση OOB error συναρτήσει mtry

Βάσει αυτής βλέπουμε ότι για mtry = 3, όπως ορίστηκε εξαρχής από τον αλγόριθμο, λαμβάνουμε τα χαμηλότερα αποτελέσματα του OOB error. Ωστόσο για να εξετάσουμε την ορθότητα της επιλογής αυτής, ελέγχουμε την αλλαγή που θα επέλθει στην προβλεπτική ικανότητα του μοντέλου μας, αν αλλάξουμε τον αριθμό του mtry σε 4 και 6 αντίστοιχα, διατηρώντας τις άλλες παραμέτρους, που ήδη έχουν οριστεί σταθερές. Τα αποτελέσματα που λαμβάνουμε από την αξιολόγηση των δύο νέων μοντέλων που προέκυψαν, εμφανίζονται στα στοιχεία 9 και 10 του πίνακα 15. Παρατηρούμε ότι και στις δύο περιπτώσεις το OOB error αυξάνεται, σε 1,07% και 1,27% αντίστοιχα, και το AUC στα δεδομένα του test set χειροτερεύει, καθώς μειώνεται σε 0,9981 και 0,9982 αντίστοιχα. Επομένως, καταλήγουμε στο συμπέρασμα ότι η αρχική επιλογή του αλγορίθμου, όπου mtry = 3 είναι αυτή, που μας δίνει τα καλύτερα αποτελέσματα πρόβλεψης των δύο κλάσεων.

Επιπρόσθετα, στο στάδιο αυτό προχωρούμε σε νέες δοκιμές του αλγορίθμου, όπου τώρα, ταυτόχρονα με την παράμετρο mtry επηρεάζουμε και την παράμετρο ntree. Σκοπός μας είναι να εξετάσουμε το ενδεχόμενο ύπαρξης ενός νέου μοντέλου, υψηλότερης προβλεπτικής ικανότητας από αυτό στο οποίο έχουμε καταλήξει προς το παρόν. Στα στοιχεία 11 και 12 του πίνακα 15 εξετάζουμε τις περιπτώσεις, όπου η παράμετρος ntree ισούται με 280 και η mtry με 4 και 6 αντίστοιχα. Σε αυτές το OOB error ισούται με 1,09% και 1,26% αντίστοιχα, ενώ το AUC του test set με 0,9981 και 0,9982 αντίστοιχα. Διαπιστώνουμε λοιπόν, ότι και στις δύο περιπτώσεις, τα αποτελέσματα που προκύπτουν είναι χειρότερα από αυτά που προέκυψαν για ntree = 300 mtry = 3. Αντίστοιχα, στα στοιχεία 13 και 14 του πίνακα 15 ελέγχουμε τα αποτελέσματα που προκύπτουν για ntree ίσο με 150 και mtry ίσο με 4 και 6 αντίστοιχα. Από την μελέτη των

αποτελεσμάτων των συγκεκριμένων δοκιμών, καταλήγουμε στο ίδιο συμπέρασμα με πριν, καθώς και πάλι τα μοντέλα που προέκυψαν έχουν χαμηλότερη προβλεπτική ικανότητα.

Τέλος, μια ακόμη παράμετρος που εξετάζουμε είναι αυτή του nodesize βάσει της οποίας προσδιορίζεται ο ελάχιστος αριθμός παρατηρήσεων, που θα πρέπει να έχει κάθε καταληκτικός κόμβος. Η default τιμή του αλγορίθμου για την συγκεκριμένη παράμετρο είναι το 1, ωστόσο εξετάζουμε τα αποτελέσματα που θα λάβουμε, αλλάζοντας την τιμή αυτή σε 5 και 3 αντίστοιχα. Ταυτόχρονα κρατάμε σταθερές τις υπόλοιπες παραμέτρους, στις οποίες καταλήξαμε προηγουμένως, δηλαδή ntree = 300 και mtry = 3. Από τα στοιχεία 15 και 16 του πίνακα 15 διαπιστώνουμε και πάλι ότι τα αποτελέσματα χειροτερεύουν, καθώς το OOB error αυξάνεται σημαντικά, ενώ αντίστοιχα το AUC των άγνωστων δεδομένων μειώνεται επίσης σημαντικά. Άρα καλύτερο μοντέλο έως τώρα είναι αυτό, που προκύπτει από την επιλογή ntree = 300, mtry = 3 και nodsize = 1.

Στο ίδιο συμπέρασμα καταλήγουμε αν πέρα από την μεταβολή του nodesize σε 5 και 3 αντίστοιχα, αλλάξουμε και το ntree σε 280, καθώς και πάλι τα αποτελέσματα της πρόβλεψης μας χειροτερεύουν, όπως φαίνεται στα στοιχεία 18 και 19 του πίνακα 15. Αντίστοιχο συμπέρασμα προκύπτει και για τα στοιχεία 22 και 23 του ίδιου πίνακα, στα οποία επιλέγουμε ntree = 150 και nodesize ίσο με 5 και 3 αντίστοιχα. Τέλος στα στοιχεία 17 και 20 ελέγχουμε την περίπτωση των μοντέλων, όπου το ntree ισούται με 300 και 280 αντίστοιχα, ενώ ταυτόχρονα αλλάζουμε και θέτουμε mtry = 4 και nodesize = 3, καθώς είναι οι τιμές των παραμέτρων που μας έδωσαν τα δεύτερα καλύτερα αποτελέσματα στους κατά μέρους ελέγχους που προηγήθηκαν. Οι δύο αυτές δοκιμές βλέπουμε ότι εξάγουν παρόμοια αποτελέσματα, τα οποία όμως δεν εξασφαλίζουν υψηλότερη ικανότητα πρόβλεψης για το μοντέλο μας.

Στον πίνακα 15 που ακολουθεί συγκεντρώνονται τα αποτελέσματα όλων των δοκιμών που πραγματοποιήσαμε παραπάνω. Λόγω των μικρών διαφορών που σημειώνονται στην πλειοψηφία των αποτελεσμάτων, χρησιμοποιούμε κλίμακα διαβάθμισης χρωμάτων, για τα αποτελέσματα του AUC στο test set και του OOB error, προκειμένου να γίνουν πιο ευδιάκριτες οι διαφορές μεταξύ των διαφορετικών μοντέλων, αλλά και πιο εύκολη η ανίχνευση του καλύτερου μοντέλου της Random Forest. Συγκεκριμένα όσο πιο υψηλή η τιμή του AUC κάποιου μοντέλου σε σχέση με τα υπόλοιπα, τόσο πιο έντονη η απόχρωση του πράσινου. Ενώ το αντίστροφο ισχύει για τις τιμές του OOB rate, όσο μικρότερη η τιμή του τόσο εντονότερη η απόχρωση του πράσινου.

	cutoff	ntree	mtry	nodesize	set	AUC	OOB error
1	(0.5,0.5)	500	3	1	training	0,999998	0,97%
					test	0,9985084	1,92%

2	(0.76,0.24)	500	3	1	training	0,9999987	1,08%
					test	0,9985616	1,92%
3	(0.76,0.24)	300	3	1	training	0,9999982	1,04%
					test	0,9985253	2,07%
4	(0.76,0.24)	280	3	1	training	0,9999981	1,07%
					test	0,9985351	
5	(0.76,0.24)	250	3	1	training	0,999998	1,11%
					test	0,9983061	2,07%
6	(0.76,0.24)	200	3	1	training	0,999998	1,09%
					test	0,9985051	2,08%
7	(0.76,0.24)	180	3	1	training	0,9999977	1,05%
					test	0,9984704	2,00%
8	(0.76,0.24)	150	3	1	training	0,9999978	1,11%
					test	0,9985029	2,10%
9	(0.76,0.24)	300	4	1	training	1	1,07%
					test	0,9981301	
10	(0.76,0.24)	300	6	1	training	1	1,27%
					test	0,9982121	
11	(0.76,0.24)	280	4	1	training	1	1,09%
					test	0,998123	
12	(0.76,0.24)	280	6	1	training	1	1,26%
					test	0,9982322	
13	(0.76,0.24)	150	4	1	training	1	1,13%
					test	0,9980877	2,19%
14	(0.76,0.24)	150	6	1	training	1	1,23%
					test	0,9978829	2,08%
15	(0.76,0.24)	300	3	5	training	0,9999291	1,20%
					test	0,9978863	
16	(0.76,0.24)	300	3	3	training	0,9999796	1,13%
					test	0,9979713	
17	(0.76,0.24)	300	4	3	training	0,9999905	1,13%
					test	0,9980796	
18	(0.76,0.24)	280	3	5	training	0,9999296	1,19%

					test	0,997893	
19	(0.76,0.24)	280	3	3	training	0,9999793	1,15%
					test	0,9979598	
20	(0.76,0.24)	280	4	3	training	0,9999906	1,12%
					test	0,9980951	
21	(0.76,0.24)	150	3	5	training	0,9999267	1,29%
					test	0,9976928	2,11%
22	(0.76,0.24)	150	3	3	training	0,9999777	1,15%
					test	0,997692	2,27%

Πίνακας 12. Συνοπτικά αποτελέσματα παραμετροποίησης Random Forest

Έτσι βάσει όλων των δοκιμών που προηγήθηκαν και της σύγκρισης των αποτελεσμάτων τους, καταλήγουμε στο συμπέρασμα ότι το μοντέλο με την υψηλότερη προβλεπτική ικανότητα, το οποίο μπορεί να γενικευτεί για την πρόβλεψη άλλων άγνωστων δεδομένων, είναι αυτό της γραμμής 3, του πίνακα 15. Συγκεκριμένα είναι αυτό για το οποίο έχουμε πλήθος δέντρων ίσο με 300, πλήθος μεταβλητών που ελέγχεται σε κάθε split ίσο με 3 και ελάχιστο πλήθος παρατηρήσεων σε κάθε κόμβο απόφασης ίσο με 1. Το συγκριμένο μοντέλο επομένως, είναι το τελικό μοντέλο στο οποίο καταλήγουμε για τον αλγόριθμο Random Forest.

Προχωρούμε λοιπόν στον υπολογισμό των μέτρων αξιολόγησης του, για τα δεδομένα του training set και του test set, όπως πράξαμε και στους άλλους δύο αλγορίθμους που προηγήθηκαν. Τα αποτελέσματα αυτών καταγράφονται στον πίνακα 16, με ακρίβεια έξι δεκαδικών ψηφίων, προκειμένου να καταστούν εμφανείς οι διαφορές μεταξύ των διαφόρων μέτρων, καθώς τα μεγέθη που σημειώνονται σε αυτά είναι ιδιαίτερα υψηλά.

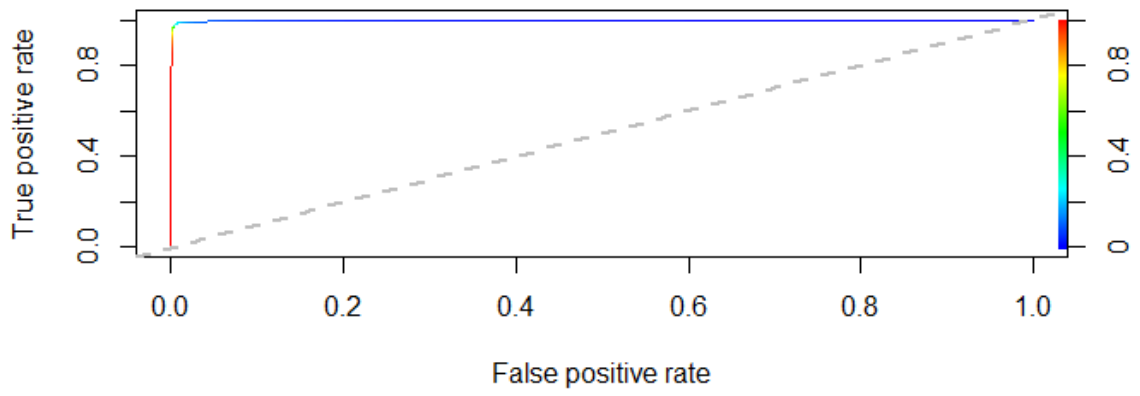
Evaluation Metric	Training set	Test set
Accuracy	0,998381	0,989778
Precision	0,993245	0,972799
Recall	1	0,984601
Specificity	0,997875	0,991396
F1 Score	0,996611	0,978664
AUC	0,999998	0,998525
Evaluation Metric		

Πίνακας 13. Μέτρα Αξιολόγησης τελικού μοντέλου Random Forest

Παρατηρώντας αρχικά τα αποτελέσματα του training set, μπορεί να δημιουργηθεί ανησυχία ότι το μοντέλο μας είναι overfitted, λόγω των υψηλών μεγεθών που σημειώνονται σε όλα τα μέτρα αυτού. Ωστόσο, ο έλεγχος των αποτελεσμάτων του test set μας καθησυχάζει, καθώς και στα

άγνωστα δεδομένα, τα μεγέθη των μέτρων αυτών είναι εξίσου υψηλά, καθιστώντας τις διαφορές μεταξύ των δύο σετ απειροελάχιστες.

Έτσι λοιπόν, βάσει της ακρίβειας (accuracy) του test set, βλέπουμε ότι το μοντέλο μας ταξινομεί σωστά το 98,9% των άγνωστων παρατηρήσεων, ενώ αξιολογεί λανθασμένα μόλις το 1,1% των παρατηρήσεων. Ταυτόχρονα η ακρίβεια του στην θετική πρόβλεψη αγγίζει το 97,3%, δηλαδή από το σύνολο αυτών που προβλέπει ότι θα αποχωρήσουν από την εταιρία, αξιολογεί σωστά το 97,3% των περιπτώσεων. Την ίδια στιγμή, εντοπίζει το 98,5% των εργαζομένων που πράγματι αποχωρούν από την εταιρία, και το 99,1% των εργαζομένων που παραμένουν σε αυτή, ποσοστά που είναι ιδιαίτερα υψηλά. Επίσης υψηλό είναι και το ποσοστό του F1 score που ισούται με 97,9%, γεγονός που δικαιολογείται από τα υψηλά ποσοστά του precision και recall που επισημάνθηκαν προηγουμένως. Τέλος το AUC ισούται με 99,8% γεγονός που επιβεβαιώνει την ύπαρξη πολύ υψηλού True Positive rate και χαμηλού False Positive rate, όπως φαίνεται ξεκάθαρα στην εικόνα 45, όπου η καμπύλη ROC «αγκαλιάζει» την πάνω αριστερά γωνία του διαγράμματος. Συμπερασματικά, μπορούμε να αναφέρουμε ότι το μοντέλο μας έχει πολύ υψηλή προβλεπτική ικανότητα στα άγνωστα δεδομένα, καθώς όλα τα μέτρα αξιολόγησης του απέχουν κατ' ελάχιστον από το 100% που αποτελεί το άριστο επίπεδο.



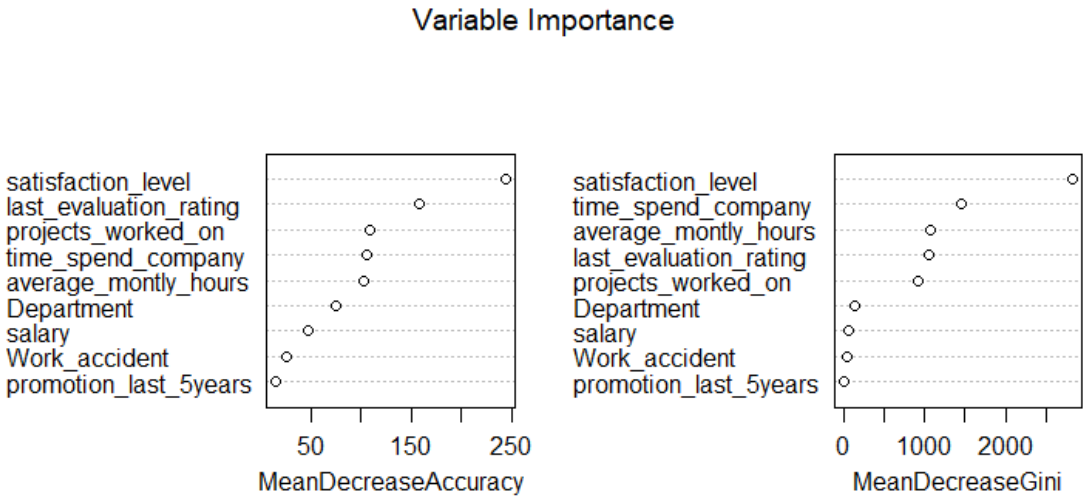
Εικόνα 45. Καμπύλη ROC μοντέλου Random Forest

Έχοντας καταλήξει στο τελικό μοντέλο του αλγορίθμου, εξετάζουμε την σημαντικότητα των μεταβλητών που συμπεριλαμβάνεται σε αυτό, βάσει των δεικτών mean decrease accuracy και mean decrease Gini, των οποίων οι απόλυτες τιμές παρατίθενται στην εικόνα 46 και η διαγραμματική τους απεικόνιση στην εικόνα 47.

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
satisfaction_level	70.278848	294.73287	244.01827	2834.009137
last_evaluation_rating	43.132003	155.73451	157.31906	1055.664903
projects_worked_on	39.931055	105.24192	108.67541	914.904334
average_montly_hours	54.819521	96.97438	102.41950	1067.902097
time_spend_company	73.630476	97.21656	105.16774	1445.316078
Work_accident	13.213025	24.31129	25.80396	44.173633
promotion_last_5years	6.060871	15.24788	15.22427	6.175264
Department	29.893971	91.51927	74.16667	130.203516
salary	26.579960	47.38403	46.65270	67.728616

Εικόνα 46. Σημαντικότητα μεταβλητών μοντέλου Random Forest

Συγκεκριμένα, ο δείκτης του mean decrease Gini αναφέρεται στην μείωση της ανομοιογένειας (impurity) που υπάρχει στο δέντρο, από την επιλογή της εκάστοτε μεταβλητής, ως κριτηρίου για τον διαχωρισμό των δεδομένων (split) σε κάθε κόμβο απόφασης. Από τα αποτελέσματα αυτών βλέπουμε λοιπόν, ποσό σημαντική είναι η κάθε μεταβλητή, σε σχέση με όλα τα splits που πραγματοποιήθηκαν. Από την άλλη το mean decrease accuracy αναφέρεται στην μείωση που θα επέλθει στην ακρίβεια του μοντέλου, αν αφαιρέσουμε κάποια από τις ανεξάρτητες μεταβλητές από αυτό (James & et al, 2013). Στην εικόνα 47 τα μεγέθη των δεικτών αυτών για κάθε μεταβλητή παρατίθενται σε φθίνουσα σειρά.



Εικόνα 47. Διαγραμματική απεικόνιση σημαντικότητας μεταβλητών Random Forest

Από τις εικόνες 46 και 47 παρατηρούμε λοιπόν, ότι το επίπεδο ικανοποίησης είναι η σημαντικότερη μεταβλητή σε όρους ακρίβειας και «κέρδους πληροφορίας», με την έννοια ότι με την χρήση αυτής εξασφαλίζεται μεγαλύτερη «καθαρότητα» στους τελικούς κόμβους και κατά συνέπεια καλύτερη πρόβλεψη για τις δύο κλάσεις. Ακολούθως, οι επόμενες πιο σημαντικές μεταβλητές για της πρόβλεψη της κλάσης κάθε εργαζομένου, αλλά σε αρκετά μικρότερο βαθμό, είναι τα έτη απασχόλησης του στην συγκεκριμένη εταιρία, οι μέσες μηνιαίες ώρες εργασίας του, ο βαθμός της τελευταίας αξιολόγησης του και το πλήθος των projects στα οποία απασχολείται. Από την άλλη, το τμήμα στο οποίο απασχολούνται, ο μισθός, η ύπαρξη

εργατικού ατυχήματος και η λήψη προαγωγής μέσα στα τελευταία πέντε χρόνια, δείχνουν να μην είναι ιδιαίτερα σημαντικές για το μοντέλο, καθώς φαίνεται πως δεν επηρεάζουν ιδιαίτερα την πρόβλεψη των αποτελεσμάτων που προκύπτουν.

3.5 ΑΞΙΟΛΟΓΗΣΗ ΜΟΝΤΕΛΩΝ

Η διαδικασία ανάλυσης που προηγήθηκε στις υποενότητες του παρόντος κεφαλαίου, έγινε με γνώμονα την κατασκευή του καλύτερου δυνατού μοντέλο για καθέναν από τους αλγορίθμους, που μελετούμε στην παρούσα εργασία. Απώτερος σκοπός αυτής της διαδικασίας είναι στο τέλος της παρούσας υποενότητας, να επιλέξουμε ένα μοντέλο που θα μπορεί να γενικευτεί με μεγαλύτερη ασφάλεια, στην πρόβλεψη άλλων άγνωστων δεδομένων του πληθυσμού. Έτσι λοιπόν, προχωρούμε στην σύγκριση των τελικών μοντέλων που προέκυψαν από τους αλγορίθμους Logistic Regression, Decision Trees και Random Forest. Η σύγκριση θα βασιστεί στα αποτελέσματα των μέτρων αξιολόγησης που προέκυψαν από την πρόβλεψη των άγνωστων δεδομένων του test set. Τα αποτελέσματα των μέτρων αξιολόγησης όλων των αλγορίθμων είναι συγκεντρωμένα στον πίνακα 17.

Evaluation Metric	Logistic Regression	Decision Trees	Random Forest
Accuracy	0,744	0,972	0,990
Precision	0,476	0,968	0,973
Recall	0,768	0,913	0,985
Specificity	0,736	0,991	0,991
F1 Score	0,588	0,940	0,979
AUC	0,826	0,973	0,999

Πίνακας 14. Συγκεντρωτικά αποτελέσματα Αλγορίθμων Ταξινόμησης

Βάσει αυτών παρατηρούμε ότι ο τρίτος αλγόριθμος του Random Forest προβλέπει σωστά την κλάση για το 99,0% των περιπτώσεων, ενώ την ίδια στιγμή αυτός του Decision Trees παρουσιάζει ελαφρώς χειρότερα αποτελέσματα, αλλά εξίσου υψηλά της τάξης του 97,2%. Αντίθετα, ο πρώτος αλγόριθμος σημειώνει σημαντικά μικρότερη ακρίβεια σε σχέση με τους άλλους δύο, καθώς εντοπίζει ορθά μόλις το 74,4% των κλάσεων. Προχωρώντας τώρα στην πιο αναλυτική μελέτη των αποτελεσμάτων που αφορούν την κλάση 1, βλέπουμε ότι στο πρώτο μοντέλο μόλις το 47,6% των στοιχείων που ταξινομήθηκαν στην κλάση 1, ανήκουν πράγματι σε αυτή, ενώ την ίδια στιγμή στο δεύτερο και τρίτο μοντέλο τα ποσοστά που σημειώνονται είναι ιδιαίτερα υψηλά, καθώς αντιστοιχούν στο 96,8% και στο 97,3% αντίστοιχα. Για να εξετάσουμε την ικανότητα των μοντέλων στον εντοπισμό των υπαλλήλων που πράγματι αποχωρούν από την εταιρία ή που μένουν σε αυτή μελετούμε τα μέτρα του recall και του specificity αντίστοιχα. Και στα δύο μέτρα ο αλγόριθμος Random Forest είναι αυτός που σημειώνει τις υψηλότερες επιδόσεις, καθώς εντοπίζει το 98,5% των εργαζομένων που πράγματι

παραιτούνται και το 99,1% αυτών που παραμένουν. Ίδια αποτελέσματα σημειώνει στο specificity και ο αλγόριθμος Decision Trees, ενώ η επίδοση του στο recall είναι μειωμένη κατά 7,2%. Εμφανώς χειρότερα είναι τα αποτελέσματα που καταγράφονται για τα μέτρα αυτά, στον αλγόριθμο Logistic Regression, καθώς αντιστοιχούν μόλις στο 76,8% και 73,6% αντίστοιχα. Το μέτρο που λαμβάνει ταυτόχρονα υπόψη του, τόσο το precision όσο και το recall που εξετάσαμε παραπάνω, καθώς και την άνιση εκπροσώπηση των κλάσεων, είναι το F1 score. Για τον λόγο αυτό χρησιμοποιείται συχνά ως μέτρο σύγκρισης μεταξύ διαφορετικών αλγορίθμων. Από αυτό διαπιστώνουμε και πάλι ότι ο αλγόριθμος Random Forest υπερτερεί έναντι των Decision Trees και Logistic Regression, σημειώνοντας ποσοστό 97,9% έναντι του 94,0% και 58,8% που καταγράφεται στους άλλους δύο. Το τελευταίο μέτρο αξιολόγησης που εξετάζουμε, άλλα ένα από τα σημαντικότερα για το πρόβλημα που μελετούμε, είναι το AUC, καθώς λαμβάνει ταυτόχρονα υπόψη, το ποσοστό των True Positive και False Positive. Στην περίπτωση του αλγορίθμου Random Forest, το AUC του μοντέλου που διαμορφώθηκε ισούται με 99,9%, δηλαδή διαφέρει κατά 0,01% από το άριστο. Από την άλλη πολύ υψηλό ποσοστό σημειώνει και το μοντέλο του Decision Trees καθώς ισούται με 97,3%, ενώ μετριότερα είναι τα αποτελέσματα για το μοντέλο του Logistic Regression καθώς ισούται με 82,6%.

Συμπερασματικά λοιπόν, μπορούμε να αναφέρουμε βάσει των παραπάνω συγκρίσεων, ότι το μοντέλου που προέκυψε από τον αλγόριθμο Random Forest είναι εμφανώς καλύτερο από τα μοντέλα των άλλων δύο αλγορίθμων. Άρα θα μπορούσε να χρησιμοποιηθεί μελλοντικά για την ταξινόμηση άλλων άγνωστων δεδομένων που απαρτίζονται από αντίστοιχα χαρακτηριστικά.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Κύριο αντικείμενο της παρούσας εργασίας αποτέλεσε η εξεύρεση των παραγόντων που συντελούν στην εργασιακή φορά ενός εργαζομένου. Στα δεδομένα που μελετήσαμε η εργασιακή φθορά ισοδυναμεί με παραίτηση, από την εταιρία στην οποία βρίσκεται κάθε εργαζόμενος. Έτσι προχωρήσαμε στην διερεύνηση αλγορίθμων στατιστικής μάθησης, που έχουν ως στόχο την επίλυση προβλημάτων ταξινόμησης δυο κλάσεων. Συγκεκριμένα χρησιμοποιήθηκαν οι αλγόριθμοι των Logistic Regression, Decision Trees και Random Forest και αναλύθηκε ο τρόπος με τον οποίο λειτουργεί καθένας από αυτούς, προκειμένου να ταξινομήσει τα δεδομένα σε κάθε κλάση. Τέλος βάσει των αποτελεσμάτων που προέκυψαν από την ταξινόμηση των άγνωστων δεδομένων, υπολογίσαμε την απόδοση κάθε αλγορίθμου βάσει των μέτρων αξιολόγησης τους.

Το συμπέρασμα στο οποίο καταλήξαμε είναι ότι το μοντέλο που προέκυψε από την εκπαίδευση του αλγορίθμου Random Forest είναι αυτό που μας δίνει με διαφορά τα καλύτερα αποτελέσματα ταξινόμησης, στο dataset που μελετήσαμε. Συγκεκριμένα είδαμε ότι όχι μόνο σημειώνει τις υψηλότερες επιδόσεις έναντι των άλλων δύο μοντέλων, αλλά ταυτόχρονα έχει και άριστη δυνατότητα πρόβλεψης των δυο κατηγοριών, δεδομένου ότι όλα τα μέτρα του προσεγγίζουν με πολύ μεγάλη ακρίβεια το 100%. Εξίσου υψηλές επιδόσεις, αλλά λίγο χαμηλότερες, παρατηρήσαμε ότι σημειώνει το μοντέλο του αλγορίθμου των Decision Trees. Όπως είδαμε το συγκεκριμένο μοντέλο υστερεί στο να προβλέψει με την ίδια ακρίβεια, σε σχέση με το μοντέλο του Random Forest, την κατηγορία των εργαζομένων που φεύγουν από την εταιρία. Τέλος διαπιστώθηκε ότι το μοντέλο που προέκυψε από τον αλγόριθμο Logistic Regression είναι αυτό που μας δίνει τα λιγότερο αξιόπιστα αποτελέσματα, έναντι των τριών μοντέλων, καθώς σε όλα τα μέτρα αξιολόγησης σημείωσε μέτριες επιδόσεις.

Έχοντας καταλήξει ότι το μοντέλο του Random Forest είναι το πιο αξιόπιστο, από τα υπόλοιπα δύο, για την πρόβλεψη της σωστής κλάσης εστιάζουμε αποκλειστικά σε αυτό, προκειμένου να ανακαλύψουμε τους παράγοντες που συμβάλλουν στην εργασιακή φθορά. Βάσει της ανάλυσης που προηγήθηκε, μπορούμε να συμπεράνουμε ότι ο καθοριστικότερος παράγοντας για την παραίτηση ενός εργαζομένου είναι το επίπεδο ικανοποίησης που του προσφέρει η εργασία του. Οι αμέσως επόμενοι πιο σημαντικοί παράγοντας είναι αυτοί του συνόλου των ετών απασχόλησης στην εταιρία, καθώς και των μέσων μηνιαίων ωρών εργασίας του. Ακολούθως η επίδοση του βάσει της τελευταίας αξιολόγησης και ο αριθμός των projects που αναλαμβάνει,

δείχνουν να ασκούν κάποια επίδραση στην απόφαση του να αποχωρήσει από την εταιρία, αλλά σε πολύ μικρότερο βαθμό από ότι οι άλλοι παράγοντες.

4.2 ΚΑΤΕΥΘΥΝΣΕΙΣ ΜΕΛΛΟΝΤΙΚΗΣ ΜΕΛΕΤΗΣ

Από τα αποτελέσματα της εργασίας αυτής προέκυψαν κάποια θέματα που αξίζει να μελετηθούν στο μέλλον. Καταρχάς θα μπορούσαν να δοκιμαστούν τα αποτελέσματα της παραπάνω μελέτης είτε σε δεδομένα διαφορετικών χρονικών παραθύρων είτε κάνοντας χρήση άλλων αλγορίθμων. Ακόμη είναι χρήσιμη η μελλοντική συγκέντρωση επιπλέον χαρακτηριστικών, για την κατανόηση της εργασιακής φθοράς, λόγω του ότι οι διαθέσιμοι παράγοντες ήταν περιορισμένοι. Κάτι τέτοιο θα βοηθήσει στην κατασκευή ενός νέου πιο εξειδικευμένου προβλεπτικού μοντέλου, ενώ ταυτόχρονα θα συμβάλλει στην κατανόηση του συγκεκριμένου προβλήματος σε μεγαλύτερο βαθμό. Τέτοιοι παράγοντες θα μπορούσαν να είναι τα επαγγελματικά ταξίδια, η οικογενειακή κατάσταση, το επίπεδο εκπαίδευσης, τα bonus κ.α. Τέλος, επιπλέον μελέτη μπορεί να διεξαχθεί κάνοντας χρήση διαφορετικών δειγμάτων, από διαφορετικούς κλάδους εταιριών, προκειμένου να επικυρωθεί η ορθότητα των αποτελεσμάτων της παρούσας εργασίας.

Το συγκεκριμένο πεδίο έρευνας είναι χρήσιμο για τα στελέχη των τμημάτων Ανθρώπινου Δυναμικού και για τους managers καθώς για κάθε εταιρία οι ανθρώπινοι πόροι αποτελούν πηγή ανταγωνιστικού πλεονεκτήματος, στην σημερινή παγκόσμια οικονομία. Ο σχεδιασμός των ανθρωπίνων πόρων αποτελεί μια από τις σημαντικότερες ευθύνες του τμήματος HR. Επομένως τα εργαλεία και τα μοντέλα που βελτιώνουν την πρόβλεψη των παραγόντων που επηρεάζουν την εργασιακή φθορά, μπορούν να αποφέρουν μεγάλη αξία στα στελέχη του HR, τα οποία πλέον καλούνται να παίξουν τον ρόλο του στρατηγικού εταίρου σε κάθε οργανισμό (Ulrich & Brockbank, 2005).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Attrition. (2018). Ανάκτηση March 2019, από mba skool.com: https://www.mbaskool.com/business-concepts/human-resources-hr-terms/1772-attrition.html

Benjamini, Y. (1988). *Opening the Box of a Boxplot.* American Statistician.

Bonett, D. G., & Wright, T. A. (2000). *Sample size requirements for estimating pearson, kendall and spearman correlations.* The Psychometric Society .

Breiman, L. (2001). *Machine Learning.* Springer US.

Breiman, L., & et al. (1984). *Classification and regression trees.* Chapman and Hall/CRC.

Carey, D., & Ogden, D. (2004). *The human side of M & A: The human side of M & A: How CEOs Leverage the Most Important Asset in Deal Making.* Oxford University Press.

Castle, N. (2018, March). *Regression vs. Classification Algorithms*. Ανάκτηση March 2019, από Oracle + DataScience.com: https://www.datascience.com/blog/regression-and-classification-machine-learning-algorithms

Chapmann, J. (2017). *Machine Learning: Fundamental Algorithms for Supervised and Unsupervised Learning With Real-world Applications.* CreateSpace Independent Publishing Platform.

Fawcett, T. (2006). An introduction to ROC analysis. *Journal Pattern Recognition Letters*.

Frost, J. (2017). *Statistics By Jim-Making statistics intuitive*. Ανάκτηση March 2019, από The Difference between Linear and Nonlinear Regression Models: https://statisticsbyjim.com/regression/difference-between-linear-nonlinear-regression-models/

Gujarati, D., & Porter, D. C. (2009). *Basic Econometrics (5th ed.).* McGraw-Hill/Irwin.

Hastie, T., & et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.).* Springer.

Hossin, M., & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2,*.

Huang, J., & Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *Journal IEEE Transactions on Knowledge and Data Engineering*.

James, G., & et al. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *14th International Joint Conference on Artificial Inrelligence (IJCAI).* Morgan Kaufmann Publishers Inc.

Liaw, A., & Wiener, M. (2002). *Classification and regression by randomForest.* R news, 2(3), 18-22.

Merhar, C. (2016). *Employee Retention - The Real Cost of Losing an Employee*. Retrieved March 2019, from PeopleKeep: https://www.peoplekeep.com/blog/bid/312123/employee-retention-the-real-cost-of-losing-an-employee

Mishra, A. (2018). *Metrics to Evaluate your Machine Learning Algorithm.* Ανάκτηση March 2019, από Towards Data Science: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Mobley, W. H. (1982). *Employee turnover: causes, consequences, and control.* Addison-Wesley Publishing.

Mohr, D. C., & et al. (2012). *Employee turnover and operational performance: the moderating effect of group-oriented organisational culture.* Human Resource Management Journal 22(2), 216-233.

Nisbet, R., & et al. (2017). *Handbook of Statistical Analysis and Data Mining Applications (2nd ed.).* Academic Press.

Quinlan, J. R. (1986). Induction of Decision Trees. *Journal Machine Learning Volume 1 Issue 1* .

Ranjan, J., & et al. (2008). Data mining techniques for better decisions in human resource management systems. International Journal of Business Information Systems (IJBIS), Volume 3, Issue 5.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach (3rd Edition).* Prentice Hall.

Sheskin, D. J. (2011). *Handbook of Parametric and Nonparametric Statistical Procedures (5th ed.).* Chapman and Hall/CRC.

Shetty, B. (2018). *Supervised Machine Learning: Classification.* Ανάκτηση March 2019, από Towards Data Science: https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d

Suits, D. B. (1957). Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association. 52 (280)*.

Ulrich, D., & Brockbank, W. (2005). *The HR Value Proposition.* Harvard Business Press.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory (2nd Ed.).* Springer.

Witten, I. H. (2011). *Data Mining: Practical machine learning tools and techniques (3rd ed.).* Morgan Kaufmann.

Πανάρετος, Ι. (2003). *Γραμμικά Μοντέλα με έμφαση στις εφαρμογές.* Ανάκτηση Μάρτιος 2019, από Οικονομικό Πανεπιστήμιο Αθηνών: http://www2.stat-athens.aueb.gr/~jpan/Linear-Models-Supplement.pdf

ΠΑΡΑΡΤΗΜΑ

#Descriptive Statistics

```
# Importing Dataset

train=read.csv('train.csv')

test=read.csv('test.csv')

#removing the ID variable from test set

test=test[,-1]


#The data set

employees=rbind(train,test)


# Dimension of dataset

dim(employees)


# stracture of dataset

str(employees)


#Missing Values

sum(is.na.data.frame(employees))


#converting integer into factor

employees$Attrition=as.factor(employees$Attrition)

employees$Work_accident=as.factor(employees$Work_accident)
```

```
employees$promotion_last_5years=as.factor(employees$promotion_last_5years)


#Summary statistics

summary(employees)


#Standard diviation

sapply(employees[,1:5],sd)


#Frequency Distribution of continuous variables

attach(employees)


satlev<-table((satisfaction_level),

        cut((satisfaction_level),

           breaks=seq(0,10,1)))


margin.table(satlev,2)


hist(satisfaction_level, main="Histogram of Satisfaction Level ",

    xlab="satisfaction_level",

    col="blue",

    freq=T,

    las=1,

    breaks=seq(0,10,1),

    ylim=c(0,5000))
```

```
lastev<-table((last_evaluation_rating),
              cut((last_evaluation_rating),
                  breaks=seq(0,10,1)))

margin.table(lastev,2)

hist(last_evaluation_rating, main="Histogram of Evaluation Rating",
     xlab="last_evaluation_rating",
     col="blue",
     freq=T,
     las=1,
     breaks=seq(0,10,1),
     ylim=c(0,7000))

proj<-table((projects_worked_on),
            cut((projects_worked_on),
                breaks=seq(0,10,1)))

margin.table(proj,2)

hist(projects_worked_on , main="Histogram of Number of Projects",
     xlab="projects_worked_on",
     col="blue",
```

```
    freq=T,
    las=1,
    breaks=seq(0,10,1),
    ylim=c(0,10000))

avghrs<-table((average_montly_hours),
            cut((average_montly_hours),
              breaks= seq(80,340,20)))

margin.table(avghrs,2)

hist(average_montly_hours , main="Histogram of Avg Monthly Hours",
    xlab="average_montly_hours",
    col="blue",
    xlim=c(0, 350),
    ylim=c(0, 7600),
    freq=T,
    las=1,
    breaks= seq(80,340,20))


yrspd<-table(time_spend_company,
           cut(time_spend_company,
             breaks=seq(0,10,1)))
```

```
margin.table(yrspd,2)


hist(time_spend_company, main="Histogram of Years in Company",
     xlab="time_spend_company",
     col="blue",
     freq=T,
     las=1,
     breaks=seq(0,10,1),
     ylim=c(0,15000))


# Frequency Distribution of categorical variables


df<-table(Work_accident)
bp<-barplot(table(Work_accident), main="Plot of Work Accidents",
    xlab="work_accident",
    col="blue",
    ylim=c(0,35000))
text(bp,df,labels=df,pos=3,offset = 0.5)


df<-table(promotion_last_5years)
bp<-barplot(table(promotion_last_5years), main="Plot of Promotion for last5years",
        xlab="promotion_last_5years",
        col="blue",
```

```
        ylim=c(0,35000))
text(bp,df,labels=df,pos=3,offset = 0.5)


df<-table(Department)
par(mar=c(8,4,4,4))
dp<-barplot(table(Department), main="Plot of Departments",
        col="blue",
        ylim=c(0,10000),
        besides=TRUE,
        axes=TRUE,axisnames=TRUE,axis.lty=0,
        las=2)
text(x=dp,df,labels=df,pos=3,offset = 0.5,tick=FALSE)



df<-table(salary)
bp<-barplot(table(salary), main="Plot of Salary",
        xlab="salary",
        col="blue",
        ylim=c(0,20000))
text(bp,df,labels=df,pos=3,offset = 0.5)


df<-table(Attrition)
bp<-barplot(table(Attrition), main="Plot for Attrition",
        xlab="attrition",
```

```
        col="blue",
        ylim=c(0,25000))
text(bp,df,labels=df,pos=3,offset = 0.5)

#Box Plots

boxplot(satisfaction_level~Attrition,
        main= "Attrition VS Satisfaction",
        xlab="attrition",
        ylab="satisfaction",
        col= "gray")

boxplot(last_evaluation_rating~Attrition,
        main= "Attrition VS Evaluation Rate",
        xlab="attrition",
        ylab="evaluation",
        col= "gray")

boxplot(projects_worked_on~Attrition,
        main= "Attrition VS Number of Projects",
        xlab="attrition",
        ylab="projects",
        col= "gray")
```

```
boxplot(average_montly_hours~Attrition,
        main= "Attrition VS Avg Montly Hours",
        xlab="attrition",
        ylab="monthly hours",
        col= "gray")


boxplot(time_spend_company~Attrition,
        main= "Attrition VS Years in Company",
        xlab="attrition",
        ylab="years",
        col= "gray")


# Correlation table of continious variables
#install.packages("corrplot")
library(corrplot)
correlation1<-employees[,c(1:5)]
correlation1<-cor(correlation1)


#hide upper triangle
lowercor1<-round(correlation1,3)
lowercor1[upper.tri(correlation1)]<-""
lowercor1<-as.data.frame(lowercor1)
lowercor1
```

```
#Correlation plot Pearson
corrplot(correlation1,method="circle",type='lower',
        tl.srt = 360,tl.col = "black")

correlation2<-employees[,c(1:5)]
correlation2<-cor(correlation2,method='spearman')

#hide upper triangle
lowercor2<-round(correlation2,3)
lowercor2[upper.tri(correlation2)]<-""
lowercor2<-as.data.frame(lowercor2)
lowercor2

#Correlation plot Spearman
corrplot(correlation2,method="circle",type='lower',
        tl.srt = 360,tl.col = "black")
```

#Logistic Regression

```
# Importing Dataset
train=read.csv('train.csv')
test=read.csv('test.csv')
#removing the ID variable from test set
test=test[,-1]
```

```
#The data set

employees=rbind(train,test)


#Converting integers into factors

employees$Attrition=as.factor(employees$Attrition)

employees$Work_accident=as.factor(employees$Work_accident)

employees$promotion_last_5years=as.factor(employees$promotion_last_5years)


#Create dummies for department and salary

Dep_ = factor(employees$Department)

dep_dummies = model.matrix(~Dep_ +0)


salary_ =factor(employees$salary)

sal_dummies= model.matrix(~salary_ +0)


#Constructing the new dataset of employees

employees=data.frame(employees,dep_dummies,sal_dummies)


#Removing columns salary,department,dep_accounting,sal_high

employees<-data.frame(employees[ ,-c(8,9,11,21)])


#Splitting the Dataset into the Training set and Test set

#install.packages('caTools')
```

```
library(caTools)
set.seed(12)
split= sample.split(employees$Attrition,SplitRatio =0.7)
training_set= subset(employees,split==TRUE)
test_set= subset(employees,split==FALSE)

#Fitting logistic regression
classifier=glm(formula=Attrition ~ .,
               family=binomial,
               data=training_set)

summary(classifier)

#Removing insignificant variables
classifier2=glm(formula=Attrition ~ .
              -Dep_support -Dep_technical - Dep_marketing - Dep_product_mng- Dep_sales,
               family=binomial,
               data=training_set)

summary(classifier2)

#Calculating events rates
table(training_set$Attrition)
table(test_set$Attrition)
```

```
######### Training Set##########

#Predicting the training set results

prob_pred= predict(classifier2,type='response',newdata =training_set[ ,-8])

y_hat= ifelse(prob_pred>0.23,1,0)


cm=as.matrix(table(training_set[,8],y_hat))

cm


accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision


recall= cm[2,2]/sum(cm[2,1:2])

recall


specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)

f1


#Plotting the ROC curve
```

```
#install.packages("ROCR")

library('ROCR')

roc.pred<-prediction(prob_pred,training_set$Attrition)

roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE,print.cutoffs.at=seq(0,1,by=0.1),

    text.adj=c(-0.2,1.7))

abline(a=0,b=1,lwd=2,lty=2,col="gray")


auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc


############## Test set ##############

#Predicting the test set results

prob_pred= predict(classifier2,type='response',newdata =test_set[ ,-8])

y_hat= ifelse(prob_pred>0.23,1,0)


cm=as.matrix(table(test_set[,8],y_hat))


accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision
```

```
recall= cm[2,2]/sum(cm[2,1:2])
recall

specificity= cm[1,1]/sum(cm[1,1:2])
specificity

f1 = 2 * precision * recall / (precision + recall)
f1

#Plotting the ROC curve
#install.packages("ROCR")
library('ROCR')
roc.pred<-prediction(prob_pred,test_set$Attrition)
roc.perf<-performance(roc.pred,'tpr','fpr')
plot(roc.perf,colorize = TRUE,print.cutoffs.at=seq(0,1,by=0.1),
    text.adj=c(-0.2,1.7))
abline(a=0,b=1,lwd=2,lty=2,col="gray")

auc <- performance(roc.pred, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

#Desicion Tree Classification

```
# Importing Dataset

train=read.csv('train.csv')

test=read.csv('test.csv')

#removing the ID variable from test set

test=test[,-1]


#The data set

employees=rbind(train,test)


#Converting integers into factors

employees$Attrition=as.factor(employees$Attrition)

employees$Work_accident=as.factor(employees$Work_accident)

employees$promotion_last_5years=as.factor(employees$promotion_last_5years)



#Splitting the Dataset into the Training set and Test set

#install.packages('caTools')

library(caTools)

set.seed(12)

split= sample.split(employees$Attrition,SplitRatio =0.7)

training_set= subset(employees,split==TRUE)

test_set= subset(employees,split==FALSE)
```

```
#Fitting Desicion Tree Classification

#install.packages('tree')

library(tree)

classifier=tree(formula=Attrition ~.,

                data=training_set)



summary(classifier)

plot(classifier)

text(classifier,pretty=0)

print(classifier)


######### Training Set##########

#Predicting the training set results

y_pred=predict(classifier,newdata = training_set[,-10],type='class')


cm=table(training_set[,10],y_pred)

print(cm)


#Evaluation metrics for training set

accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])
```

```
precision


recall= cm[2,2]/sum(cm[2,1:2])

recall


specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)

f1


#Plotting the ROC curve

library('ROCR')

prob_pred=predict(classifier,type='vector',newdata = training_set[,-10])[,2]

roc.pred<-prediction(prob_pred,training_set$Attrition)

roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE,print.cutoffs.at=seq(0,1,by=0.1),

    text.adj=c(-0.2,1.7))

abline(a=0,b=1,lwd=2,lty=2,col="gray")


auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc
```

```
############# Test set #############

#Predicting the test set results

y_pred=predict(classifier,newdata = test_set[,-10],type='class')

cm=table(test_set[,10],y_pred)

print(cm)


#Evaluation metrics for test set

accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision


recall= cm[2,2]/sum(cm[2,1:2])

recall


specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)

f1


#Plotting the ROC curve

prob_pred= predict(classifier,type='vector',newdata =test_set[,-10])[,2]
```

```
roc.pred<-prediction(prob_pred,test_set$Attrition)

roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE)

abline(a=0,b=1,lwd=2,lty=2,col="gray")


auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc


# Pruning the tree

set.seed(12)

cv.classifier=cv.tree(classifier, FUN=prune.misclass)

print(cv.classifier)


par(mfrow =c(1,2))

plot(cv.classifier$size,cv.classifier$dev,type="b")

plot(cv.classifier$size,cv.classifier$k,type="b")


set.seed(12)

prune.classifier=prune.misclass(classifier,best=10)

plot(prune.classifier)

text(prune.classifier,pretty=0)


#Evaluation metrics for training set after pruning
```

```
set.seed(12)

y_pred=predict(prune.classifier,newdata = training_set[,-10],type='class')

cm=table(training_set[,10],y_pred)

print(cm)


accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision


recall= cm[2,2]/sum(cm[2,1:2])

recall


specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)

f1

#Plotting the ROC curve

prob_pred= predict(prune.classifier,type='vector',newdata =training_set[,-10])[,2]

roc.pred<-prediction(prob_pred,training_set$Attrition)

roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE)
```

```
abline(a=0,b=1,lwd=2,lty=2,col="gray")


auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc


#Evaluation metrics for test set after pruning

y_pred=predict(prune.classifier,newdata = test_set[,-10],type='class')

cm=table(test_set[,10],y_pred)

print(cm)


accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision


recall= cm[2,2]/sum(cm[2,1:2])

recall


specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)
```

```
f1

#Plotting the ROC curve

prob_pred= predict(prune.classifier,type='vector',newdata =test_set[,-10])[,2]

roc.pred<-prediction(prob_pred,test_set$Attrition)

roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE,print.cutoffs.at=seq(0,1,by=0.1),

    text.adj=c(-0.2,1.7))

abline(a=0,b=1,lwd=2,lty=2,col="gray")


auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc
```

#Random Forest Classification

```
# Importing Dataset

train=read.csv('train.csv')

test=read.csv('test.csv')

#removing the ID variable from test set

test=test[,-1]


#The data set

employees=rbind(train,test)
```

```
#Converting integers into factors
employees$Attrition=as.factor(employees$Attrition)
employees$Work_accident=as.factor(employees$Work_accident)
employees$promotion_last_5years=as.factor(employees$promotion_last_5years)

#Splitting the Dataset into the Training set and Test set
library(caTools)
set.seed(12)
split= sample.split(employees$Attrition,SplitRatio =0.7)
training_set= subset(employees,split==TRUE)
test_set= subset(employees,split==FALSE)

#Fitting Random Forest Classification
#install.packages("randomForest")
library(randomForest)
set.seed(12)
classifier=randomForest(formula=Attrition ~.,
                        data = training_set,
                        type=classification,
                        importance= TRUE)

classifier
plot(classifier)
importance(classifier)
```

```
varImpPlot(classifier,
           sort = T,
           main="Variable Importance")



classifier
plot(classifier)
importance(classifier)
varImpPlot(classifier,
           sort = T,
           main="Variable Importance")

#Defining the threshold to 0.24
classifier=randomForest(formula=as.factor(Attrition) ~.,
                        data = training_set,
                        type=classification,
                        importance= TRUE,
                        cutoff=c(0.76,0.24))
classifier

#Selecting the best number of each parameter
set.seed(12)
classifier=randomForest(formula=as.factor(Attrition) ~.,
                        data = training_set,
```

```
                type=classification,
                importance= TRUE,
                cutoff=c(0.76,0.24),
                ntree=280,
                mtry=3,
                nodesize=3)
classifier

#Tuning the best number of predictors at each split
set.seed(12)
mtry=tuneRF(training_set[,-10],
        training_set$Attrition,
        mtryStart = 3,
        stepFactor = 2,
        improve = 0.001,
        ntreeTry = 300,
        cutoff=c(0.76,0.24),
        plot=TRUE,
        trace = TRUE)

######### Training Set##########

#Predicting the results for the training set
y_hat=predict(classifier,newdata = training_set[,-10])
```

```
cm=table(training_set[,10],y_hat)

print(cm)



#Evaluation metrics for training set

accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision


recall= cm[2,2]/sum(cm[2,1:2])

recall


specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)

f1


#Plotting the ROC curve

library('ROCR')

prob_pred= predict(classifier,type='prob',newdata =training_set[,-10])[,2]

roc.pred<-prediction(prob_pred,training_set$Attrition)
```

```
auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc


roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE)

abline(a=0,b=1,lwd=2,lty=2,col="gray")


############# Test set #############


#Predicting the results for the test set

y_pred=predict(classifier,newdata = test_set[,-10])

cm=table(test_set[,10],y_pred)

print(cm)


#Evaluation metrics for test set

accuracy = (sum(diag(cm)))/sum(cm)

accuracy


precision= cm[2,2]/sum(cm[1:2,2])

precision


recall= cm[2,2]/sum(cm[2,1:2])

recall
```

```
specificity= cm[1,1]/sum(cm[1,1:2])

specificity


f1 = 2 * precision * recall / (precision + recall)

f1


#Plotting the ROC curve

prob_pred= predict(classifier,type='prob',newdata =test_set[,-10])[,2]

roc.pred<-prediction(prob_pred,test_set$Attrition)

auc <- performance(roc.pred, measure = "auc")

auc <- auc@y.values[[1]]

auc


roc.perf<-performance(roc.pred,'tpr','fpr')

plot(roc.perf,colorize = TRUE)

abline(a=0,b=1,lwd=2,lty=2,col="gray")
```