



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

Πανεπιστήμιο Πειραιώς

Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφορικής»

**ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ ΤΕΧΝΗΤΗ
ΝΟΗΜΟΣΥΝΗ**

Όνοματεπώνυμο : Μαρία Καούνη

Αριθμός Μητρώου : ΜΠΠΛ2219

Όνοματεπώνυμο : Μαρία Αμοργιανού

Αριθμός Μητρώου : ΜΠΠΛ2205

Abstract

Ο τομέας των τηλεπικοινωνιών είναι ένα ταχέως αναπτυσσόμενος τομέας όπου ένα μεγάλο, κρίσιμο και δύσκολο πρόβλημα αποτελεί η παρακίνηση του πελάτη. Μεγάλο ενδιαφέρον αποτελεί για τους ερευνητές η πρόβλεψη της συμπεριφοράς των πελατών προκειμένου να μπορούν να είναι σε θέση να ξέρουν τον αριθμό των πελατών που χάνονται και αυτών που μένουν στην εταιρεία. Το μεγαλύτερο κίνητρο είναι η έντονη ανάγκη των επιχειρήσεων να διατηρούν τους πελάτες που έχουν μαζί με υψηλό κόστος και η προσπάθεια τους να αποκτήσουν κι' άλλους επιπλέον πελάτες. Τεράστια πρόκληση αποτελεί πλέον η διατήρηση της υπάρχουσας πελατείας, για να ανταποκριθούν στην ανάγκη να επιβιώσουν μέσα στο ανταγωνιστικό περιβάλλον της εποχής που επικρατεί. Έχει παρατηρηθεί έλλειψη αποτελεσματικών προσεγγίσεων του Customer Churn Prediction (CCP) στο τομέα των τηλεπικοινωνιών. Σε έρευνα που έχει πραγματοποιηθεί στον τομέα των τηλεπικοινωνιών έχει παρουσιαστεί ότι το κόστος απόκτησης νέου πελάτη είναι πολύ περισσότερο από το να διατηρηθεί η υπάρχουσα πελατεία. Έτσι με την συλλογή πληροφοριών από τις βιομηχανίες τηλεπικοινωνιών μπορεί να βρεθεί μια πρόβλεψη του πελάτη για το εάν θα εγκαταλείψει την εταιρεία ή όχι.

Περιεχόμενα

Abstract	2
Περιεχόμενα.....	3
Εισαγωγή	4
Περιγραφή.....	5
Θεωρία	7
Σχεδιαστικές Αποφάσεις	9
Logistic Regression	9
Decision Tree	10
K Nearest Neighbor Classifier	11
Random Forest Classifier	12
Ανάπτυξη	14
Εκτέλεση.....	26
Συμπεράσματα	68
Βιβλιογραφία	69

Εισαγωγή

Στη συγκεκριμένη εργασία, εξετάζεται το πρόβλημα της χαμένης πελατείας (churn) σε μια εταιρεία τηλεπικοινωνιών. Η αύξηση του ανταγωνισμού στην αγορά έχει οδηγήσει τις εταιρείες να επικεντρώνονται στη διατήρηση των πελατών. Η έρευνα αναζητά λύσεις πρόβλεψης του ποσοστού χαμένης πελατείας, διερευνώντας διάφορες μεθόδους, όπως logistic regression, LDA, QDA, KNN και decision tree.

Τα δεδομένα περιλαμβάνουν πληροφορίες για πελάτες, όπου η εξαρτημένη μεταβλητή είναι η churn. Στο σύνολο των 3333 δεδομένων και 20 μεταβλητών, αξιολογούνται διάφορες χαρακτηριστικά και εξετάζεται η σχέση τους με την απώλεια πελατών.

Η ανάλυση περιλαμβάνει τον μετασχηματισμό κάποιων μεταβλητών αν θα χρειαστεί, όπως η μετατροπή αριθμητικών μεταβλητών σε factor για καλύτερη επεξεργασία. Μέθοδοι όπως logistic regression, LDA, QDA, KNN και decision tree χρησιμοποιούνται για την πρόβλεψη, και τέλος επιλέγονται δύο υποψήφια προβλεπτικά μοντέλα ως τα βέλτιστα.

Περιγραφή

Λόγω της γρήγορης ανάπτυξης της επικοινωνίας δεδομένων δικτύου και την πρόοδο των πληροφοριών η τεχνολογία έχει διαθέσιμο ένα μεγάλο όγκο δεδομένων. Με την αύξηση του ανταγωνισμού στην αγορά οι εταιρείες έχουν αφιερώσει περισσότερο χρόνο για να κάνουν τους προηγούμενους πελάτες τους να πείσουν τους νέους πελάτες. Οι πελάτες παίζουν σημαντικό ρόλο στην επιβίωση και την ανάπτυξη μιας βιομηχανίας τηλεπικοινωνιών. Η διαχείριση βοηθά στη σύλληψη των πληροφοριών των καταναλωτών και ο οργανισμός χρησιμοποιεί περαιτέρω αυτές τις πληροφορίες για την βελτίωση και την ανάλυση της απόκτησης πελατών ακόμα και για την διατήρηση τους. Στην συγκεκριμένη εργασία θα ερευνήσουμε αυτό το πρόβλημα της χαμένης και μη χαμένης πελατείας. Μια εταιρεία τηλεπικοινωνιών μας ζήτησε να προβλέψουμε ένα μοντέλο και να μπορούμε να είμαστε σε θέση να προβλέψουμε πόσοι και ποιο είναι το ποσοστό απόρριψης του πελάτη σε μια εταιρεία ή σε απλούστερες λέξεις ταχύτητα με την οποία ο πελάτης θα εγκαταλείψει την εταιρεία ή την υπηρεσία για κάποιους δικούς τους λόγους. Παραδείγματα παραλήψεων πελατών περιλαμβάνουν την ακύρωση συνδρομής, το κλείσιμο λογαριασμού, η μη ανανέωση σύμβασης ή σύμβαση παροχής υπηρεσιών η ακόμα και η απόφαση αγοράς σε άλλο κατάστημα.

Το *churn* μπορεί να συμβεί εξαιτίας πολλών διαφορετικών λόγων και η ανάλυση του βοηθά να εντοπιστεί η αιτία αυτού του προβλήματος ανοίγοντας ευκαιρίες για την εφαρμογή αποτελεσματικών στρατηγικών διατήρησης. Τα δεδομένα που μας παρέχονται από την εταιρεία τηλεπικοινωνιών αποτελούν ένα σύνολο δεδομένων το οποίο αναφέρεται σε πληροφορίες πελατών της εταιρείας που κρατάνε για τους πελάτες τους και σε αυτά περιλαμβάνεται και μια μεταβλητή για την χαμένη πελατεία τους. Η μεταβλητή αυτή που βρίσκεται διαθέσιμη στο σύνολο δεδομένων διαχωρίζεται σε δύο τιμές <<True>> και <<False>>, όπου στο πρώτο αντιστοιχεί το 1 δηλαδή εννοεί πως οι πελάτες φεύγουν από την εταιρεία και το δεύτερο στο 0 δηλαδή πως οι πελάτες δεν παραμένουν στην εταιρεία. Περιλαμβάνει σαν factor μεταβλητές

οι οποίες αναγράφουν την σύμβαση του πελάτη, η καταγωγή του πελάτη και το σχέδιο φωνητικού ταχυδρομείου. Περιέχει ακόμα αριθμητικές μεταβλητές οι οποίες είναι ο τηλεφωνικός κωδικός του, τον αριθμό των μηνών που ο πελάτης έχει λογαριασμό στην εταιρεία, τις συνολικές κλήσεις, χρεώσεις και λεπτά⁴ των ημερών, των τηλεφωνικών κλήσεων κατά την διάρκεια της απογευματινής και της βραδινής ώρας και τον αριθμό των λεπτών που πέρασε στο τηλέφωνο, σε διεθνείς κλήσεις σε ένα συγκεκριμένο μήνα καθώς και τον αριθμό των κλήσεων εξυπηρέτησης πελατών που ο πελάτης πραγματοποίησε ώστε να επιλυθεί ένα συγκεκριμένο ζήτημα το οποίο μπορεί να τον απασχολούσε σε ένα συγκεκριμένο θέμα.

Στην έρευνα που θα παρουσιάσουμε θα συμπεριλάβουμε ένα μεγάλο εύρος από classification μεθόδους με σκοπό να προτείνουμε μοντέλα πρόβλεψης για την εταιρεία. Συγκεκριμένα οι μέθοδοι που θα χρησιμοποιήσουμε είναι logistic regression, LDA, QDA, KNN και decision tree method. Στην ανάλυση που θα παρουσιάσουμε στις επόμενες ενότητες, θα παρουσιάσουμε πιο αναλυτικά τα δεδομένα μας, θα επιλέξουμε τις μεταβλητές τις οποίες είναι χρήσιμες για την πρόβλεψη του μοντέλου, θα αξιολογήσουμε τα μοντέλα τα οποία θα βρούμε και θα προβάλουμε όλα τα βήματα που θα ακολουθήσουμε. Στην συνέχεια θα καταλήξουμε σε δύο καλύτερα υποψήφια προβλεπτικά μοντέλα.

Το σύνολο δεδομένων που έχουμε αποτελείται από 3333 δεδομένα και 20 μεταβλητές. Η εξαρτημένη μεταβλητή μας όπως είναι λογικό αφού είναι και η μεταβλητή που θέλουμε να εξετάσουμε, θα είναι η δίτιμη μεταβλητή churn και όλα τα άλλα χαρακτηριστικά που έχουμε για τους πελάτες μας αποτελούν τις ανεξάρτητες μεταβλητές. Στο σύνολο δεδομένων που μας παρέχεται θα μετασχηματίσουμε μόνο την μεταβλητή account.length area.code, number.vmail.messages, total.day.calls, total.eve.calls, total.night.calls, total.intl.calls, customer.service.calls οι οποίες εμφανίζονται ως integer και για καλύτερη διευκόλυνση μας θα τις μετατρέψουμε σε factor.

Θεωρία

Η πρόβλεψη μοντέλων, ειδικά στον τομέα της μηχανικής μάθησης και της στατιστικής, επικεντρώνεται στην ανάπτυξη αλγορίθμων που μπορούν να προβλέψουν την τιμή μιας εξαρτημένης μεταβλητής με βάση τις τιμές πολλών ανεξάρτητων μεταβλητών. Ακολούθως, παρέχω έναν επισκόπηση της θεωρίας πίσω από διάφορες μεθόδους πρόβλεψης μοντέλων:

1. Λογιστική Παλινδρόμηση (Logistic Regression):

- **Αντικείμενο:** Χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι δυαδική.
- **Λειτουργία:** Προσαρμόζει ένα γραμμικό μοντέλο στα δεδομένα, εφαρμόζοντας τη λογιστική συνάρτηση για πρόβλεψη πιθανοτήτων.

2. Ανάλυση Διακύμανσης (ANOVA):

- **Αντικείμενο:** Κατάλληλη για πολλαπλές κατηγορίες εξαρτημένης μεταβλητής.
- **Λειτουργία:** Αξιολογεί τη διακύμανση μεταξύ των μέσων των διάφορων ομάδων και τη διακύμανση εντός των ομάδων.

3. Κανονικά Διακυμάνσεων Διαφορετικών Στατιστικών (QDA, LDA):

- **Αντικείμενο:** Κατάλληλα για πολλαπλές κατηγορίες εξαρτημένης μεταβλητής.
- **Λειτουργία:** Βασίζονται στην παραδοχή κανονικής κατανομής των δεδομένων και προσπαθούν να κατασκευάσουν συναρτήσεις διακύμανσης μεταξύ των κατηγοριών.

4. K-Κοντινότεροι Γείτονες (K-Nearest Neighbors - KNN):

- **Αντικείμενο:** Απλή μέθοδος, κυρίως για ταξινόμηση και παλινδρόμηση.
- **Λειτουργία:** Βασίζεται στην ιδέα ότι οι παρόμοιες περιπτώσεις έχουν παρόμοιες εξαρτημένες τιμές.

5. Δέντρα Απόφασης (Decision Trees):

- **Αντικείμενο:** Κατάλληλα για ταξινόμηση και παλινδρόμηση.
- **Λειτουργία:** Δημιουργούν δομές δέντρου με βάση τα χαρακτηριστικά των δεδομένων, διαχωρίζοντας τα σε διακλαδώσεις.

Κάθε μέθοδος έχει τα πλεονεκτήματα και τους περιορισμούς της, και η επιλογή εξαρτάται από τη φύση των δεδομένων και το είδος του προβλήματος που επιδιώκεται να λυθεί.

Σχεδιαστικές Αποφάσεις

Logistic Regression

Η πρόβλεψη με τη Λογιστική Παλινδρόμηση (Logistic Regression) είναι μια ισχυρή τεχνική που χρησιμοποιείται ευρέως σε προβλήματα ταξινόμησης και πρόβλεψης με δυαδικές εξαρτημένες μεταβλητές. Αν και το όνομά της περιέχει τη λέξη "παλινδρόμηση", πρόκειται πραγματικά για έναν αλγόριθμο ταξινόμησης.

Εδώ είναι μερικά σημαντικά στοιχεία για την προβλεπτική δύναμη της Λογιστικής Παλινδρόμησης:

- Μοντέλο Πιθανοτήτων:** Η Λογιστική Παλινδρόμηση εκτιμά πιθανότητες, μετατρέποντας τα αποτελέσματα σε ένα εύρος μεταξύ 0 και 1. Αυτές οι πιθανότητες αντιπροσωπεύουν την πιθανότητα μιας εισόδου να ανήκει σε μια συγκεκριμένη κατηγορία.
- Συνάρτηση Απόφασης:** Με βάση τις πιθανότητες που υπολογίζονται, λαμβάνεται μια απόφαση σχετικά με την κατηγορία προσανατολίζοντας την απόφαση σε μια από τις δύο δυνατές κατηγορίες.
- Κατώφλι Απόφασης:** Το μοντέλο μπορεί να προσαρμοστεί για να λάβει αποφάσεις με βάση ένα κατώφλι πιθανοτήτων. Εάν η πιθανότητα υπερβεί αυτό το κατώφλι, τότε η είσοδος ταξινομείται ως μια κατηγορία, ειδάλλως σαν την άλλη.
- Αξιολόγηση Επίδοσης:** Η απόδοση του μοντέλου αξιολογείται συχνά με χρήση μετρικών όπως η ακρίβεια, η ευαισθησία, η ειδικότητα και η καμπύλη ROC. Αυτές οι μετρικές παρέχουν μια κατανόηση της ικανότητας του μοντέλου να προβλέπει σωστά και να αντιμετωπίζει τις λάθος προβλέψεις.
- Διαχείριση Συσχέτισης:** Εάν υπάρχουν υψηλά επίπεδα συσχέτισης μεταξύ των χαρακτηριστικών, η Λογιστική Παλινδρόμηση μπορεί να είναι ευαίσθητη. Η διαχείριση των συσχετίσεων απαιτεί προσεκτική επιλογή των χαρακτηριστικών ή χρήση τεχνικών όπως η επιλογή μεταβλητών.

Η Λογιστική Παλινδρόμηση παρέχει, λοιπόν, μια ευέλικτη και ισχυρή μέθοδο πρόβλεψης, και η κατανόηση των παραμέτρων του μοντέλου και των αποτελεσμάτων της είναι καθοριστική για την αποτελεσματική χρήση της.

Decision Tree

Τα Δέντρα Απόφασης (Decision Trees) αποτελούν έναν ισχυρό αλγόριθμο μηχανικής μάθησης που χρησιμοποιείται τόσο για προβλήματα ταξινόμησης όσο και παλινδρόμησης. Εδώ παρέχουμε μερικές βασικές πληροφορίες σχετικά με τα Δέντρα Απόφασης:

1. Δομή Απόφασης:

- Κόμβοι:* Κάθε δέντρο απόφασης ξεκινά με έναν αρχικό κόμβο, που αντιπροσωπεύει ολόκληρο το σύνολο δεδομένων.
- Κλάδια:* Σε κάθε κόμβο, το δέντρο κλαδεύει σε υποκόμβους βάσει κάποιου χαρακτηριστικού.
- Φύλλα:* Οι τελικοί υποκόμβοι, ή φύλλα, αντιστοιχούν σε τελικές αποφάσεις.

2. Λειτουργία:

- Διαχωρισμός:* Οι κόμβοι διαχωρίζονται βάσει των χαρακτηριστικών που έχουν τη μεγαλύτερη ικανότητα διαχωρισμού των δεδομένων.
- Επιλογή Χαρακτηριστικών:* Κατά τη διαδικασία εκπαίδευσης, το δέντρο επιλέγει τα χαρακτηριστικά που βοηθούν περισσότερο στην κατηγοριοποίηση ή πρόβλεψη.
- Κριτήρια Απόφασης:* Τα δέντρα χρησιμοποιούν κριτήρια όπως η Gini impurity ή η εντροπία για να επιλέξουν τη βέλτιστη απόφαση.

3. Πλεονεκτήματα:

- Ερμηνευσιμότητα:* Είναι ευανάγνωστα και κατανοητά από τον άνθρωπο.
- Χειρισιμότητα:* Δεν απαιτούν πολύ προεπεξεργασία των δεδομένων.

4. Προκλήσεις:

- a. *Προεπιλογή*: Η υπερβολική ανάπτυξη μπορεί να οδηγήσει σε υπερεκπαίδευση.
- b. *Εναισθησία στην Αρχική Κατάτμηση*: Μπορεί να είναι ευαίσθητα στην αρχική κατάτμηση των δεδομένων.

Συνολικά, τα Δέντρα Απόφασης αποτελούν έναν εύκολο και αποτελεσματικό τρόπο για την αντιμετώπιση προβλημάτων ταξινόμησης και παλινδρόμησης, παρέχοντας παράλληλα ερμηνευσιμότητα στα αποτελέσματα.

K Nearest Neighbor Classifier

Στην στατιστική, ο αλγόριθμος K Nearest Neighbors (KNN ή K Πλησιέστεροι Γείτονες) είναι μια μη παραμετρική μέθοδος μάθησης με επίβλεψη. Υλοποιήθηκε αρχικά από την Evelyn Fix και τον Joseph Hodges το 1951 και στην πορεία επεκτάθηκε από τον Thomas Cover. Ο αλγόριθμος KNN χρησιμοποιείται για λόγους ταξινόμησης και παλινδρόμησης, χρησιμοποιώντας ως είσοδο, ένα σύνολο k από κοντινά δεδομένα-παραδείγματα του dataset. Με το KNN εκμεταλλευόμαστε τις προηγούμενες εμφανίσεις δεδομένων, οι οποίες έχουν γνωστές τιμές εξόδου, ώστε να προβλέψουμε άγνωστη τιμή εξόδου σε νέα στιγμιότυπα δεδομένων.

Ο αλγόριθμος K Nearest Neighbor είναι μια πολύ γνωστή και ευρέως χρησιμοποιούμενη τεχνική κατηγοριοποίησης στην μηχανική μάθηση. Στον κόσμο της ταξινόμησης, η KNN επιδιώκει να αναθέσει μια κατηγορία σε ένα σημείο δεδομένων βασιζόμενη στις κατηγορίες των "κοντινότερων γειτόνων" του. Συνήθως, ορίζουμε έναν παράμετρο k, που αντιπροσωπεύει τον αριθμό των κοντινότερων γειτόνων που θα ληφθούν υπόψη. Η KNN υπολογίζει την απόσταση μεταξύ του σημείου που πρόκειται να ταξινομηθεί και των υπόλοιπων δειγμάτων στο σύνολο εκπαίδευσης και επιλέγει τις k εγγυημένα πλησιέστερες.

Η λειτουργία του KNN περιγράφεται πιο αναλυτικά παρακάτω:

1. Επιλογή της Μετρικής Απόστασης:

Κατά την εκκίνηση του αλγορίθμου, πρέπει να οριστεί μια μετρική απόσταση (π.χ., Ευκλείδεια απόσταση, Μανχάταν, Χάμινγκ, κ.λπ.). Η μετρική απόσταση

χρησιμοποιείται για να υπολογίσει το "κόστος" ή την απόσταση μεταξύ δύο σημείων στον χώρο χαρακτηριστικών.

2. Κατάταξη Κοντινότερων Γειτόνων:

Για να ταξινομήσουμε ένα νέο δείγμα, υπολογίζουμε την απόστασή του από όλα τα άλλα δείγματα στο σύνολο εκπαίδευσης, χρησιμοποιώντας την επιλεγμένη μετρική απόσταση. Στη συνέχεια, επιλέγουμε τα k δείγματα με την ελάχιστη απόσταση.

3. Ψηφοφορία:

Ταξινομούμε το νέο δείγμα στην κατηγορία που εμφανίζεται συχνότερα μεταξύ των k κοντινότερων γειτόνων. Αυτή η διαδικασία είναι γνωστή ως "ψηφοφορία των γειτόνων."

Η KNN λειτουργεί καλά σε περιπτώσεις όπου τα δείγματα της ίδιας κατηγορίας σχηματίζουν συγκεντρώσεις στον χώρο χαρακτηριστικών. Ωστόσο, πρέπει να λαμβάνεται υπόψη η επιλογή του κατάλληλου αριθμού k, καθώς και η επιλογή της κατάλληλης μετρικής απόστασης, ανάλογα με τη φύση των δεδομένων.

Για την διεξαγωγή της εργασίας μας, χρησιμοποιήσαμε μερικές μεθόδους από την κλάση «KNeighborsClassifier» της βιβλιοθήκης scikit-learn σε κώδικα python.

Random Forest Classifier

Ένας ακόμα δημοφιλής αλγόριθμος επιβλεπόμενης μηχανικής μάθησης είναι ο αλγόριθμος Random Forest, ο οποίος χρησιμοποιείται στις περιπτώσεις που έχουμε ως στόχο μια μεταβλητή με ετικέτα. Ο αλγόριθμος Random Forest είναι μια μέθοδος συνόλου, κάτι που σημαίνει ότι συνδυάζει προβλέψεις με άλλα μοντέλα και μπορεί να χρησιμοποιηθεί για την επίλυση ζητημάτων παλινδρόμησης και ταξινόμησης. Κάθε ένα από τα μικρότερα μοντέλα στο σύνολο του Random Forest είναι ένα decision tree.

Για την προσπάθεια επίλυσης του προβλήματος της εργασίας μας χρησιμοποιήσαμε τον αλγόριθμο μόνο για την επίλυση ζητημάτων ταξινόμησης. Η λειτουργία του Random Forest περιγράφεται πιο αναλυτικά παρακάτω:

1. Επιλογή Δειγμάτων:

Κατά την εκπαίδευση, ο random forest χρησιμοποιεί τη μέθοδο Bootstrap Aggregating (Bagging). Αυτό σημαίνει ότι δημιουργεί διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης με επανατοποθέτηση, προσπαθώντας να δημιουργήσει ποικιλομορφία μεταξύ των δέντρων.

2. Επιλογή Χαρακτηριστικών:

Κάθε φορά που δημιουργείται ένας κόμβος σε ένα δέντρο αποφάσεων, επιλέγεται τυχαία ένα υποσύνολο των διαθέσιμων χαρακτηριστικών για τον κόμβο. Αυτό επιπλέον αυξάνει την ποικιλομορφία μεταξύ των δέντρων.

3. Δημιουργία Δέντρων Αποφάσεων:

Για κάθε υποσύνολο δεδομένων, δημιουργείται ένα δέντρο αποφάσεων (decision tree). Τα δέντρα αυτά λαμβάνουν αποφάσεις βάσει των χαρακτηριστικών που επιλέγονται τυχαία σε κάθε κόμβο.

4. Ψηφοφορία:

Κατά την ταξινόμηση ενός νέου δείγματος, όλα τα δέντρα στο τυχαίο δάσος συμβάλλουν στην απόφαση. Για προβλήματα ταξινόμησης, χρησιμοποιείται η ψηφοφορία των δέντρων για να επιλεγεί η τελική κατηγορία.

Η συνολική διαδικασία του random forest βοηθά στον περιορισμό του overfitting, ενισχύοντας τη γενίκευση του μοντέλου. Επίσης, η τυχαιότητα στην επιλογή δειγμάτων και χαρακτηριστικών εξασφαλίζει ότι κάθε δέντρο είναι μοναδικό και συνεισφέρει με διαφορετικό τρόπο στο τελικό μοντέλο.

Για την διεξαγωγή της εργασίας μας, χρησιμοποιήσαμε μερικές μεθόδους από την κλάση «RandomForestClassifier» της βιβλιοθήκης scikit-learn σε κώδικα python.

Ανάπτυξη

Η υλοποίηση της εργασίας μας έγινε στο ολοκληρωμένο περιβάλλον ανάπτυξης (Integrated Development Environment – IDE) PyCharm, όπου χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python.

Αρχικά, εισαγάγαμε τις απαραίτητες βιβλιοθήκες για την υλοποίηση:

```
1 import pandas as pd
2 import numpy as np
3 import time
4 import matplotlib.pyplot as plt
5 import seaborn as sns
6 from sklearn import metrics
7 from sklearn.tree import DecisionTreeClassifier, export_text, export_graphviz
8 from statsmodels.graphics.gofplots import qqplot
9 import scipy.stats as stats
10 from scipy.stats import pearsonr, spearmanr, norm, chi2_contingency, kurtosis, skew
11 from sklearn.covariance import EllipticEnvelope
12 from sklearn.linear_model import LogisticRegression
13 from sklearn.model_selection import train_test_split
14 from sklearn.metrics import confusion_matrix, accuracy_score, precision_score
15 from scipy.stats import anderson
16 #KNN Class import from Scikit Learn
17 from sklearn.neighbors import KNeighborsClassifier
18 #Random Forest class import from Scikit Learn
19 from sklearn.ensemble import RandomForestClassifier
```

Εικόνα 1 – Βιβλιοθήκες

Στην συνέχεια, προσθέσαμε το dataset που θέλαμε να χρησιμοποιήσουμε στον κώδικα μας και κάναμε κάποιες μετατροπές στα δεδομένα που θα μας ήταν χρήσιμες στην πορεία:

```

21  ► data = pd.read_csv('telecom_churn.csv')
22  print(data.columns)
23  print(data.describe())
24  print(data.dtypes)
25
26  # Converts data to numbers - errors='coerce' is used for any data that can't be converted - NaN
27  data['Account length'] = pd.to_numeric(data['Account length'], errors='coerce')
28  data['Area code'] = pd.to_numeric(data['Area code'], errors='coerce')
29  data['Number vmail messages'] = pd.to_numeric(data['Number vmail messages'], errors='coerce')
30  data['Total day calls'] = pd.to_numeric(data['Total day calls'], errors='coerce')
31  data['Total eve calls'] = pd.to_numeric(data['Total eve calls'], errors='coerce')
32  data['Total night calls'] = pd.to_numeric(data['Total night calls'], errors='coerce')
33  data['Total intl calls'] = pd.to_numeric(data['Total intl calls'], errors='coerce')
34  data['Customer service calls'] = pd.to_numeric(data['Customer service calls'], errors='coerce')
35
36  print(data.info())
37  print(data.isnull().sum())

```

Εικόνα 2 - Προετοιμασία dataset

Παρακάτω, δημιουργήσαμε κάποια γραφήματα για να κατανοήσουμε καλύτερα τα δεδομένα που έχουμε διαθέσιμα:

```

39  #The percentage of each state in the dataset
40  state_counts = data['State'].value_counts(normalize=True) * 100
41  plt.figure(figsize=(10, 6))
42  sns.barplot(x=state_counts.index, y=state_counts.values)
43  plt.xlabel('Πολιτεία')
44  plt.ylabel('Ποσοστό (%)')
45  plt.title('Πολιτεία')
46  plt.xticks(rotation=90)
47  plt.show()
48
49  time.sleep(5)
50
51  #The percentage of the people having vs not having international plan in the dataset
52  international_plan_counts = (data['International plan'].value_counts(normalize=True) * 100).sort_index()
53  plt.figure(figsize=(10, 6))
54  plt.bar(international_plan_counts.index, international_plan_counts.values, color='blue')
55  plt.xlabel('International Plan')
56  plt.ylabel('Ποσοστό (%)')
57  plt.title('Ποσοστό International Plan')
58  plt.xticks(rotation=90)
59  plt.show()
60
61  time.sleep(5)

```

Εικόνα 3 - Ποσοστά (Πολιτεία & International Plan)

```

63 #The percentage of people having vs not having a voice mail plan in the dataset
64 voice_mail_plan_counts = (data['Voice mail plan'].value_counts(normalize=True) * 100).sort_index()
65 plt.figure(figsize=(10, 6))
66 plt.bar(voice_mail_plan_counts.index, voice_mail_plan_counts.values, color='green')
67 plt.xlabel('Voice Mail Plan')
68 plt.ylabel('Ποσοστό (%)')
69 plt.title('Ποσοστό Voice Mail Plan')
70 plt.xticks(rotation=90)
71 plt.show()
72
73 time.sleep(5)
74
75 #The percentage of churn values in the dataset
76 churn_counts = (data['Churn'].value_counts(normalize=True) * 100).sort_index()
77 plt.figure(figsize=(6, 6))
78 plt.pie(churn_counts.values, labels=churn_counts.index, autopct='%1.1f%%', startangle=140)
79 plt.title('Ποσοστό Churn')
80 plt.show()
81
82 time.sleep(5)

```

Εικόνα 4 - Ποσοστά (Voice Mail Plan & Churn)

```

84 #The frequency of each Account Length
85 plt.figure(figsize=(10, 6))
86 plt.hist(data['Account length'], bins=30, color='gray', alpha=0.7, edgecolor='black')
87 plt.axvline(np.mean(data['Account length']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
88 plt.axvline(np.median(data['Account length']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
89 plt.xlabel('Μήκος Λογαριασμού')
90 plt.ylabel('Συχνότητα')
91 plt.title('Κατανούμε Μήκους Λογαριασμού')
92 plt.legend()
93 plt.show()
94 plt.figure(figsize=(8, 6))
95 stats.probplot(data['Account length'], dist='norm', plot=plt)
96 plt.title('QQ-plot για Μήκος Λογαριασμού')
97 plt.show()
98 time.sleep(5)
99 #The frequency of each Area code
100 plt.figure(figsize=(10, 6))
101 plt.hist(data['Area code'], bins=30, color='gray', alpha=0.7, edgecolor='black')
102 plt.axvline(np.mean(data['Area code']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
103 plt.axvline(np.median(data['Area code']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
104 plt.xlabel('Κωδικός Περιοχής')
105 plt.ylabel('Συχνότητα')
106 plt.title('Κατανούμε Κωδικού Περιοχής')
107 plt.legend()
108 plt.show()
109 plt.figure(figsize=(8, 6))
110 stats.probplot(data['Area code'], dist='norm', plot=plt)
111 plt.title('QQ-plot για Κωδικό Περιοχής')
112 plt.show()
113 time.sleep(5)

```

Εικόνα 5 - Συχνότητα (Account Length & Area Code)

```

118 #The frequency of total day minutes
119 plt.figure(figsize=(10, 6))
120 plt.hist(data['Total day minutes'], bins=30, color='gray', alpha=0.7, edgecolor='black')
121 plt.axvline(np.mean(data['Total day minutes']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
122 plt.axvline(np.median(data['Total day minutes']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
123 plt.xlabel('Συνολικά Λεπτά Ημέρας')
124 plt.ylabel('Συχνότητα')
125 plt.title('Κατανομή Συνολικών Λεπτών Ημέρας')
126 plt.legend()
127 plt.show()
128 plt.figure(figsize=(8, 6))
129 stats.probplot(data['Total day minutes'], dist='norm', plot=plt)
130 plt.title('QQ-plot για Συνολικά Λεπτά Ημέρας')
131 plt.show()
132 time.sleep(5)
133 #The frequency of total day calls
134 plt.figure(figsize=(10, 6))
135 plt.hist(data['Total day calls'], bins=30, color='gray', alpha=0.7, edgecolor='black')
136 plt.axvline(np.mean(data['Total day calls']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
137 plt.axvline(np.median(data['Total day calls']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
138 plt.xlabel('Συνολικά Τηλέφωνα Ημέρας')
139 plt.ylabel('Συχνότητα')
140 plt.title('Κατανομή Συνολικών Τηλεφώνων Ημέρας')
141 plt.legend()
142 plt.show()
143 plt.figure(figsize=(8, 6))
144 stats.probplot(data['Total day calls'], dist='norm', plot=plt)
145 plt.title('QQ-plot για Συνολικά Τηλέφωνα Ημέρας')
146 plt.show()
147 time.sleep(5)

```

Εικόνα 6 - Συχνότητα (Total Day Min & Total Day Calls)

```

152 #The frequency of total day charge
153 plt.figure(figsize=(10, 6))
154 plt.hist(data['Total day charge'], bins=30, color='gray', alpha=0.7, edgecolor='black')
155 plt.axvline(np.mean(data['Total day charge']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
156 plt.axvline(np.median(data['Total day charge']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
157 plt.xlabel('Συνολική Χρέωση Ημέρας')
158 plt.ylabel('Συχνότητα')
159 plt.title('Κατανομή Συνολικής Χρέωσης Ημέρας')
160 plt.legend()
161 plt.show()
162 plt.figure(figsize=(8, 6))
163 stats.probplot(data['Total day charge'], dist='norm', plot=plt)
164 plt.title('QQ-plot για Συνολική Χρέωση Ημέρας')
165 plt.show()
166 time.sleep(5)
167 #The frequency of total eve calls
168 plt.figure(figsize=(10, 6))
169 plt.hist(data['Total eve calls'], bins=30, color='gray', alpha=0.7, edgecolor='black')
170 plt.axvline(np.mean(data['Total eve calls']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
171 plt.axvline(np.median(data['Total eve calls']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
172 plt.xlabel('Συνολικά Τηλέφωνα Απογεύματος')
173 plt.ylabel('Συχνότητα')
174 plt.title('Κατανομή Συνολικών Τηλεφώνων Απογεύματος')
175 plt.legend()
176 plt.show()
177 plt.figure(figsize=(8, 6))
178 stats.probplot(data['Total eve calls'], dist='norm', plot=plt)
179 plt.title('QQ-plot για Συνολικά Τηλέφωνα Απογεύματος')
180 plt.show()
181 time.sleep(5)

```

Εικόνα 7 - Συχνότητα (Total Day Charge & Total Eve Calls)

```

186 #The frequency of total eve minutes
187 plt.figure(figsize=(10, 6))
188 plt.hist(data['Total eve minutes'], bins=30, color='gray', alpha=0.7, edgecolor='black')
189 plt.axvline(np.mean(data['Total eve minutes']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
190 plt.axvline(np.median(data['Total eve minutes']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
191 plt.xlabel('Συνολικά Λεπτά Απογεύματος')
192 plt.ylabel('Συχνότητα')
193 plt.title('Κατανομή Συνολικών Λεπτών Απογεύματος')
194 plt.legend()
195 plt.show()
196 plt.figure(figsize=(8, 6))
197 stats.probplot(data['Total eve minutes'], dist='norm', plot=plt)
198 plt.title('QQ-plot για Συνολικά Λεπτά Απογεύματος')
199 plt.show()
200 time.sleep(5)
201 #The frequency of total eve charge
202 plt.figure(figsize=(10, 6))
203 plt.hist(data['Total eve charge'], bins=30, color='gray', alpha=0.7, edgecolor='black')
204 plt.axvline(np.mean(data['Total eve charge']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
205 plt.axvline(np.median(data['Total eve charge']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
206 plt.xlabel('Συνολική Χρέωση Απογεύματος')
207 plt.ylabel('Συχνότητα')
208 plt.title('Κατανομή Συνολικής Χρέωσης Απογεύματος')
209 plt.legend()
210 plt.show()
211 plt.figure(figsize=(8, 6))
212 stats.probplot(data['Total eve charge'], dist='norm', plot=plt)
213 plt.title('QQ-plot για Συνολική Χρέωση Απογεύματος')
214 plt.show()
215 time.sleep(5)

```

Εικόνα 8 - Συχνότητα (Total Eve Mins & Total Eve Charge)

```

220 #The frequency of total night minutes
221 plt.figure(figsize=(10, 6))
222 plt.hist(data['Total night minutes'], bins=30, color='gray', alpha=0.7, edgecolor='black')
223 plt.axvline(np.mean(data['Total night minutes']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
224 plt.axvline(np.median(data['Total night minutes']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
225 plt.xlabel('Συνολικά Λεπτά το Βράδυ')
226 plt.ylabel('Συχνότητα')
227 plt.title('Κατανομή Συνολικών Λεπτών το Βράδυ')
228 plt.legend()
229 plt.show()
230 plt.figure(figsize=(8, 6))
231 stats.probplot(data['Total night minutes'], dist='norm', plot=plt)
232 plt.title('QQ-plot για Συνολικά Λεπτά το Βράδυ')
233 plt.show()
234 time.sleep(5)
235 #The frequency of total night calls
236 plt.figure(figsize=(10, 6))
237 plt.hist(data['Total night calls'], bins=30, color='gray', alpha=0.7, edgecolor='black')
238 plt.axvline(np.mean(data['Total night calls']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
239 plt.axvline(np.median(data['Total night calls']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
240 plt.xlabel('Συνολικά Τηλεφώνα το Βράδυ')
241 plt.ylabel('Συχνότητα')
242 plt.title('Κατανομή Συνολικών Τηλεφώνων το Βράδυ')
243 plt.legend()
244 plt.show()
245 plt.figure(figsize=(8, 6))
246 stats.probplot(data['Total night calls'], dist='norm', plot=plt)
247 plt.title('QQ-plot για Συνολικά Τηλέφωνα το Βράδυ')
248 plt.show()
249 time.sleep(5)

```

Εικόνα 9 - Συχνότητα (Total Night Mins & Total Night Calls)

```

254 #The frequency of total night charge
255 plt.figure(figsize=(10, 6))
256 plt.hist(data['Total night charge'], bins=30, color='gray', alpha=0.7, edgecolor='black')
257 plt.axvline(np.mean(data['Total night charge']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
258 plt.axvline(np.median(data['Total night charge']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
259 plt.xlabel('Συνολική Χρέωση το Βράδυ')
260 plt.ylabel('Συχνότητα')
261 plt.title('Κατανομή Συνολικής Χρέωσης το Βράδυ')
262 plt.legend()
263 plt.show()
264 plt.figure(figsize=(8, 6))
265 stats.probplot(data['Total night charge'], dist='norm', plot=plt)
266 plt.title('QQ-plot για Συνολική Χρέωση το Βράδυ')
267 plt.show()
268 time.sleep(5)
269 #The frequency of total international calls
270 plt.figure(figsize=(10, 6))
271 plt.hist(data['Total intl calls'], bins=30, color='gray', alpha=0.7, edgecolor='black')
272 plt.axvline(np.mean(data['Total intl calls']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
273 plt.axvline(np.median(data['Total intl calls']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
274 plt.xlabel('Συνολικός Αριθμός Διεθνών κλήσεων')
275 plt.ylabel('Συχνότητα')
276 plt.title('Συνολικός Αριθμός Διεθνών κλήσεων')
277 plt.legend()
278 plt.show()
279 plt.figure(figsize=(8, 6))
280 stats.probplot(data['Total intl calls'], dist='norm', plot=plt)
281 plt.title('QQ-plot για Συνολικό Αριθμό Διεθνών κλήσεων')
282 plt.show()
283 time.sleep(5)

```

Εικόνα 10 - Συχνότητα (Total Night Charge & Total International Calls)

```

288 #The frequency of total international minutes
289 plt.figure(figsize=(10, 6))
290 plt.hist(data['Total intl minutes'], bins=30, color='gray', alpha=0.7, edgecolor='black')
291 plt.axvline(np.mean(data['Total intl minutes']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
292 plt.axvline(np.median(data['Total intl minutes']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
293 plt.xlabel('Συνολικά Λεπτά Διεθνών κλήσεων')
294 plt.ylabel('Συχνότητα')
295 plt.title('Συνολικά Λεπτά Διεθνών κλήσεων')
296 plt.legend()
297 plt.show()
298 plt.figure(figsize=(8, 6))
299 stats.probplot(data['Total intl minutes'], dist='norm', plot=plt)
300 plt.title('QQ-plot για Συνολικά Λεπτά Διεθνών κλήσεων')
301 plt.show()
302 time.sleep(5)
303 #The frequency of total international charge
304 plt.figure(figsize=(10, 6))
305 plt.hist(data['Total intl charge'], bins=30, color='gray', alpha=0.7, edgecolor='black')
306 plt.axvline(np.mean(data['Total intl charge']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
307 plt.axvline(np.median(data['Total intl charge']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
308 plt.xlabel('Συνολική Χρέωση Διεθνών κλήσεων')
309 plt.ylabel('Συχνότητα')
310 plt.title('Συνολική Χρέωση Διεθνών κλήσεων')
311 plt.legend()
312 plt.show()
313 plt.figure(figsize=(8, 6))
314 stats.probplot(data['Total intl charge'], dist='norm', plot=plt)
315 plt.title('QQ-plot για Συνολική Χρέωση Διεθνών κλήσεων')
316 plt.show()
317 time.sleep(5)

```

Εικόνα 11 - Συχνότητα (Total International Calls & Total International Charge)

```

322 #The frequency of customer service calls
323 plt.figure(figsize=(10, 6))
324 plt.hist(data['Customer service calls'], bins=30, color='gray', alpha=0.7, edgecolor='black')
325 plt.axvline(np.mean(data['Customer service calls']), color='red', linestyle='dashed', linewidth=2, label='Μέση')
326 plt.axvline(np.median(data['Customer service calls']), color='blue', linestyle='dashed', linewidth=2, label='Διάμεσο')
327 plt.xlabel('Συνολικός Αριθμός Τηλεφώνων προς την Εξυπρέτηση Πελατών')
328 plt.ylabel('Συχνότητα')
329 plt.title('Συνολικός Αριθμός Τηλεφώνων προς την Εξυπρέτηση Πελατών')
330 plt.legend()
331 plt.show()
332 plt.figure(figsize=(8, 6))
333 stats.probplot(data['Customer service calls'], dist='norm', plot=plt)
334 plt.title('QQ-plot για Συνολικό Αριθμό Τηλεφώνων προς την Εξυπρέτηση Πελατών')
335 plt.show()
336 |
337 time.sleep(5)

```

Eικόνα 12 - Συχνότητα Customer Service Calls

```

399 # Correlation
400 numeric_columns = data.select_dtypes(include=np.number).columns
401 summary_stats = data[numeric_columns].describe().transpose()
402 print(round(summary_stats, 2))
403 rounded_data = np.round(data[numeric_columns], 2)
404 print("Rounded data:")
405 print(rounded_data.describe())
406 # Υπολογισμός πίνακα συσχέτισης Pearson
407 correlation_pearson = data[numeric_columns].corr(method='pearson')
408 print("Correlation Pearson:")
409 print(correlation_pearson)
410 # Σχεδίαση πίνακα συσχέτισης
411 plt.figure(figsize=(10, 8))
412 sns.heatmap(correlation_pearson, annot=True, cmap='coolwarm', fmt=".2f")
413 plt.title("Pearson Correlation Heatmap")
414 plt.show()
415 # Υπολογισμός πίνακα συσχέτισης Spearman
416 correlation_spearman = data[numeric_columns].corr(method='spearman')
417 print(correlation_spearman)
418 # Σχεδίαση πίνακα συσχέτισης Spearman
419 plt.figure(figsize=(10, 8))
420 sns.heatmap(correlation_spearman, annot=True, cmap='coolwarm', fmt=".2f")
421 plt.title("Spearman Correlation Heatmap")
422 plt.show()

```

Eικόνα 13 - Correlation & Spearman

```

429 # Υπολογισμός συσχέτισης και εκτέλεση του Τεστ
430 correlation_minutes_charge = pearsonr(data['Total day minutes'], data['Total day charge'])
431 print(f"Correlation between Total day minutes and Total day charge: {correlation_minutes_charge[0]:.4f}")
432 # Τεστ για συσχέτιση με χρόνο cor.test
433 cor_test_minutes_charge = spearmanr(data['Total day minutes'], data['Total day charge'])
434 print(cor_test_minutes_charge)
435 # Εύρεση εκτός των κύριων ακτίνων με Mahalanobis
436 features = ['Total day minutes', 'Total day calls', 'Total day charge']
437 envelope = EllipticEnvelope()
438 envelope.fit(data[features])
439 outliers = envelope.predict(data[features])
440 # Ποσοστό εκτός των κύριων ακτίνων
441 outlier_percentage = (sum(outliers == -1) / len(data)) * 100
442 print(f"Percentage of outliers: {outlier_percentage:.1f}%")
443 # Εγγύηση boxplot
444 plt.figure(figsize=(10, 6))
445 sns.boxplot(data=data[numerical_columns])
446 plt.title("Boxplot of Numeric Variables")
447 plt.xticks(rotation=45)
448 plt.show()

```

Εικόνα 14 - Correlation & Spearman (2)

```

455 # Εύρεση ακραίων τιμών
456 for col in numerical_columns:
457     # Πάρτε τις γραμμές του boxplot
458     lines = sns.boxplot(data=data[col], showfliers=False).get_lines()
459
460     # Πάρτε τα δεδουμένα των ακραίων τιμών
461     outliers = lines[0].get_ydata()
462
463     if len(outliers) != 0:
464         print('-----')
465         print(f'Outliers for variable {col}')
466         print(f'{len(outliers)} outliers')
467         print(f'{round(100 * len(outliers) / len(data[col]), 1)}% outliers')
468         print(outliers)
469

```

Εικόνα 15 - Εύρεση ακραίων τιμών

```

471 # --- Εκτέλεση των Chi-Square Tests και εμφάνιση των contingency tables
472
473 @mpl2205 *
474 def chi_square_test(cross_tab):
475     chi2, p, _, _ = chi2_contingency(cross_tab)
476     print(f"Chi-Square Value: {chi2:.4f}")
477     print(f"P-value: {p:.4f}")
478     print("")
479
480 # Churn vs State
481 cross_tab_state = pd.crosstab(data['Churn'], data['State'])
482 chi_square_test(cross_tab_state)
483
484 # Churn vs International Plan
485 cross_tab_international_plan = pd.crosstab(data['Churn'], data['International plan'])
486 chi_square_test(cross_tab_international_plan)
487
488 # Churn vs Voice Mail Plan
489 cross_tab_voice_mail_plan = pd.crosstab(data['Churn'], data['Voice mail plan'])
490 chi_square_test(cross_tab_voice_mail_plan)
491

```

Εικόνα 16 - Chi-Square Tests

```

531     #--- Skewness & kurtosis
532
533     skewness = data[numerical_columns].apply(skew)
534     kurt = data[numerical_columns].apply(kurtosis)
535
536     print("Skewness:")
537     print(skewness)
538
539     print("\nKurtosis:")
540     print(kurt)

```

Eικόνα 17 - Skewness & Kurtosis

```

543     # --- Testing for normality
544     numerical_columns = data.select_dtypes(include=np.number)
545     y = numerical_columns
546
547     for col in y.columns:
548         ks_stat, ks_p_value = stats.kstest(y[col], 'norm')
549         print(f"Kolmogorov-Smirnov test for {col}: KS Statistic = {ks_stat}, p-value = {ks_p_value}")
550
551     anderson_test_results = y.apply(lambda x: anderson(x).statistic)
552     print("Anderson-Darling test results:")
553     print(anderson_test_results)
554
555     shapiro_test_results = y.apply(stats.shapiro)
556     print("Shapiro-Wilk test results:")
557     print(shapiro_test_results)
558
559     t_test_results = [stats.ttest_1samp(y[col], 0) for col in y.columns]
560     print("One-sample t-test results:")
561     print(t_test_results)

```

Eικόνα 18 - Normality Test

Στην συνέχεια, υλοποιήσαμε τον κώδικα για τέσσερις αλγορίθμους που χρησιμοποιήσαμε για την διεξαγωγή της εργασίας:

1. Logistic Regression

```

565 # Logistic Regression
566
567 # Διαχωρισμός των ανεξάρτητων (X) και εξαρτημένης (y) μεταβλητής
568 data_encoded = pd.get_dummies(data, columns=['International plan', 'Voice mail plan', 'State'])
569
570 # Διαχωρισμός των ανεξάρτητων (X) και εξαρτημένης (y) μεταβλητής
571 X = data_encoded.drop('Churn', axis=1)
572 y = data_encoded['Churn']
573
574 # Ορισμός του μοντέλου Logistic Regression
575 logreg = LogisticRegression(solver='liblinear', max_iter=1000)
576
577 # Εκπαίδευση του μοντέλου
578 logreg.fit(X, y)
579
580 # Εκτύπωση των συντελεστών
581 print("Coefficients:")
582 print(logreg.coef_)
583
584 # Προβλέψεις πιθανοτήτων
585 y_pred_probs = logreg.predict_proba(X)[:, 1]
586 print("Predicted Probabilities:")
587 print(y_pred_probs[:10])

```

Eικόνα 19 - Logistic Regression (1)

```

598 # Υπολογισμός του classification report
599 classification_report = metrics.classification_report(y, y_pred_class)
600 print("Classification Report:")
601 print(classification_report)
602
603 # Υπολογισμός του training error rate
604 train_error_rate = 1 - logreg.score(X, y)
605 print(f"Training Error Rate: {train_error_rate}")
606
607 # Διαχωρισμός των δεδομένων σε training και test set
608 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=1)
609
610 # Εκπαίδευση του μοντέλου στα training data
611 logreg.fit(X_train, y_train)
612
613 # Προβλέψεις πιθανοτήτων για τα test data
614 y_pred_probs_test = logreg.predict_proba(X_test)[:, 1]
615
616 # Κατωφλιοποίηση των προβλέψεων για τα test data
617 y_pred_class_test = np.where(y_pred_probs_test > threshold, 1, 0)
618
619 # Υπολογισμός του confusion matrix για τα test data
620 conf_matrix_test = metrics.confusion_matrix(y_test, y_pred_class_test)
621 print("Confusion Matrix (Test Data):")
622 print(conf_matrix_test)

```

Eικόνα 20 - Logistic Regression (2)

2. Decision Trees

```

626 # Fitting Classification Trees
627
628 tree_mod = DecisionTreeClassifier(random_state=42)
629
630 data_encoded = pd.get_dummies(data, columns=['International plan', 'Voice mail plan', 'State'])
631
632 X = data_encoded.drop('Churn', axis=1)
633 y = data_encoded['Churn']
634 tree_mod.fit(X, y)
635
636 # Display summary
637 tree_rules = export_text(tree_mod, feature_names=list(X.columns))
638 print(tree_rules)
639
640 # Display the tree
641 export_graphviz(tree_mod, out_file='tree.dot', feature_names=list(X.columns),
642                  class_names=['False', 'True'], filled=True, rounded=True,
643                  special_characters=True)
644
645 # Calculate test classification
646 np.random.seed(2)
647 train_idx = np.random.choice(range(len(data)), int(len(data)/3), replace=False)
648 test_idx = np.setdiff1d(range(len(data)), train_idx)
649
650 data_test = data.iloc[test_idx]
651 churn_test = data['Churn'].iloc[test_idx]

```

Etkóva 21 - Decision Trees (1)

```

653 tree_pred = tree_mod.predict(X.iloc[test_idx])
654 conf_matrix = confusion_matrix(churn_test, tree_pred)
655 print(conf_matrix)
656 accuracy = np.sum(np.diag(conf_matrix)) / np.sum(conf_matrix)
657 print(f'Accuracy: {accuracy}')
658
659 # Pruning the tree
660 np.random.seed(3)
661 tree_cv = DecisionTreeClassifier(random_state=42)
662 tree_cv.fit(X.iloc[train_idx], y.iloc[train_idx])
663
664 prune_path = tree_cv.cost_complexity_pruning_path(X.iloc[train_idx], y.iloc[train_idx])
665 alphas = prune_path ccp_alphas
666
667 for alpha in alphas:
668     pruned_tree = DecisionTreeClassifier(random_state=42, ccp_alpha=alpha)
669     pruned_tree.fit(X.iloc[train_idx], y.iloc[train_idx])
670
671     prune_pred = pruned_tree.predict(X.iloc[test_idx])
672     prune_conf_matrix = confusion_matrix(churn_test, prune_pred)
673     accuracy = np.sum(np.diag(prune_conf_matrix)) / np.sum(prune_conf_matrix)
674     print(f'Alpha: {alpha}, Accuracy: {accuracy}')
675

```

Etkóva 22 - Decision Trees (2)

3. K-Nearest Neighbor

```
678 #K-Nearest Neighbor Classification
679 print("\nK-Nearest Neighbors Classifier\n")
680 data_encoded = pd.get_dummies(data, columns=['International plan', 'Voice mail plan', 'State'])
681
682 X = data_encoded.drop('Churn', axis=1)
683 y = data_encoded['Churn']
684
685 random_state = int(time.time())
686 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_state)
687
688 neigh = KNeighborsClassifier(n_neighbors=3)
689 neigh.fit(X_train, y_train)
690
691 # Predictions on the test set
692 y_pred = neigh.predict(X_test)
693
694 # Print the predicted labels for the test set
695 print("Predicted Labels for the Test Set:")
696 print(y_pred)
697
698 # Compute accuracy
699 accuracy = accuracy_score(y_test, y_pred)
700 print(f"Accuracy: {accuracy:.2%}")
701
702 # Compute precision
703 precision = precision_score(y_test, y_pred)
704 print(f"Precision: {precision:.2%}")
```

Eukόva 23 - K-Nearest Neighbor

4. Random Forest

```
706 #Random Forest Classification
707 print("\nRandom Forest Classifier\n")
708 data_encoded = pd.get_dummies(data, columns=['International plan', 'Voice mail plan', 'State'])
709
710 X = data_encoded.drop('Churn', axis=1)
711 y = data_encoded['Churn']
712
713 random_state = int(time.time())
714 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_state)
715
716 # Initialize and train the Random Forest Classifier
717 rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
718 rf_classifier.fit(X_train, y_train)
719
720 # Predictions on the test set
721 y_pred = rf_classifier.predict(X_test)
722 print("Predicted Labels for the Test Set:")
723 print(y_pred)
724
725 # Compute accuracy
726 accuracy = accuracy_score(y_test, y_pred)
727 print(f"Accuracy: {accuracy:.2%}")
728
729 # Compute precision
730 precision = precision_score(y_test, y_pred)
731 print(f"Precision: {precision:.2%}")
```

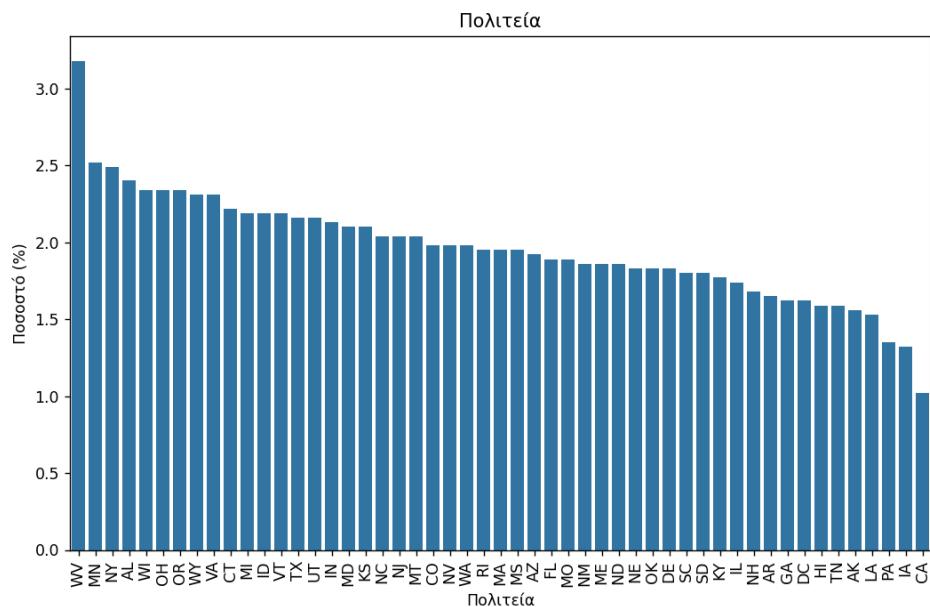
Eukόva 24 - Random Forest

Εκτέλεση

Παρακάτω, παραθέτουμε τα αποτελέσματα από την εκτέλεση του κώδικα μας.

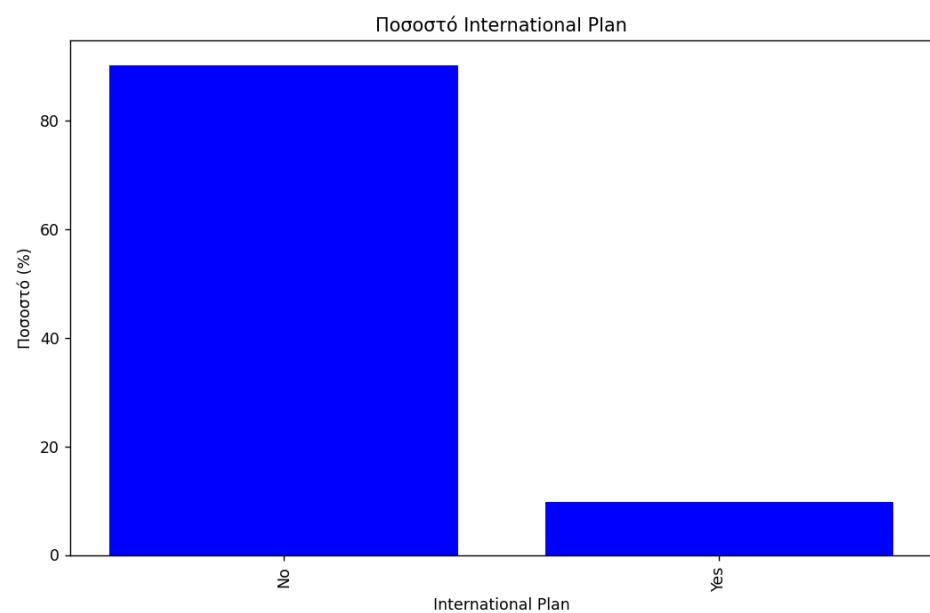
1. Γραφήματα:

Figure 1



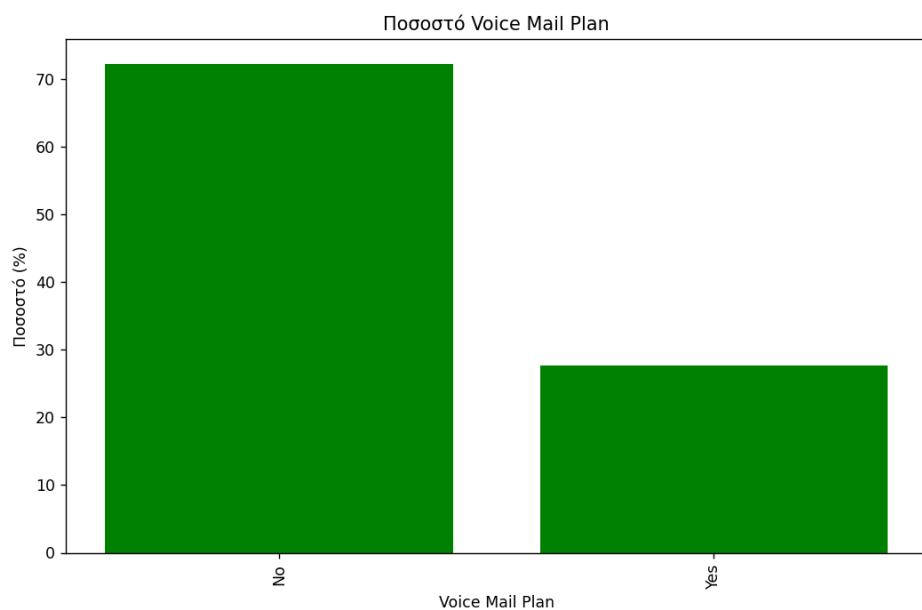
Εικόνα 25 - Ποσοστά Πολιτειών

Figure 1



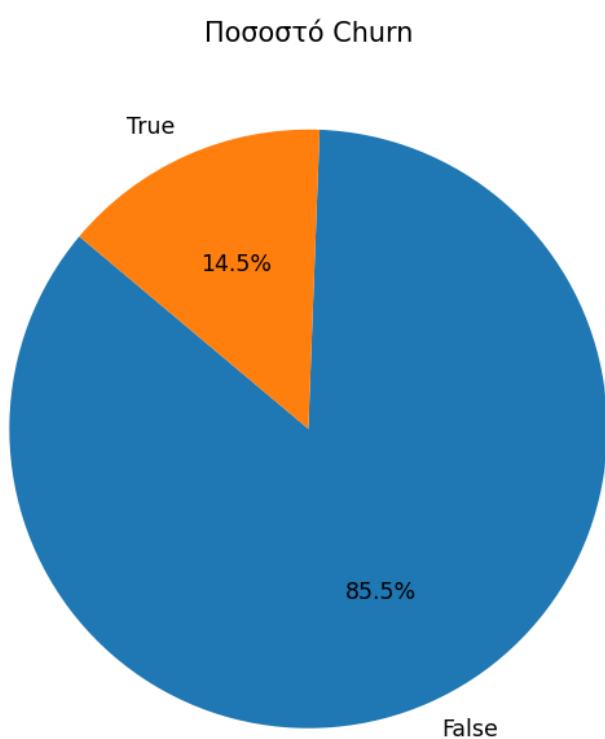
Εικόνα 26 - Ποσοστό International Plan

Figure 1



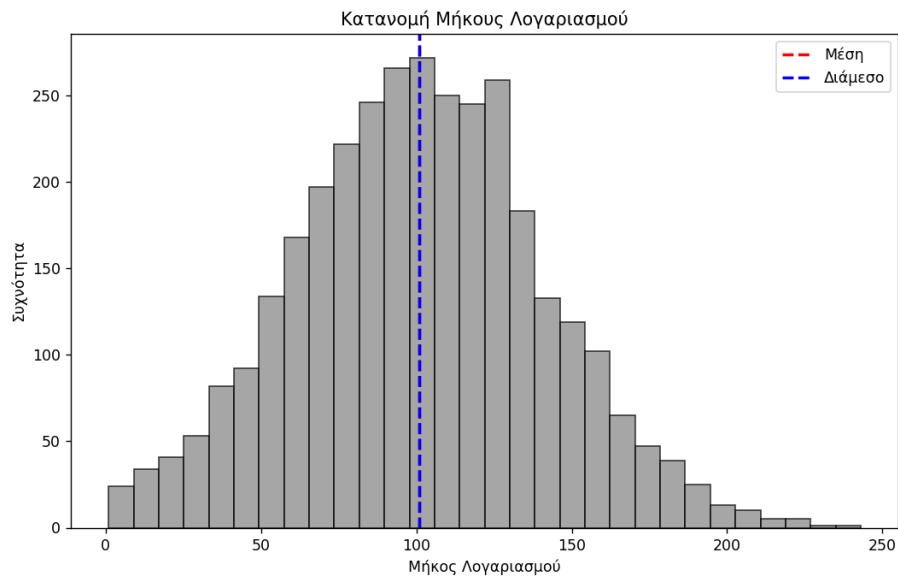
Εικόνα 27 - Ποσοστό Voice Mail Plan

Figure 1



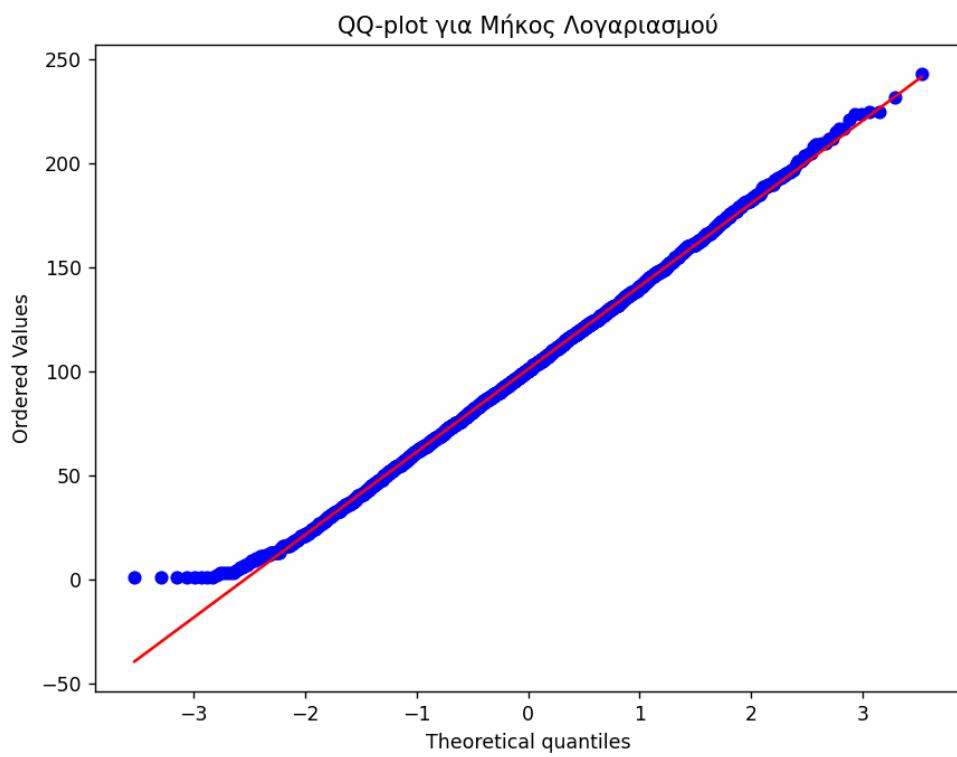
Εικόνα 28 - Ποσοστό Churn (Target Value)

Figure 1



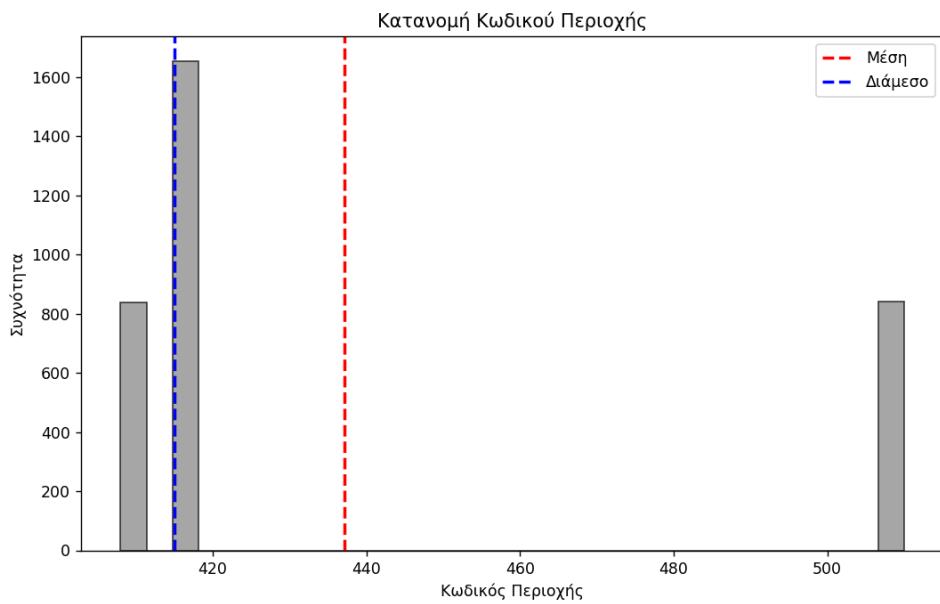
Εικόνα 29 - Κατανομή Μήκους Λογαριασμού

Figure 1



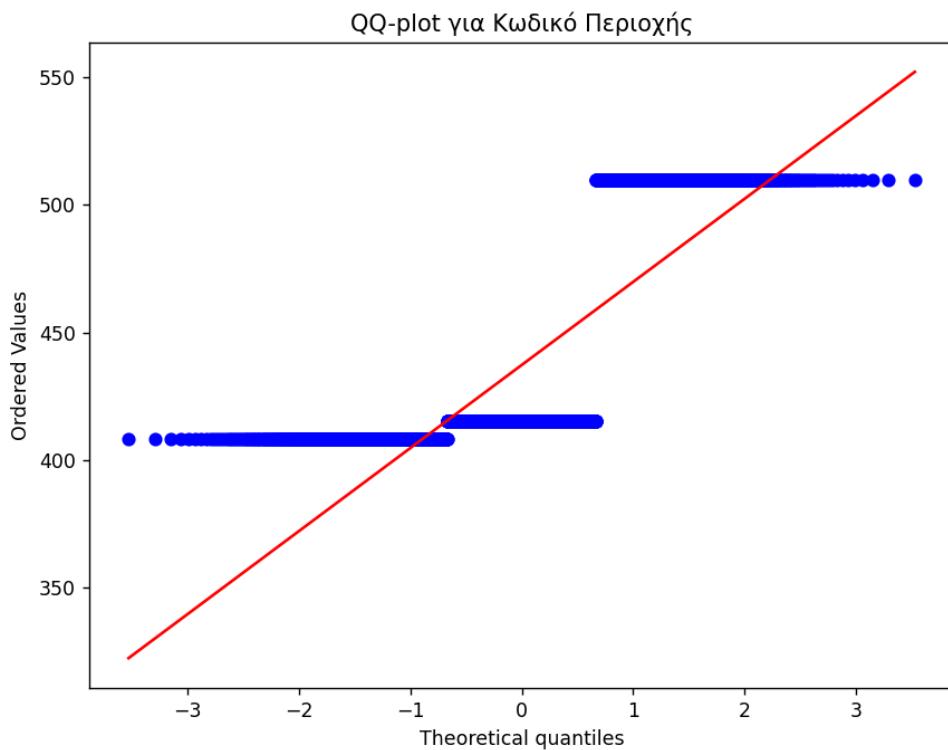
Εικόνα 30 - QQ Plot (Μήκος Λογαριασμού)

Figure 1



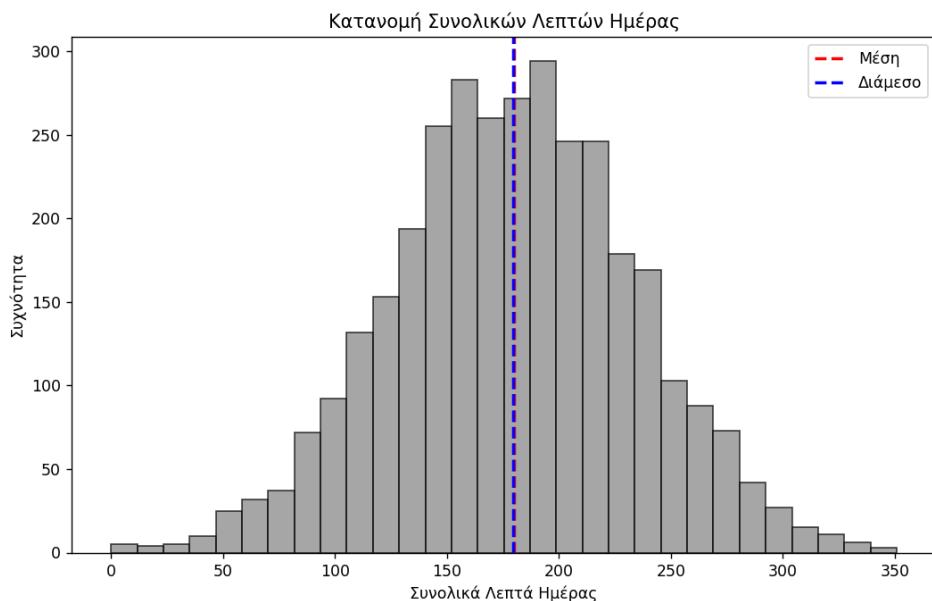
Εικόνα 31 - Κατανομή Κωδικού Περιοχής

Figure 1



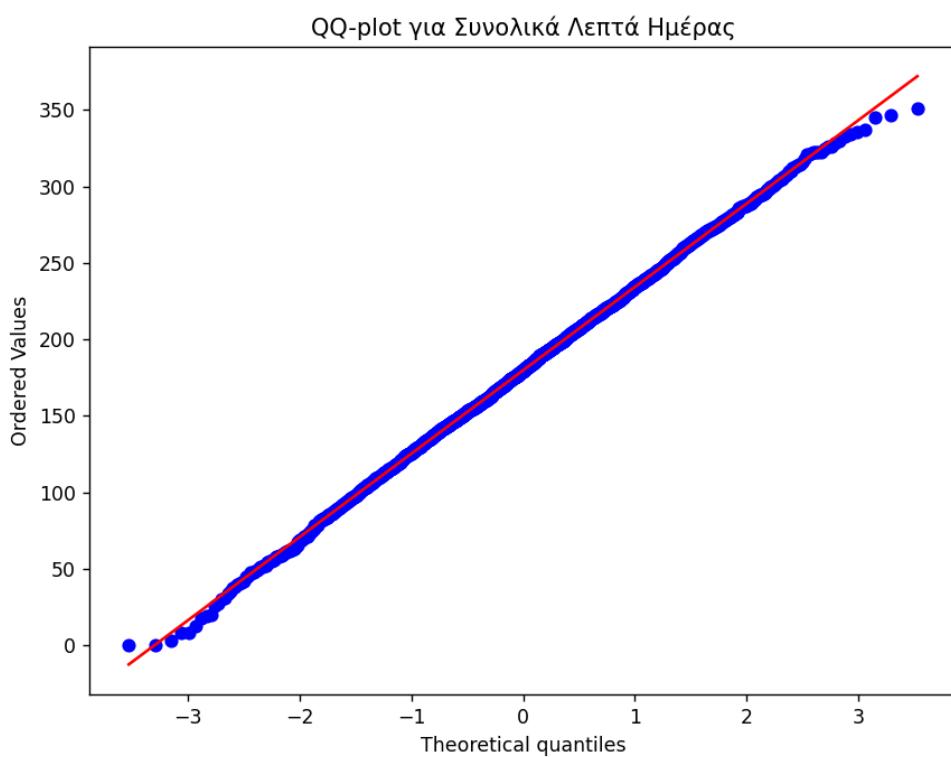
Εικόνα 32 - QQ Plot (Κωδικός Περιοχής)

Figure 1



Εικόνα 33 - Κατανομή Συνολικών Λεπτών Ημέρας

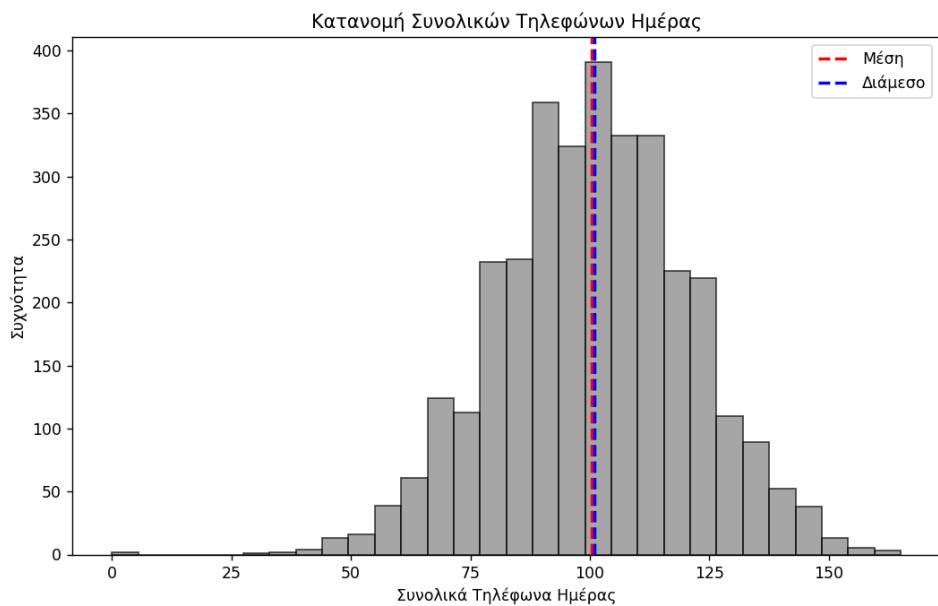
Figure 1



Εικόνα 34 - QQ Plot (Συνολικά Λεπτά Ημέρας)

Figure 1

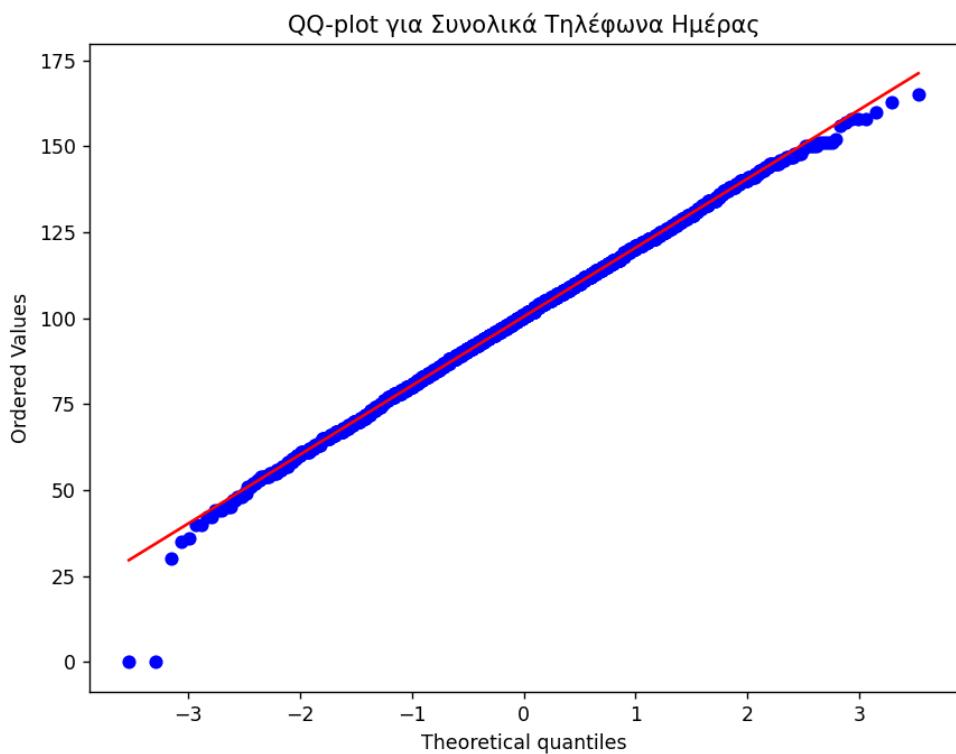
- □ ×



Εικόνα 35 - Κατανομή Συνολικών Τηλεφώνων Ημέρας

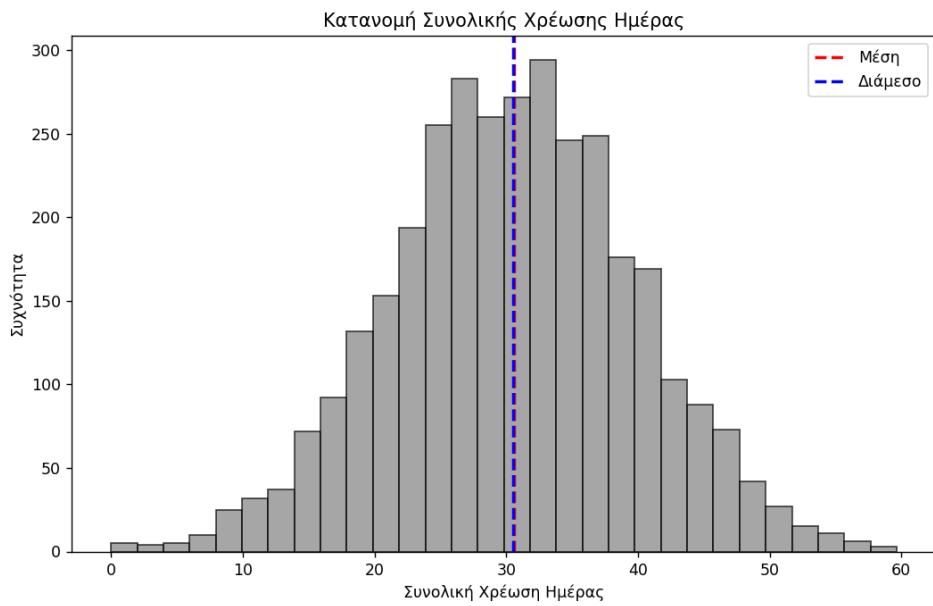
Figure 1

- □ ×



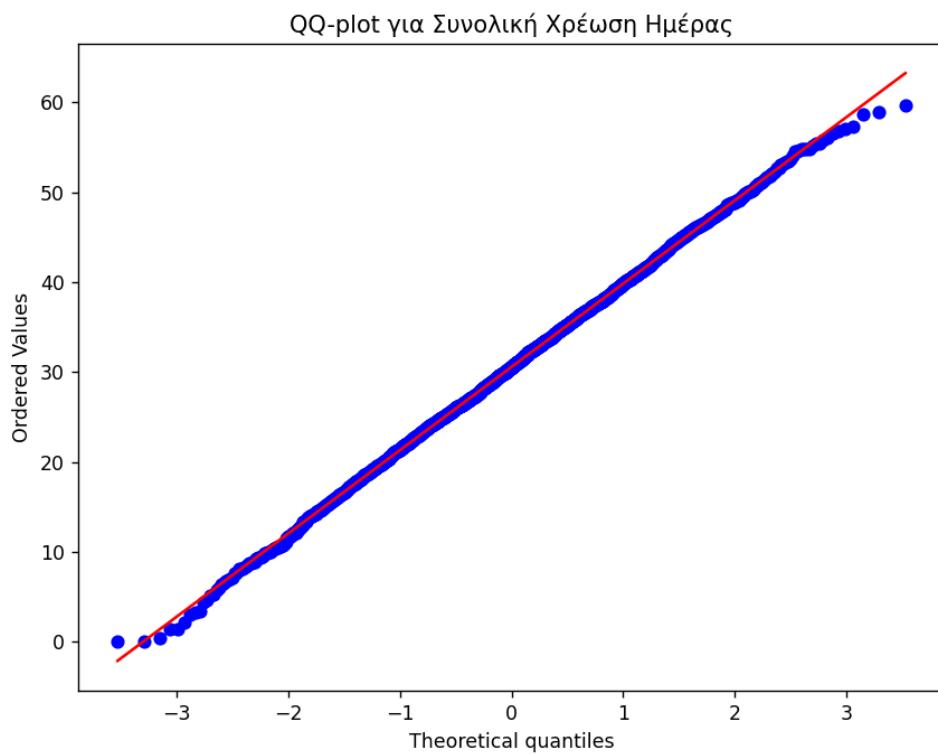
Εικόνα 36 - QQ Plot (Συνολικά Τηλέφωνα Ημέρας)

Figure 1



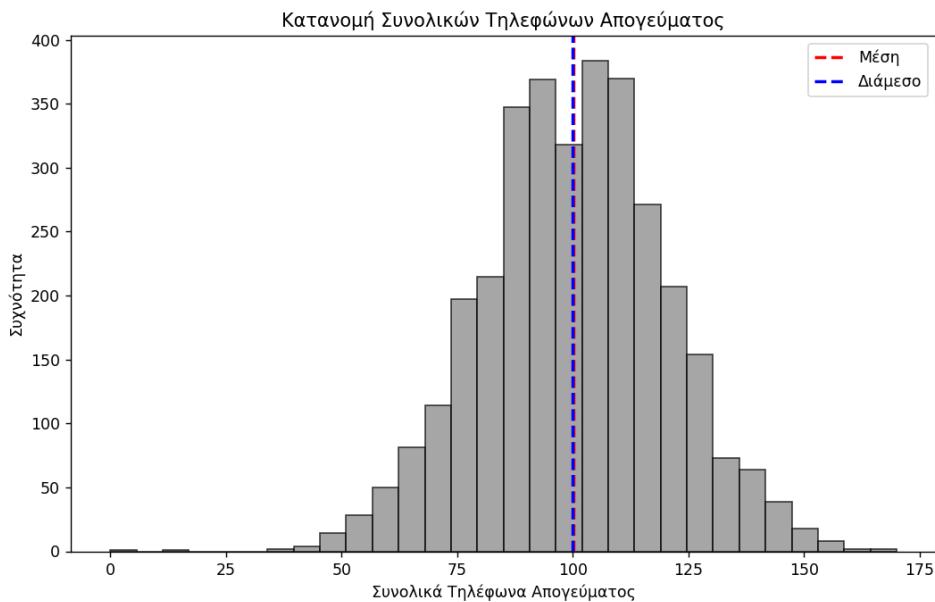
Εικόνα 37 - Κατανομή Συνολικής Χρέωσης Ημέρας

Figure 1



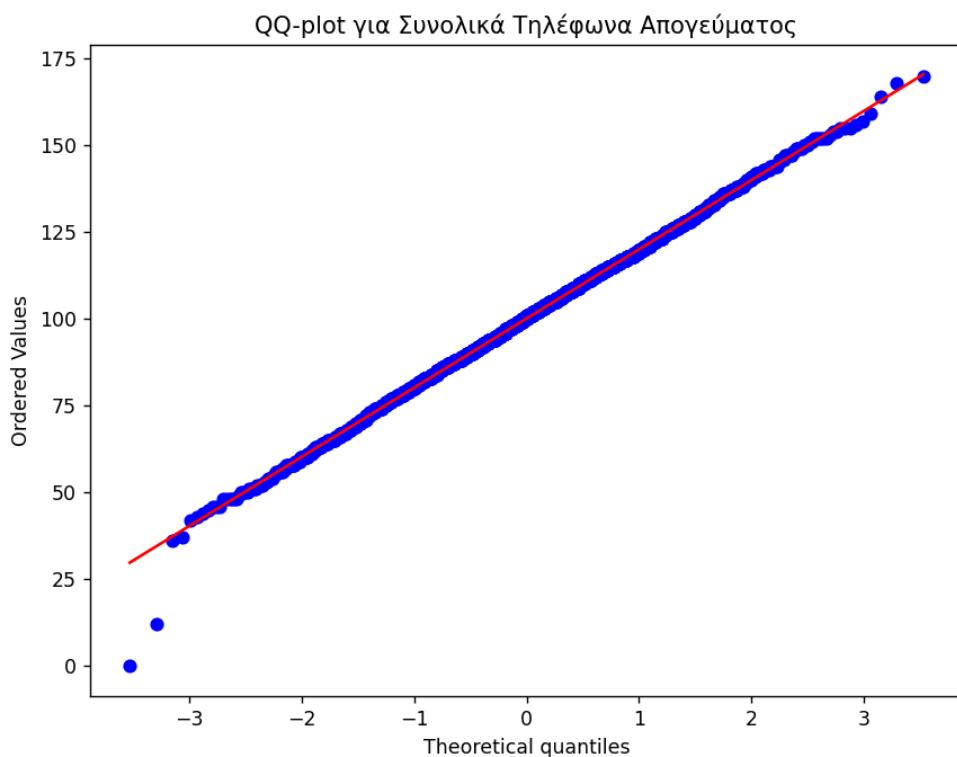
Εικόνα 38 - QQ Plot (Συνολική Χρέωση Ημέρας)

Figure 1



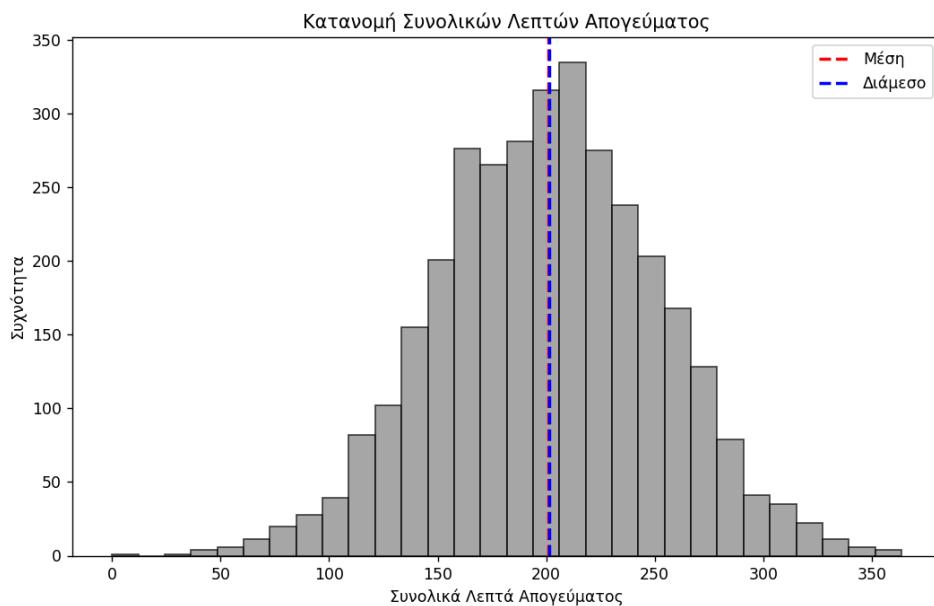
Εικόνα 39 - Κατανομή Συνολικών Απογευματινών Τηλεφώνων

Figure 1



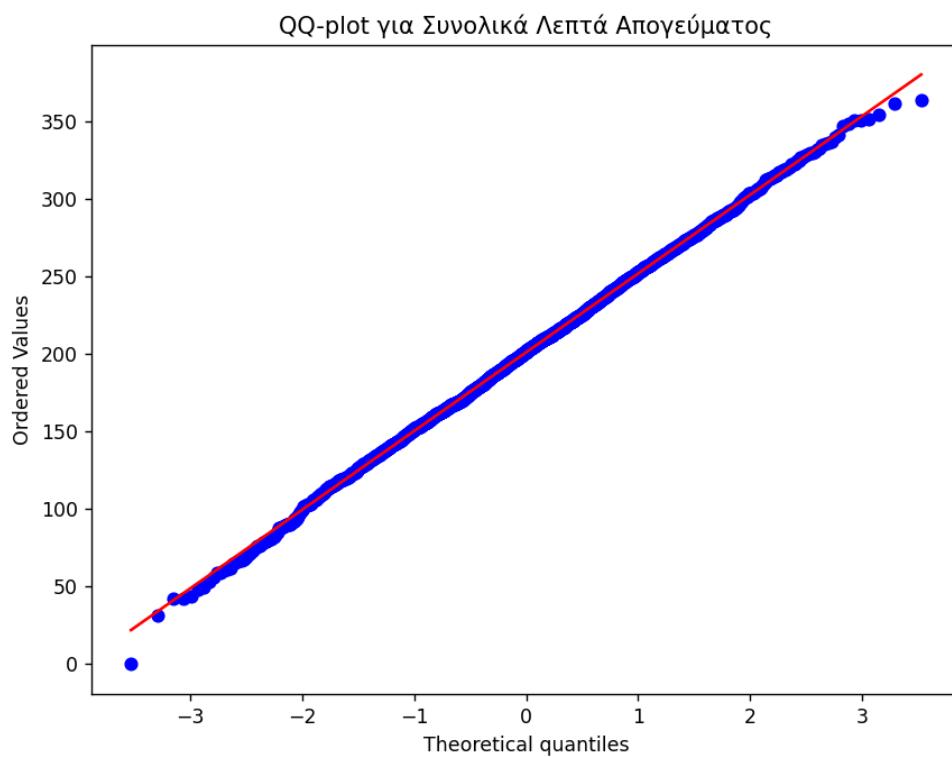
Εικόνα 40 - QQ Plot(Συνολικά Απογευματινά Τηλέφωνα)

Figure 1



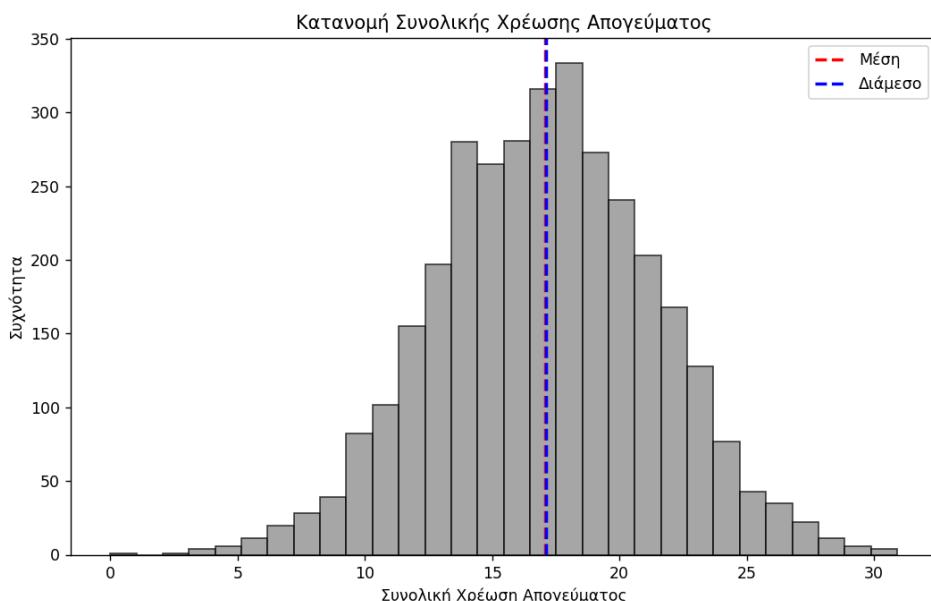
Εικόνα 41 - Κατανομή Συνολικών Λεπτών Απογεύματος

Figure 1



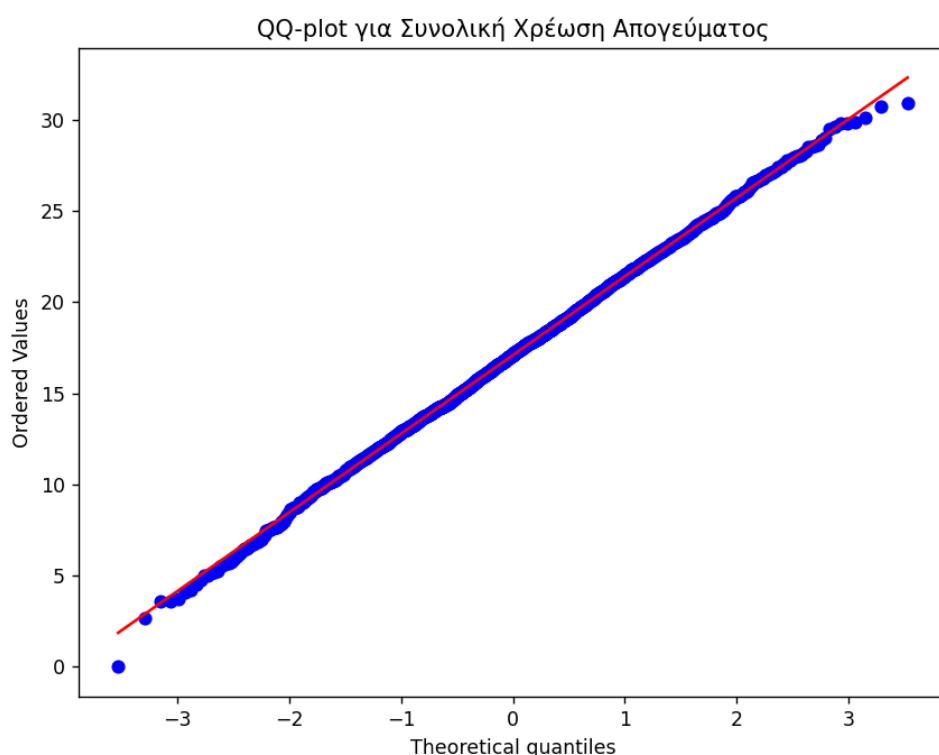
Εικόνα 42 - QQ Plot (Συνολικά Λεπτά Απογεύματος)

Figure 1



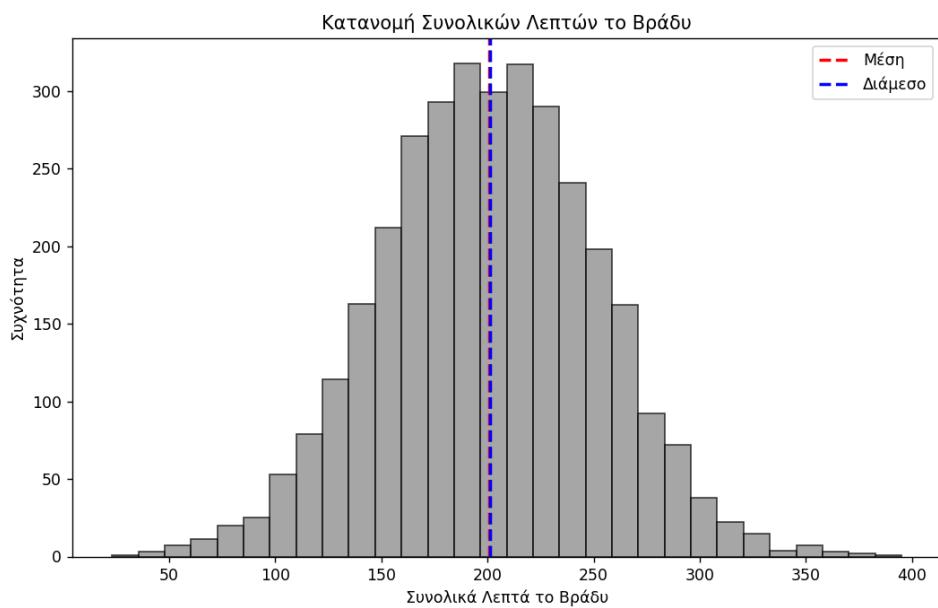
Εικόνα 43 - Κατανομή Συνολικής Χρέωσης Απογεύματος

Figure 1



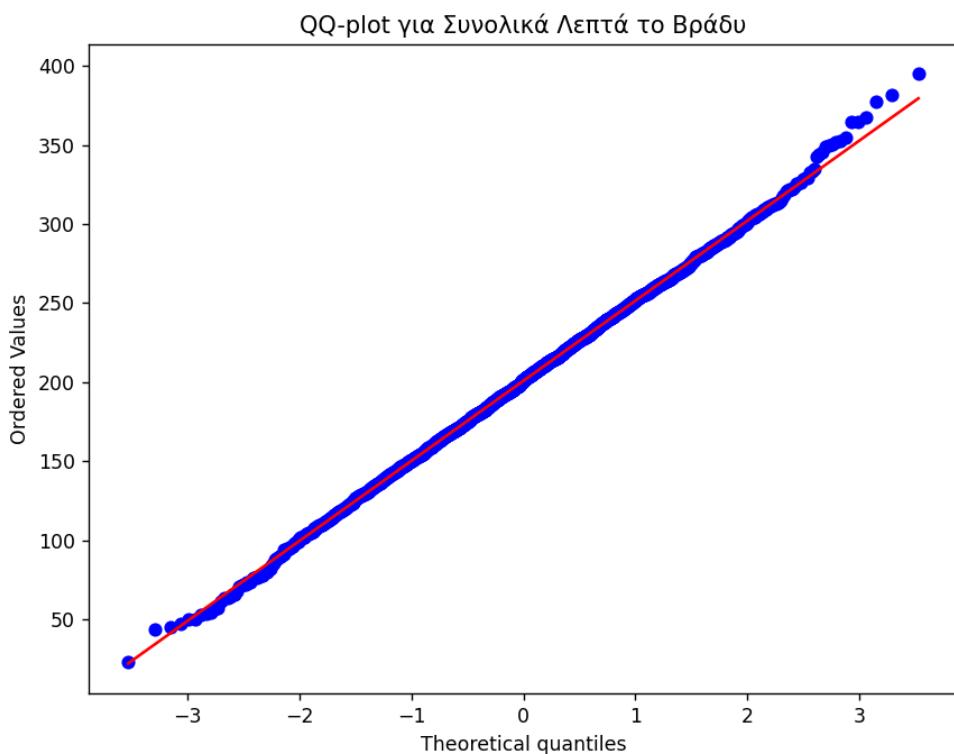
Εικόνα 44 - QQ Plot (Συνολική Χρέωση Απογεύματος)

Figure 1



Εικόνα 45 - Κατανομή Συνολικών Βραδινών Λεπτών

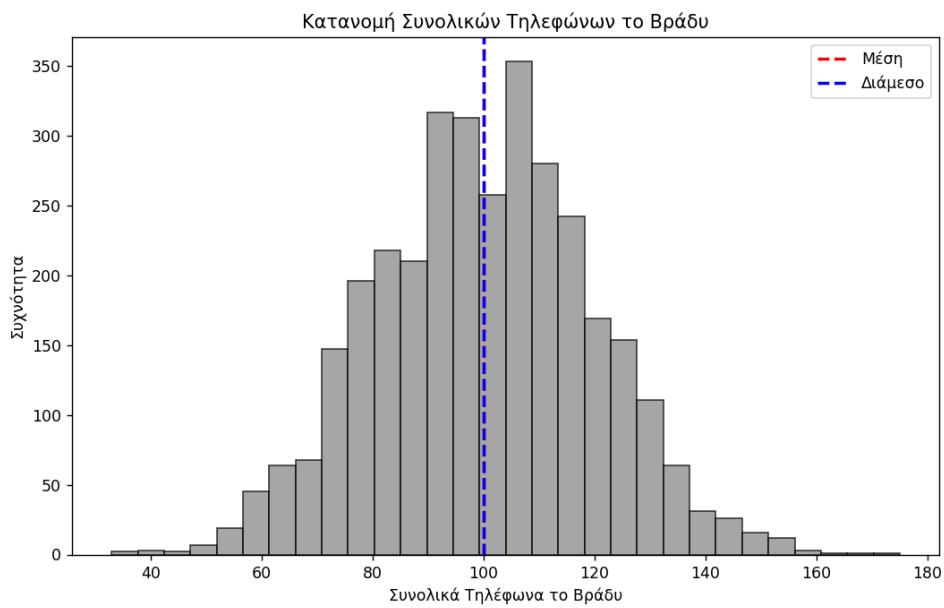
Figure 1



Εικόνα 46 - QQ Plot (Συνολικά Βραδινά Λεπτά)

Figure 1

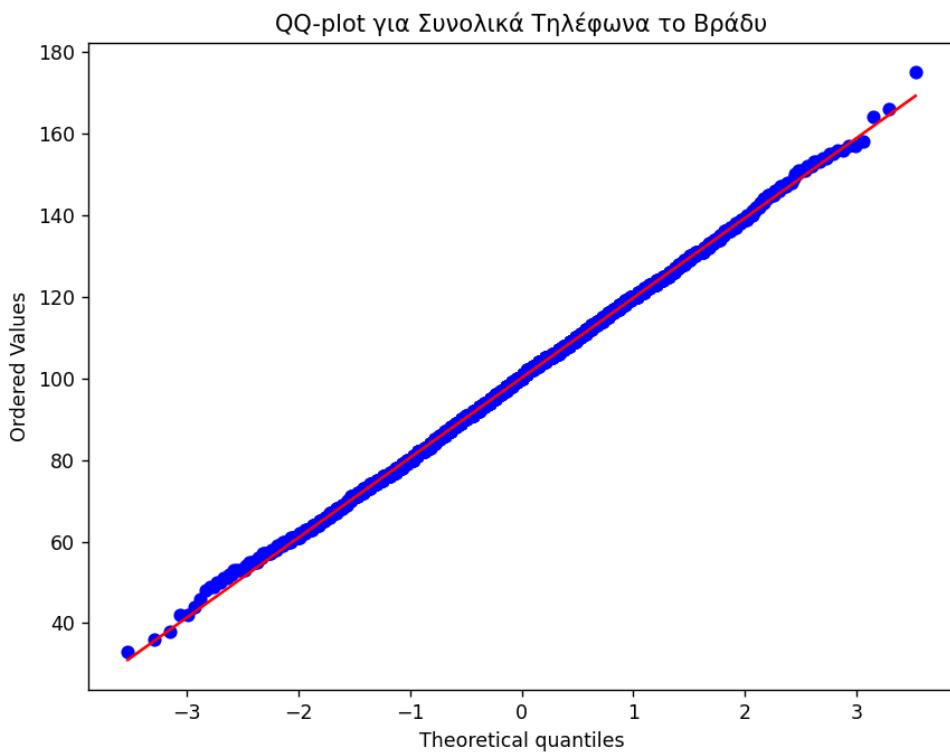
- □ ×



Εικόνα 47 - Κατανομή Συνολικών Βραδινών Λεπτών

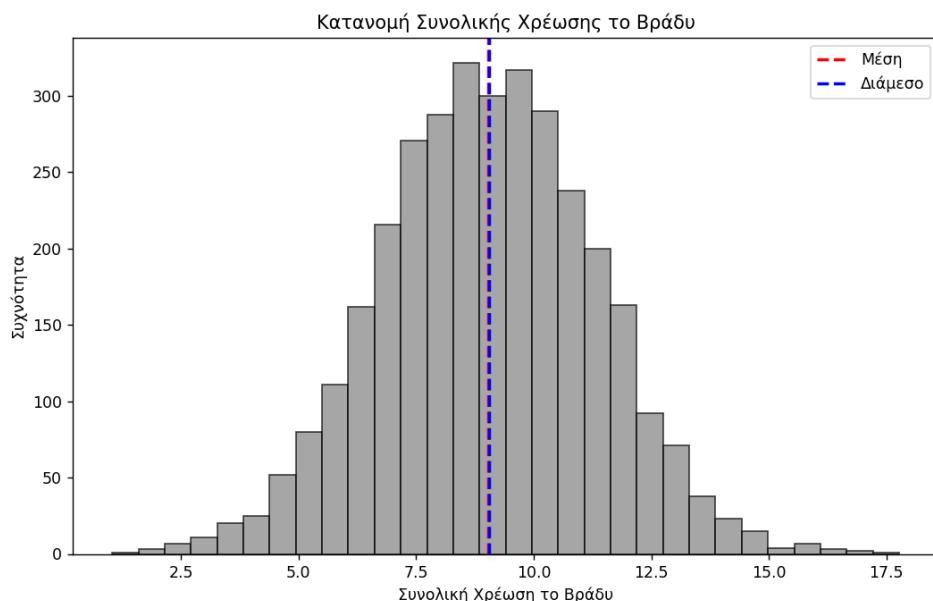
Figure 1

- □ ×



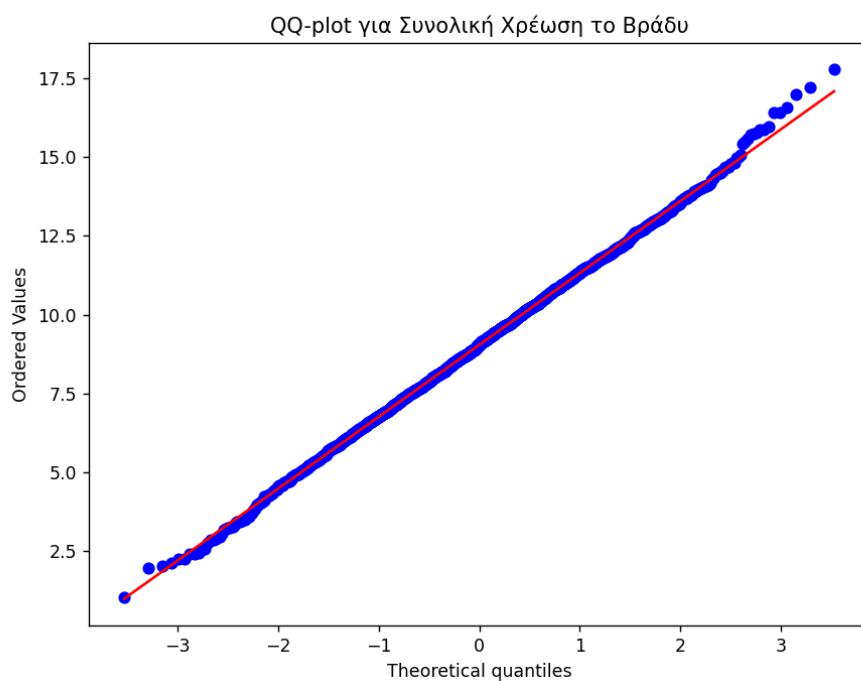
Εικόνα 48 - QQ Plot (Συνολικά Βραδινά Τηλέφωνα)

Figure 1



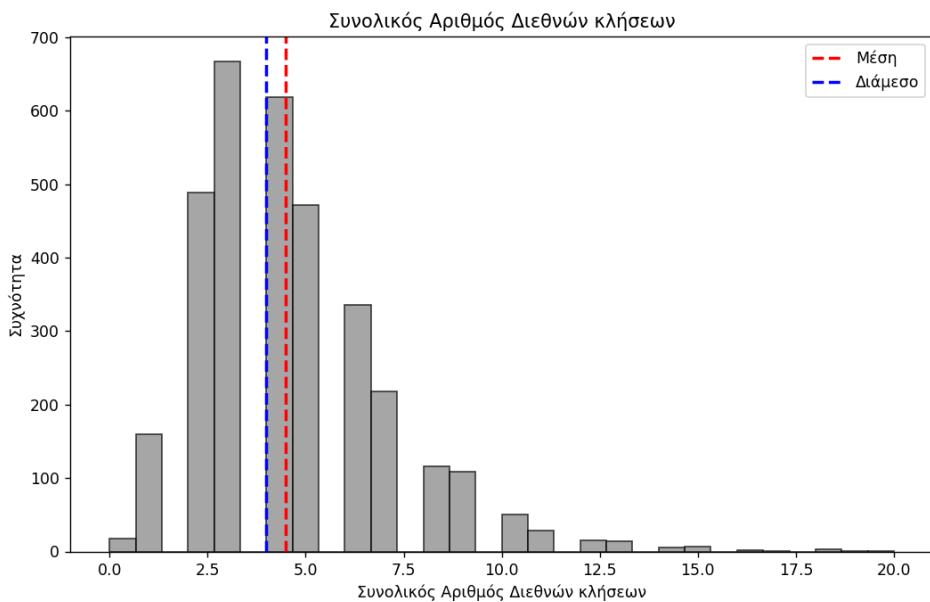
Εικόνα 49 - Κατανομή Συνολικής Βραδινής Χρέωσης

Figure 1



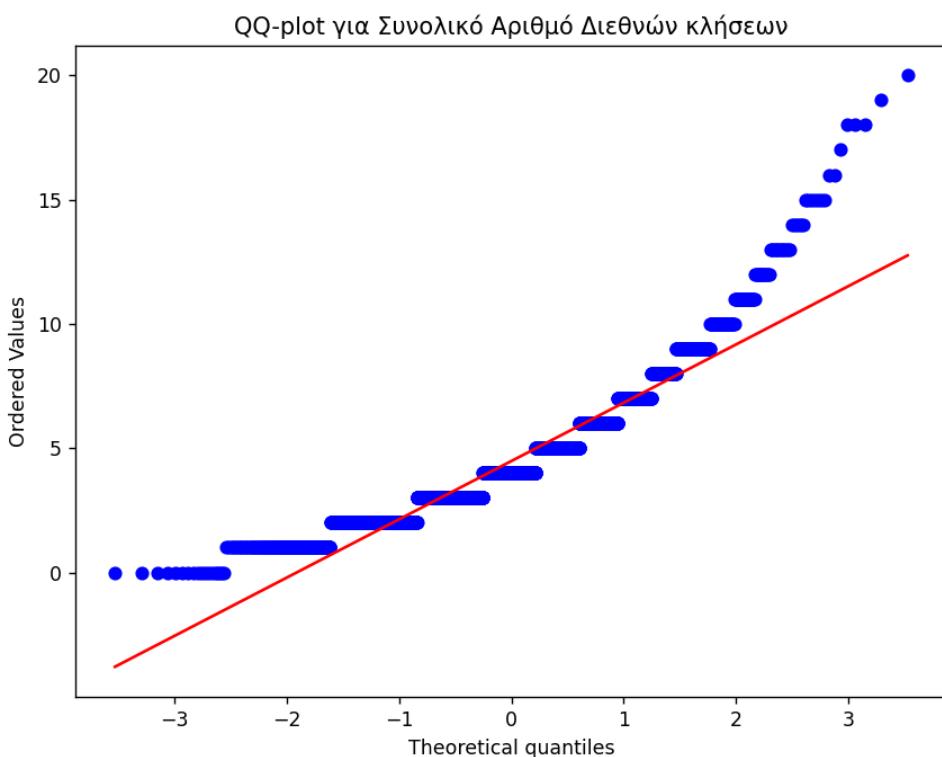
Εικόνα 50 - QQ Plot (Συνολική Βραδινή Χρέωση)

Figure 1



Εικόνα 51 - Κατανομή Συνολικών Διεθνών Κλήσεων

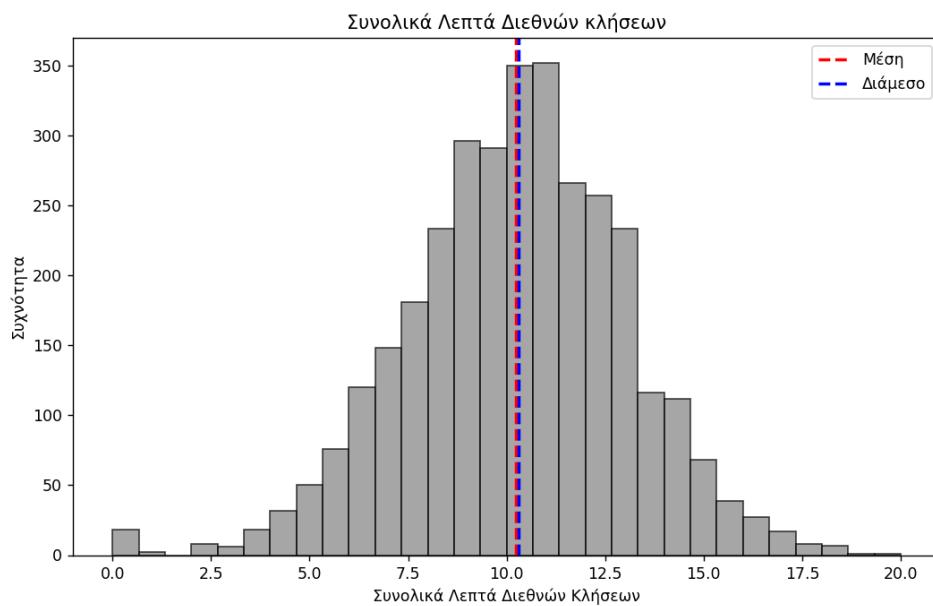
Figure 1



Εικόνα 52 - QQ Plot (Συνολικός Αριθμός Διεθνών κλήσεων)

Figure 1

- □ ×



Εικόνα 53 - Κατανομή Συνολικών Λεπτών Διεθνών Κλήσεων

Figure 1

- □ ×

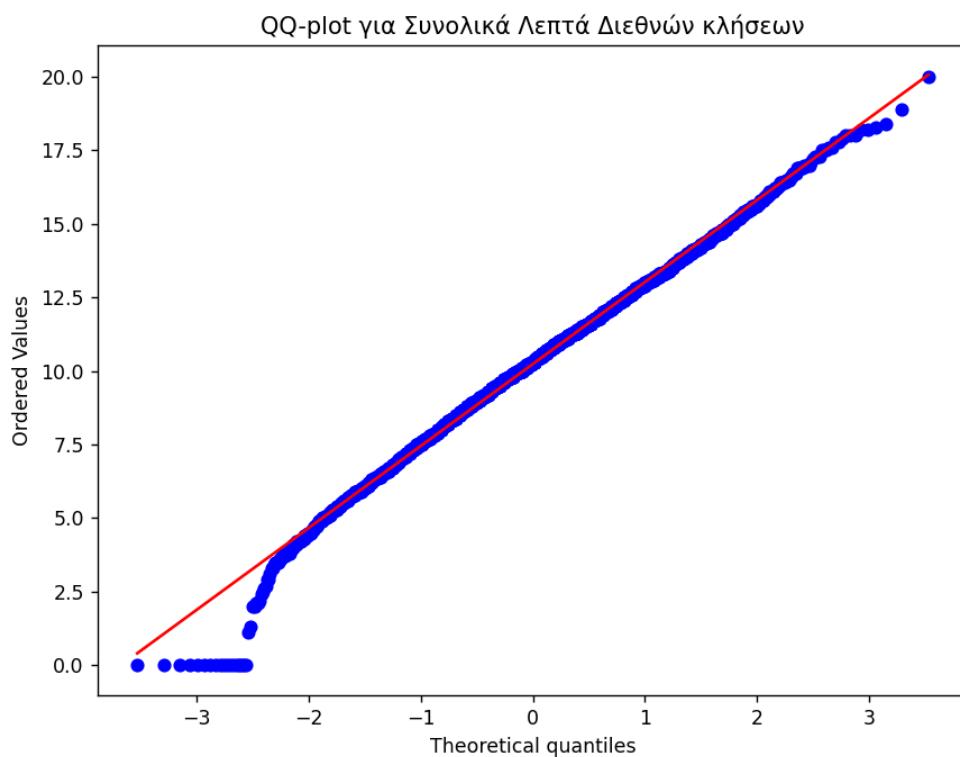
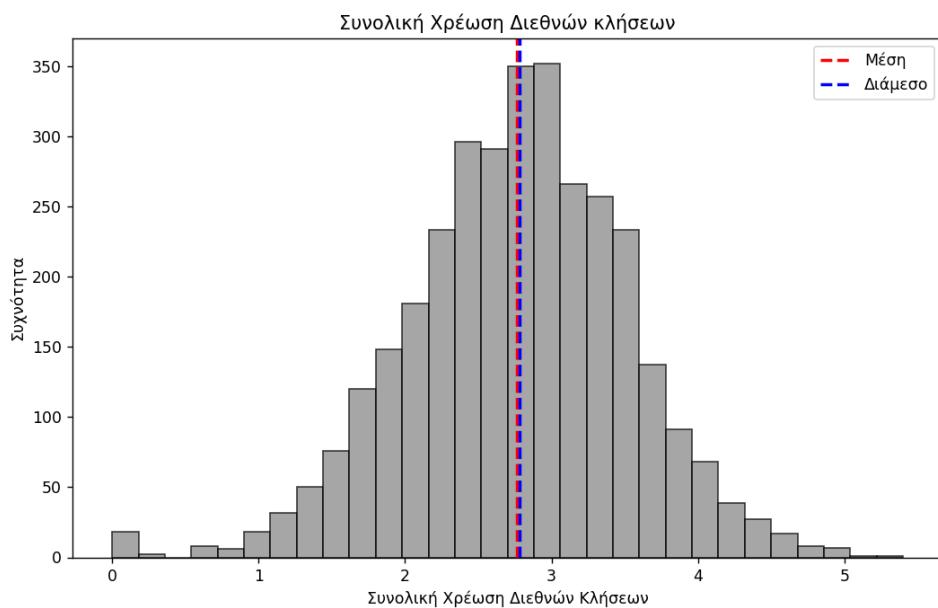
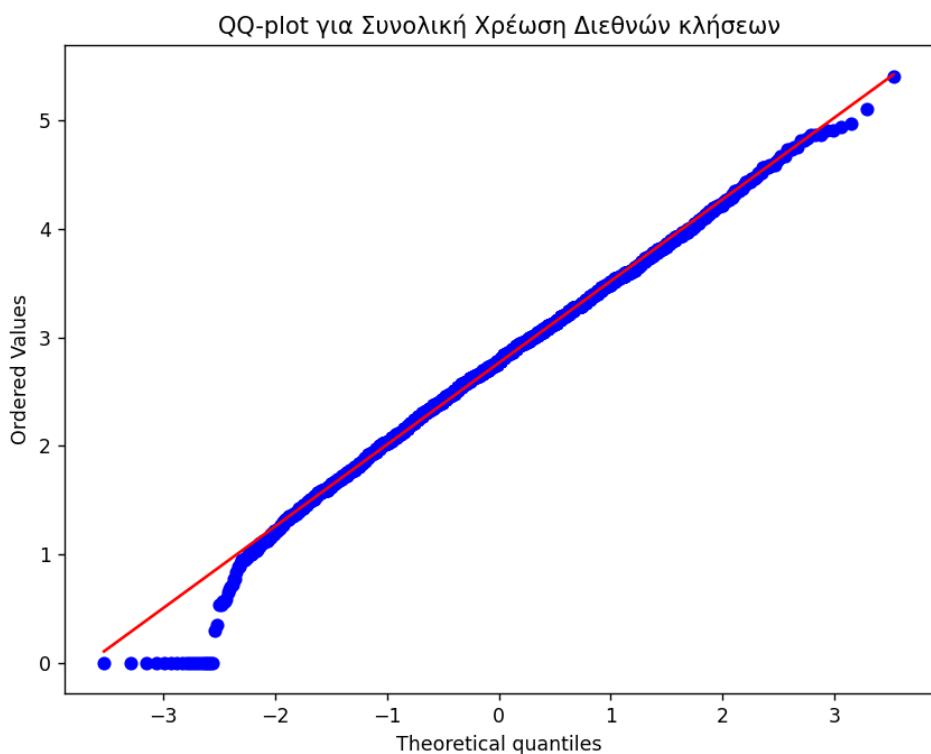


Figure 1



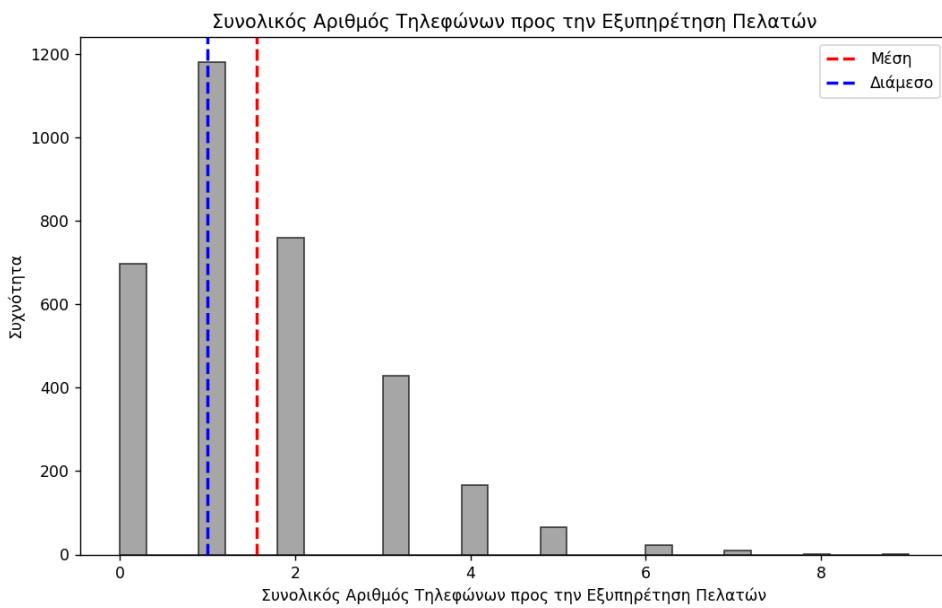
Εικόνα 54 - Κατανομή Συνολικής Χρέωσης Διεθνών Κλήσεων

Figure 1



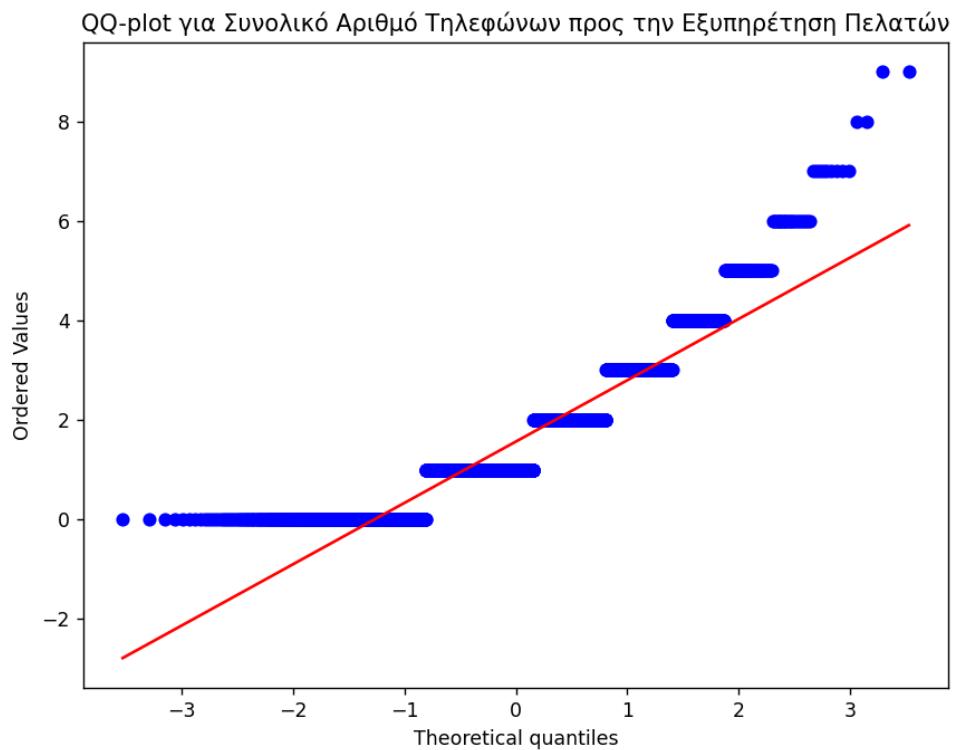
Εικόνα 55 - QQ Plot (Συνολική Χρέωση Διεθνών κλήσεων)

Figure 1



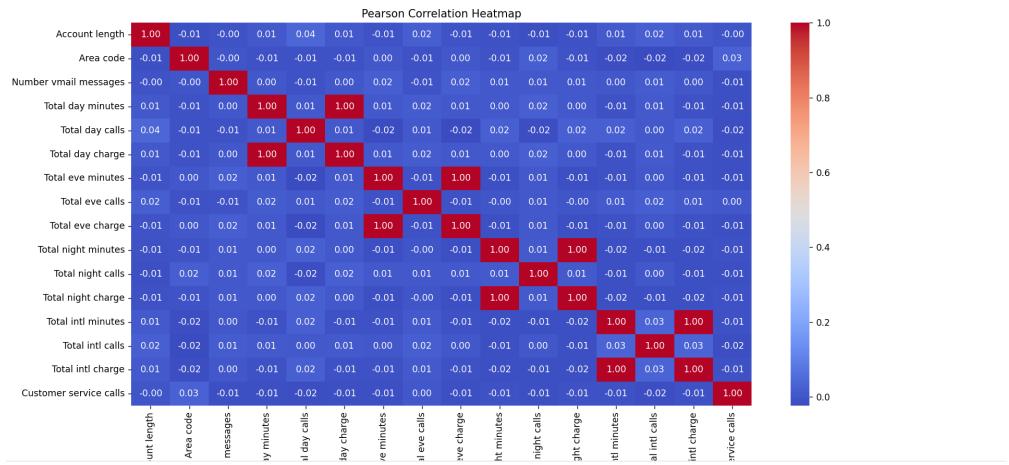
Εικόνα 56 - Κατανομή Συνόλου Τηλεφώνων προς την Εξυπηρέτηση Πελατών

Figure 1

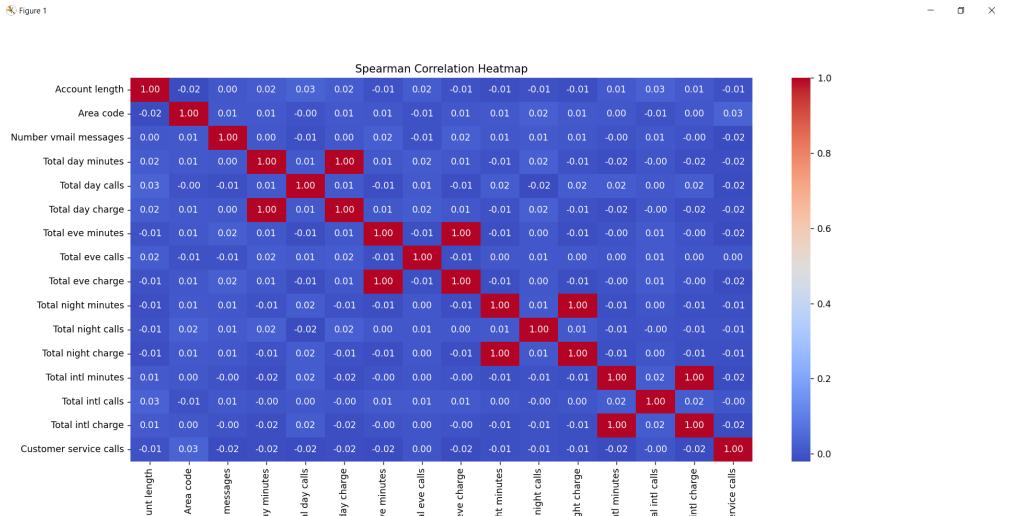


Εικόνα 57 - QQ Plot (Συνολικός Αριθμός Τηλεφώνων προς την Εξυπηρέτηση Πελατών)

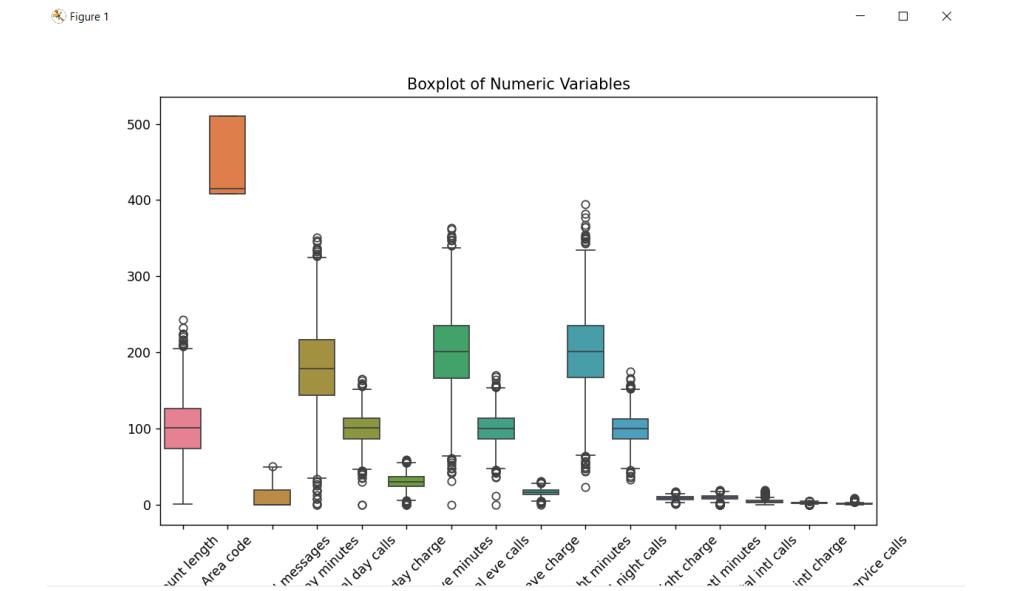
Figure 1



Εικόνα 58 - Pearson Correlation Heatmap



Εικόνα 59 - Spearman Correlation Heatmap



Εικόνα 60 - Boxplot Αριθμητικών Δεδομένων

2. Κονσόλα:

```
Index(['State', 'Account length', 'Area code', 'International plan',
       'Voice mail plan', 'Number vmail messages', 'Total day minutes',
       'Total day calls', 'Total day charge', 'Total eve minutes',
       'Total eve calls', 'Total eve charge', 'Total night minutes',
       'Total night calls', 'Total night charge', 'Total intl minutes',
       'Total intl calls', 'Total intl charge', 'Customer service calls',
       'Churn'],
      dtype='object')
   Account length    Area code ...  Total intl charge  Customer service calls
count      3333.000000  3333.000000 ...      3333.000000      3333.000000
mean      101.064806  437.182418 ...      2.764581      1.562856
std       39.822106  42.371290 ...      0.753773      1.315491
min       1.000000  408.000000 ...      0.000000      0.000000
25%      74.000000  408.000000 ...      2.300000      1.000000
50%     101.000000  415.000000 ...      2.780000      1.000000
75%     127.000000  510.000000 ...      3.270000      2.000000
max     243.000000  510.000000 ...      5.400000      9.000000
```

Εικόνα 61 - Περιεχόμενα Dataset (1)

```
[8 rows x 16 columns]
   State                      object
   Account length            int64
   Area code                  int64
   International plan        object
   Voice mail plan           object
   Number vmail messages     int64
   Total day minutes         float64
   Total day calls            int64
   Total day charge           float64
   Total eve minutes          float64
   Total eve calls             int64
   Total eve charge           float64
   Total night minutes         float64
   Total night calls            int64
   Total night charge           float64
   Total intl minutes          float64
   Total intl calls             int64
   Total intl charge           float64
   Customer service calls     int64
   Churn                      bool
dtype: object
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
```

Εικόνα 62 - Περιεχόμενα Dataset (2)

```

Data columns (total 20 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   State            3333 non-null   object  
 1   Account length   3333 non-null   int64  
 2   Area code         3333 non-null   int64  
 3   International plan 3333 non-null   object  
 4   Voice mail plan  3333 non-null   object  
 5   Number vmail messages 3333 non-null   int64  
 6   Total day minutes 3333 non-null   float64 
 7   Total day calls   3333 non-null   int64  
 8   Total day charge  3333 non-null   float64 
 9   Total eve minutes 3333 non-null   float64 
 10  Total eve calls   3333 non-null   int64  
 11  Total eve charge  3333 non-null   float64 
 12  Total night minutes 3333 non-null   float64 
 13  Total night calls  3333 non-null   int64  
 14  Total night charge 3333 non-null   float64 
 15  Total intl minutes 3333 non-null   float64 
 16  Total intl calls   3333 non-null   int64  
 17  Total intl charge  3333 non-null   float64 
 18  Customer service calls 3333 non-null   int64  
 19  Churn             3333 non-null   bool    
dtypes: bool(1), float64(8), int64(8), object(3) 
memory usage: 498.1+ KB
None

```

Εικόνα 63 - Περιεχόμενα Dataset (3)

State	0
Account length	0
Area code	0
International plan	0
Voice mail plan	0
Number vmail messages	0
Total day minutes	0
Total day calls	0
Total day charge	0
Total eve minutes	0
Total eve calls	0
Total eve charge	0
Total night minutes	0
Total night calls	0
Total night charge	0
Total intl minutes	0
Total intl calls	0
Total intl charge	0
Customer service calls	0
Churn	0
dtype:	int64

Εικόνα 64 - Σύνολο null τιμών στο dataset

	count	mean	std	...	50%	75%	max
Account length	3333.0	101.06	39.82	...	101.00	127.00	243.00
Area code	3333.0	437.18	42.37	...	415.00	510.00	510.00
Number vmail messages	3333.0	8.10	13.69	...	0.00	20.00	51.00
Total day minutes	3333.0	179.78	54.47	...	179.40	216.40	350.80
Total day calls	3333.0	100.44	20.07	...	101.00	114.00	165.00
Total day charge	3333.0	30.56	9.26	...	30.50	36.79	59.64
Total eve minutes	3333.0	200.98	50.71	...	201.40	235.30	363.70
Total eve calls	3333.0	100.11	19.92	...	100.00	114.00	170.00
Total eve charge	3333.0	17.08	4.31	...	17.12	20.00	30.91
Total night minutes	3333.0	200.87	50.57	...	201.20	235.30	395.00
Total night calls	3333.0	100.11	19.57	...	100.00	113.00	175.00
Total night charge	3333.0	9.04	2.28	...	9.05	10.59	17.77
Total intl minutes	3333.0	10.24	2.79	...	10.30	12.10	20.00
Total intl calls	3333.0	4.48	2.46	...	4.00	6.00	20.00
Total intl charge	3333.0	2.76	0.75	...	2.78	3.27	5.40
Customer service calls	3333.0	1.56	1.32	...	1.00	2.00	9.00

Εικόνα 65 - Σύνολο numeric columns

Rounded data:						
	Account length	Area code	...	Total intl charge	Customer service calls	
count	3333.000000	3333.000000	...	3333.000000	3333.000000	
mean	101.064806	437.182418	...	2.764581	1.562856	
std	39.822106	42.371290	...	0.753773	1.315491	
min	1.000000	408.000000	...	0.000000	0.000000	
25%	74.000000	408.000000	...	2.300000	1.000000	
50%	101.000000	415.000000	...	2.780000	1.000000	
75%	127.000000	510.000000	...	3.270000	2.000000	
max	243.000000	510.000000	...	5.400000	9.000000	

[8 rows x 16 columns]

Εικόνα 66 - Στρογγυλοποίηση αριθμητικών δεδομένων

Correlation Pearson:						
	Account length	...	Customer service calls			
Account length	1.000000	...	-0.003796			
Area code	-0.012463	...	0.027572			
Number vmail messages	-0.004628	...	-0.013263			
Total day minutes	0.006216	...	-0.013423			
Total day calls	0.038470	...	-0.018942			
Total day charge	0.006214	...	-0.013427			
Total eve minutes	-0.006757	...	-0.012985			
Total eve calls	0.019260	...	0.002423			
Total eve charge	-0.006745	...	-0.012987			
Total night minutes	-0.008955	...	-0.009288			
Total night calls	-0.013176	...	-0.012802			
Total night charge	-0.008960	...	-0.009277			
Total intl minutes	0.009514	...	-0.009640			
Total intl calls	0.020661	...	-0.017561			
Total intl charge	0.009546	...	-0.009675			
Customer service calls	-0.003796	...	1.000000			

[16 rows x 16 columns]

Εικόνα 67 - Correlation Pearson

	Account length	...	Customer service calls
Account length	1.000000	...	-0.005942
Area code	-0.017439	...	0.031850
Number vmail messages	0.003077	...	-0.019639
Total day minutes	0.017884	...	-0.015032
Total day calls	0.032690	...	-0.020957
Total day charge	0.017884	...	-0.015032
Total eve minutes	-0.007954	...	-0.017805
Total eve calls	0.018378	...	0.002697
Total eve charge	-0.007950	...	-0.017800
Total night minutes	-0.013643	...	-0.012713
Total night calls	-0.007669	...	-0.008087
Total night charge	-0.013654	...	-0.012709
Total intl minutes	0.014761	...	-0.017374
Total intl calls	0.027453	...	-0.000598
Total intl charge	0.014761	...	-0.017374
Customer service calls	-0.005942	...	1.000000

[16 rows x 16 columns]

Εικόνα 68 - Correlation Spearman

```
Correlation between Total day minutes and Total day charge: 1.0000
SignificanceResult(statistic=1.0, pvalue=0.0)
```

Εικόνα 69 - Υπολογισμός Συσχέτισης

Percentage of outliers: 10.0%

Εικόνα 70 - Ποσοστό εκτός των κύριων ακτινών

```
Outliers for variable Account length
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Area code
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Number vmail messages
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total day minutes
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total day calls
2 outliers
0.1% outliers
[74. 1.]
```

Εικόνα 71 - Δεδομένα Ακραίων Τιμών (1)

```
Outliers for variable Total day charge
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total eve minutes
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total eve calls
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total eve charge
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total night minutes
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total night calls
2 outliers
0.1% outliers
[74. 1.]
```

Εικόνα 72 - Δεδομένα Ακραίων Τιμών (2)

```
Outliers for variable Total night charge
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total intl minutes
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total intl calls
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Total intl charge
2 outliers
0.1% outliers
[74. 1.]
-----
Outliers for variable Customer service calls
2 outliers
0.1% outliers
[74. 1.]
```

Εικόνα 73 - Δεδομένα Ακραίων Τιμών (3)

Chi-Square Value: 83.0438
P-value: 0.0023

Chi-Square Value: 222.5658
P-value: 0.0000

Chi-Square Value: 34.1317
P-value: 0.0000

Εικόνα 74 - Chi-Square Tests

```

Skewness:
Account length           0.096563
Area code                 1.126316
Number vmail messages    1.264254
Total day minutes        -0.029064
Total day calls          -0.111736
Total day charge         -0.029070
Total eve minutes        -0.023867
Total eve calls          -0.055538
Total eve charge         -0.023847
Total night minutes      0.008917
Total night calls         0.032485
Total night charge        0.008882
Total intl minutes       -0.245026
Total intl calls          1.320883
Total intl charge         -0.245176
Customer service calls   1.090868
dtype: float64

```

Eikόνα 75 – Skewness

```

Kurtosis:
Account length           -0.109474
Area code                 -0.706374
Number vmail messages    -0.052852
Total day minutes        -0.021710
Total day calls          0.241017
Total day charge         -0.021582
Total eve minutes        0.023792
Total eve calls          0.204048
Total eve charge         0.023650
Total night minutes      0.083888
Total night calls         -0.073711
Total night charge        0.083735
Total intl minutes       0.606472
Total intl calls          3.077165
Total intl charge         0.606897
Customer service calls   1.726518
dtype: float64

```

Eikόνα 76 – Kurtosis

```

Kolmogorov-Smirnov test for Account length: KS Statistic = 0.9959498319413672, p-value = 0.0
Kolmogorov-Smirnov test for Area code: KS Statistic = 1.0, p-value = 0.0
Kolmogorov-Smirnov test for Number vmail messages: KS Statistic = 0.5, p-value = 0.0
Kolmogorov-Smirnov test for Total day minutes: KS Statistic = 0.999099909990996, p-value = 0.0
Kolmogorov-Smirnov test for Total day calls: KS Statistic = 0.9993999399939995, p-value = 0.0
Kolmogorov-Smirnov test for Total day charge: KS Statistic = 0.9972943174290896, p-value = 0.0
Kolmogorov-Smirnov test for Total eve minutes: KS Statistic = 0.9996999699969997, p-value = 0.0
Kolmogorov-Smirnov test for Total eve calls: KS Statistic = 0.9996999699969997, p-value = 0.0
Kolmogorov-Smirnov test for Total eve charge: KS Statistic = 0.9992346010132793, p-value = 0.0
Kolmogorov-Smirnov test for Total night minutes: KS Statistic = 1.0, p-value = 0.0
Kolmogorov-Smirnov test for Total night calls: KS Statistic = 1.0, p-value = 0.0
Kolmogorov-Smirnov test for Total night charge: KS Statistic = 0.9942136785090832, p-value = 0.0
Kolmogorov-Smirnov test for Total intl minutes: KS Statistic = 0.9899156157616066, p-value = 0.0
Kolmogorov-Smirnov test for Total intl calls: KS Statistic = 0.9238445275177674, p-value = 0.0
Kolmogorov-Smirnov test for Total intl charge: KS Statistic = 0.8895872238393948, p-value = 0.0
Kolmogorov-Smirnov test for Customer service calls: KS Statistic = 0.6322238339773338, p-value = 0.0

```

Eukóva 77 - Kolmogorov - Smirnov Results

Anderson-Darling test results:	
Account length	0.426158
Area code	687.939568
Number vmail messages	634.895083
Total day minutes	0.210056
Total day calls	0.607521
Total day charge	0.210281
Total eve minutes	0.291521
Total eve calls	0.638074
Total eve charge	0.293427
Total night minutes	0.208816
Total night calls	0.681179
Total night charge	0.212092
Total intl minutes	2.425119
Total intl calls	79.715713
Total intl charge	2.453685
Customer service calls	132.759008
dtype:	float64

Eukóva 78 - Anderson-Darling Results

Shapiro-Wilk test results:					
	Account length	Area code	...	Total intl charge	Customer service calls
0	0.998279	0.59027	...	9.937097e-01	8.767108e-01
1	0.001158	0.00000	...	7.640023e-11	1.401298e-45

[2 rows x 16 columns]

Eukóva 79 - Shapiro-Wilk Results

```

One-sample t-test results:
[TtestResult(statistic=146.51880985204102, pvalue=0.0, df=3332), TtestResult(statistic=599.6740124725692, pvalue=0.0, df=3332), TtestResult(statistic=34.158437587695325,
pvalue=1.5880197922094533e-219, df=3332), TtestResult(statistic=198.55073626795428, pvalue=0.0, df=3332), TtestResult(statistic=288.9202409507132, pvalue=0.0, df=3332),
TtestResult(statistic=198.5545519445562, pvalue=0.0, df=3332), TtestResult(statistic=228.794045108337, pvalue=0.0, df=3332), TtestResult(statistic=290.1130432272334,
pvalue=0.0, df=3332), TtestResult(statistic=228.97352684266, pvalue=0.0, df=3332), TtestResult(statistic=229.308374645952, pvalue=0.0, df=3332), TtestResult(statistic=295.
342088247208837, pvalue=0.0, df=3332), TtestResult(statistic=229.30884877127825, pvalue=0.0, df=3332), TtestResult(statistic=211.695868657311315, pvalue=0.0, df=3332),
TtestResult(statistic=105.073385656362219, pvalue=0.0, df=3332), TtestResult(statistic=211.774184545938708, pvalue=0.0, df=3332), TtestResult(statistic=68.588105099505059,
pvalue=0.0, df=3332)]

```

Eukóva 80 - One-sample t-test Results

```

Logistic Regression

Coefficients:
[[ 2.42028164e-04 -3.11239628e-03  1.20122522e-02  1.17562240e-02
   1.37724920e-03  2.15986639e-03  7.01536425e-03 -1.48258804e-03
  -6.11216095e-03  2.76374641e-03 -2.40365446e-03  3.28094991e-03
   5.83219997e-02 -9.42060729e-02  2.94979347e-02  5.07381286e-01
  -2.40221391e+00 -2.63274800e-01 -7.00713974e-01 -1.96477474e+00
  -6.71789119e-01 -4.86768571e-01  6.12926186e-02 -6.50799296e-01
   9.27021630e-01 -2.04655521e-01  1.76351726e-01 -1.51450147e-01
  -8.43479675e-02 -1.52046894e-01 -7.99486074e-02 -8.26916226e-01
  -4.35999125e-01  2.22897507e-02 -8.92786508e-01 -2.99903471e-01
   2.19480020e-01 -7.75997159e-02 -2.90408674e-01  2.89643231e-01
  2.68248944e-01  4.89587752e-01  4.52795507e-01  3.07025132e-01
  -2.23221390e-01  4.83022389e-01  8.47058580e-01 -1.72761623e-01
  -5.68502679e-01 -4.29379746e-01  2.59835530e-01  6.19197261e-01
  -3.78714738e-01  3.75731510e-01  2.60709677e-01 -1.28019941e-01
  9.41017847e-02 -5.33652808e-02  2.42467841e-01 -7.33519555e-01
  8.98345518e-01 -6.06414063e-03 -5.37671358e-01  6.21775065e-01
  2.19044120e-01 -1.05431102e+00 -6.07767735e-01  5.13388861e-01
  -5.25808981e-01 -1.86391931e-01 -4.02983200e-01]]

Predicted Probabilities:
[0.13111491 0.03760928 0.1076792 0.53599293 0.57751115 0.2000364
 0.19045884 0.09322771 0.16172247 0.40701894]

Confusion Matrix:
[[2776 74]
 [ 364 119]]
```

Eικόνα 81 - Logistic Regression (1)

```

Classification Report:
      precision    recall    f1-score   support
  False        0.88      0.97      0.93     2850
   True        0.62      0.25      0.35      483

  accuracy                           0.87     3333
  macro avg       0.75      0.61      0.64     3333
weighted avg       0.85      0.87      0.84     3333

Training Error Rate: 0.1314131413141314
Confusion Matrix (Test Data):
[[905 38]
 [123 34]]
```

Eικόνα 82 - Logistic Regression (2)

Decision Trees Classification

```
|--- Total day charge <= 44.96
|   |--- Customer service calls <= 3.50
|   |   |--- International plan_No <= 0.50
|   |   |   |--- Total intl calls <= 2.50
|   |   |   |   |--- class: True
|   |   |   |   |--- Total intl calls >  2.50
|   |   |   |   |--- Total intl minutes <= 13.10
|   |   |   |   |   |--- Total day minutes <= 251.90
|   |   |   |   |   |   |--- Total eve minutes <= 324.85
|   |   |   |   |   |   |   |--- Total night charge <= 3.77
|   |   |   |   |   |   |   |   |--- class: True
|   |   |   |   |   |   |   |   |--- Total night charge >  3.77
|   |   |   |   |   |   |   |   |--- class: False
|   |   |   |   |   |   |   |   |--- Total eve minutes >  324.85
|   |   |   |   |   |   |   |   |--- Total eve calls <= 113.00
|   |   |   |   |   |   |   |   |   |--- class: True
|   |   |   |   |   |   |   |   |   |--- Total eve calls >  113.00
|   |   |   |   |   |   |   |   |   |--- class: False
|   |   |   |   |   |   |   |   |--- Total day minutes >  251.90
|   |   |   |   |   |   |   |   |--- Total eve minutes <= 245.75
|   |   |   |   |   |   |   |   |   |--- Total day minutes <= 253.20
|   |   |   |   |   |   |   |   |   |   |--- class: True
|   |   |   |   |   |   |   |   |   |--- Total day minutes >  253.20
|   |   |   |   |   |   |   |   |   |   |--- class: False
|   |   |   |   |   |   |   |   |--- Total eve minutes >  245.75
|   |   |   |   |   |   |   |   |   |--- class: True
```

Etkόva 83 - Decision Trees (1)

```

|   |   |   |--- Total intl minutes > 13.10
|   |   |   |--- class: True
|   |--- International plan_No > 0.50
|   |--- Total day minutes <= 223.25
|   |   |--- Total eve charge <= 27.99
|   |   |--- Account length <= 224.50
|   |   |   |--- State_TX <= 0.50
|   |   |   |   |--- Number vmail messages <= 43.50
|   |   |   |   |--- Total night calls <= 104.50
|   |   |   |   |   |--- Total night minutes <= 353.55
|   |   |   |   |   |--- Total day charge <= 36.05
|   |   |   |   |   |--- truncated branch of depth 12
|   |   |   |   |   |--- Total day charge > 36.05
|   |   |   |   |   |--- truncated branch of depth 7
|   |   |   |   |   |--- Total night minutes > 353.55
|   |   |   |   |   |--- Total eve minutes <= 265.40
|   |   |   |   |   |--- class: False
|   |   |   |   |   |--- Total eve minutes > 265.40
|   |   |   |   |   |--- class: True
|   |   |   |   |--- Total night calls > 104.50
|   |   |   |   |--- Account length <= 152.50
|   |   |   |   |   |--- Total eve minutes <= 168.70
|   |   |   |   |   |--- class: False
|   |   |   |   |   |--- Total eve minutes > 168.70
|   |   |   |   |   |--- truncated branch of depth 15
|   |   |   |   |   |--- Account length > 152.50
|   |   |   |   |--- Total night calls <= 143.00

```

Εικόνα 84 - Decision Trees (2)

```

| | | | | | | | | | | |--- truncated branch of depth 9
| | | | | | | | | | |--- Total night calls > 143.00
| | | | | | | | | | |--- class: True
| | | | | | |--- Number vmail messages > 43.50
| | | | | | |--- Account length <= 140.50
| | | | | | |--- State_UT <= 0.50
| | | | | | |--- class: False
| | | | | | |--- State_UT > 0.50
| | | | | | |--- class: True
| | | | | | |--- Account length > 140.50
| | | | | | |--- class: True
| | | | | |--- State_TX > 0.50
| | | | | |--- Total day minutes <= 216.00
| | | | | |--- Total intl calls <= 1.50
| | | | | | |--- Customer service calls <= 1.50
| | | | | | |--- class: False
| | | | | | |--- Customer service calls > 1.50
| | | | | | |--- class: True
| | | | | | |--- Total intl calls > 1.50
| | | | | | |--- Total night charge <= 11.27
| | | | | | |--- class: False
| | | | | | |--- Total night charge > 11.27
| | | | | | |--- Total night minutes <= 256.00
| | | | | | |--- class: True
| | | | | | |--- Total night minutes > 256.00
| | | | | | |--- class: False
| | | | | | |--- Total day minutes > 216.00
| | | | | | |--- class: True
| | | | | |--- Account length > 224.50

```

Εικόνα 85 - Decision Trees (3)

```

| | | | | | | --- State_MI <= 0.50
| | | | | | | --- class: False
| | | | | | | --- State_MI > 0.50
| | | | | | | --- class: True
| | | | | --- Total eve charge > 27.99
| | | | | | --- Total night charge <= 10.39
| | | | | | | --- class: False
| | | | | | | --- Total night charge > 10.39
| | | | | | | --- Total day minutes <= 131.75
| | | | | | | --- class: False
| | | | | | | --- Total day minutes > 131.75
| | | | | | | --- class: True
| | | | | --- Total day minutes > 223.25
| | | | | | --- Total eve minutes <= 259.80
| | | | | | | --- Total eve minutes <= 242.35
| | | | | | | --- Total day minutes <= 223.40
| | | | | | | | --- class: True
| | | | | | | --- Total day minutes > 223.40
| | | | | | | | --- Total intl charge <= 4.65
| | | | | | | | | --- Account length <= 204.00
| | | | | | | | | | --- Total night minutes <= 233.35
| | | | | | | | | | | --- Total day calls <= 147.50
| | | | | | | | | | | | --- truncated branch of depth 8
| | | | | | | | | | | | | --- Total day calls > 147.50
| | | | | | | | | | | | | --- class: True
| | | | | | | | | | | | | --- Total night minutes > 233.35
| | | | | | | | | | | | | --- Total day charge <= 43.50
| | | | | | | | | | | | | | --- truncated branch of depth 5
| | | | | | | | | | | | | | --- Total day charge > 43.50

```

Etkόva 86 - Decision Trees (4)

```

| | | | | | | | | | | | |--- truncated branch of depth 2
| | | | | | | | | | |--- Account length > 204.00
| | | | | | | | | | |--- class: True
| | | | | | | | | | |--- Total intl charge > 4.65
| | | | | | | | | | |--- class: True
| | | | | | |--- Total eve minutes > 242.35
| | | | | | |--- Total night minutes <= 201.05
| | | | | | |--- Total day charge <= 44.37
| | | | | | |--- class: False
| | | | | | |--- Total day charge > 44.37
| | | | | | |--- class: True
| | | | | | |--- Total night minutes > 201.05
| | | | | | |--- Number vmail messages <= 11.50
| | | | | | |--- Account length <= 52.00
| | | | | | |--- Total night calls <= 91.50
| | | | | | |--- class: True
| | | | | | |--- Total night calls > 91.50
| | | | | | |--- class: False
| | | | | | |--- Account length > 52.00
| | | | | | |--- class: True
| | | | | | |--- Number vmail messages > 11.50
| | | | | | |--- class: False
| | | | | |--- Total eve minutes > 259.80
| | | | | |--- Voice mail plan_Yes <= 0.50
| | | | | |--- Total night charge <= 6.62
| | | | | |--- Total day charge <= 41.91
| | | | | |--- class: False
| | | | | |--- Total day charge > 41.91

```

Etkόva 87 - Decision Trees (5)

```

|   |   |   |   |   |   |   |--- class: True
|   |   |   |   |   |--- Total night charge >  6.62
|   |   |   |   |   |--- Total intl minutes <= 7.90
|   |   |   |   |   |--- Total intl minutes <= 6.45
|   |   |   |   |   |   |--- Total intl calls <= 1.00
|   |   |   |   |   |   |--- class: False
|   |   |   |   |   |   |--- Total intl calls >  1.00
|   |   |   |   |   |   |--- class: True
|   |   |   |   |   |   |--- Total intl minutes >  6.45
|   |   |   |   |   |   |--- class: False
|   |   |   |   |   |--- Total intl minutes >  7.90
|   |   |   |   |   |--- class: True
|   |   |   |   |--- Voice mail plan_Yes >  0.50
|   |   |   |   |--- class: False
|--- Customer service calls >  3.50
|   |--- Total day minutes <= 160.20
|   |--- Total eve charge <= 19.83
|   |--- State_IL <= 0.50
|   |--- State_CO <= 0.50
|   |--- class: True
|   |--- State_CO >  0.50
|   |--- Total intl calls <= 6.50
|   |--- class: True
|   |--- Total intl calls >  6.50
|   |--- class: False
|   |--- State_IL >  0.50
|   |--- class: False

```

Etkόνα 88 - Decision Trees (6)

```

|   |   |   |--- Total eve charge >  19.83
|   |   |   |--- Total day minutes <= 120.50
|   |   |   |   |--- class: True
|   |   |   |--- Total day minutes >  120.50
|   |   |   |--- Total night minutes <= 207.35
|   |   |   |   |--- class: False
|   |   |   |--- Total night minutes >  207.35
|   |   |   |--- Total day charge <= 23.36
|   |   |   |   |--- class: True
|   |   |   |--- Total day charge >  23.36
|   |   |   |--- Total eve calls <= 95.50
|   |   |   |   |--- class: True
|   |   |   |--- Total eve calls >  95.50
|   |   |   |   |--- class: False
|   |--- Total day minutes >  160.20
|   |--- Total eve charge <= 12.05
|   |--- Total eve calls <= 125.00
|   |--- Total intl calls <= 7.00
|   |--- State_WA <= 0.50
|   |   |--- class: True
|   |--- State_WA >  0.50
|   |   |--- class: False
|   |--- Total intl calls >  7.00
|   |   |--- class: False
|   |--- Total eve calls >  125.00
|   |   |--- class: False
|   |--- Total eve charge >  12.05
|   |--- Total day charge <= 29.88
|   |--- Total eve minutes <= 212.15

```

Etkόνα 89 - Decision Trees (7)

```

|   |   |   |   |   |--- State_AL <= 0.50
|   |   |   |   |   |--- Total intl minutes <= 16.80
|   |   |   |   |   |--- class: True
|   |   |   |   |   |--- Total intl minutes > 16.80
|   |   |   |   |   |--- class: False
|   |   |   |   |--- State_AL > 0.50
|   |   |   |   |   |--- class: False
|   |   |   |   |--- Total eve minutes > 212.15
|   |   |   |   |--- class: False
|   |   |   |   |--- Total day charge > 29.88
|   |   |   |   |--- International plan_Yes <= 0.50
|   |   |   |   |   |--- Total intl charge <= 1.28
|   |   |   |   |   |--- class: True
|   |   |   |   |--- Total intl charge > 1.28
|   |   |   |   |--- State_MS <= 0.50
|   |   |   |   |   |--- State_OR <= 0.50
|   |   |   |   |   |   |--- State_WY <= 0.50
|   |   |   |   |   |   |--- State_TX <= 0.50
|   |   |   |   |   |   |--- class: False
|   |   |   |   |   |   |--- State_TX > 0.50
|   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |--- State_WY > 0.50
|   |   |   |   |   |   |--- Total night charge <= 10.20
|   |   |   |   |   |   |--- class: False
|   |   |   |   |   |   |--- Total night charge > 10.20
|   |   |   |   |   |   |--- class: True
|   |   |   |   |   |--- State_OR > 0.50
|   |   |   |   |   |--- Total night calls <= 99.50
|   |   |   |   |   |--- class: False

```

Etkόνα 90 - Decision Trees (8)

```

|   |   |   |   |   |   |   |   |--- Total night calls > 99.50
|   |   |   |   |   |   |   |--- class: True
|   |   |   |   |   |--- State_MS > 0.50
|   |   |   |   |   |--- class: True
|   |   |   |   |--- International plan_Yes > 0.50
|   |   |   |   |--- Total eve minutes <= 243.05
|   |   |   |   |--- Total day minutes <= 187.50
|   |   |   |   |--- class: True
|   |   |   |   |--- Total day minutes > 187.50
|   |   |   |   |--- class: False
|   |   |   |   |--- Total eve minutes > 243.05
|   |   |   |   |--- class: True
|--- Total day charge > 44.96
|   |--- Voice mail plan_Yes <= 0.50
|   |   |--- Total eve charge <= 15.96
|   |   |--- Total day minutes <= 277.70
|   |   |   |--- Total night minutes <= 257.60
|   |   |   |--- State_NE <= 0.50
|   |   |   |--- class: False
|   |   |   |--- State_NE > 0.50
|   |   |   |--- class: True
|   |   |   |--- Total night minutes > 257.60
|   |   |   |--- Total eve charge <= 11.49
|   |   |   |--- class: False
|   |   |   |--- Total eve charge > 11.49
|   |   |   |--- class: True
|   |   |--- Total day minutes > 277.70
|   |   |--- Total eve minutes <= 144.35
|   |   |   |--- Total night calls <= 81.00
|   |   |   |--- class: True

```

Eukóva 91 - Decision Trees (9)

```

|   |   |   |   |   |--- Total night calls >  81.00
|   |   |   |   |   |--- class: False
|   |   |   |   |--- Total eve minutes >  144.35
|   |   |   |   |   |--- Total night charge <= 6.84
|   |   |   |   |   |--- Account length <= 123.00
|   |   |   |   |   |   |--- class: False
|   |   |   |   |   |--- Account length >  123.00
|   |   |   |   |   |   |--- class: True
|   |   |   |   |   |--- Total night charge >  6.84
|   |   |   |   |   |--- class: True
|   |   |--- Total eve charge >  15.96
|   |   |   |--- Total night minutes <= 127.00
|   |   |   |   |--- Total day minutes <= 277.15
|   |   |   |   |   |--- class: False
|   |   |   |   |--- Total day minutes >  277.15
|   |   |   |   |   |--- class: True
|   |   |   |--- Total night minutes >  127.00
|   |   |   |   |--- State_ID <= 0.50
|   |   |   |   |   |--- class: True
|   |   |   |--- State_ID >  0.50
|   |   |   |   |   |--- Total intl minutes <= 8.50
|   |   |   |   |   |--- class: False
|   |   |   |   |   |--- Total intl minutes >  8.50
|   |   |   |   |   |--- class: True
|--- Voice mail plan_Yes >  0.50
|   |--- International plan_Yes <= 0.50
|   |   |--- Total day charge <= 54.23
|   |   |   |--- class: False
|   |   |--- Total day charge >  54.23

```

```

|   |   |   |   |--- class: True
|   |   |--- International plan_Yes >  0.50
|   |   |--- Total day minutes <= 276.30
|   |   |   |--- class: True
|   |   |   |--- Total day minutes >  276.30
|   |   |   |--- State_ME <= 0.50
|   |   |   |   |--- class: False
|   |   |   |--- State_ME >  0.50
|   |   |   |   |--- class: True

[[1896    0]
 [  0  326]]
Accuracy: 1.0

```

Eukόva 93 - Decision Trees (11)

```

Alpha: 0.0, Accuracy: 0.90999099909991
Alpha: 0.0008308523160008305, Accuracy: 0.90999099909991
Alpha: 0.0008365707636194488, Accuracy: 0.9212421242124212
Alpha: 0.0008685079034219214, Accuracy: 0.923042304230423
Alpha: 0.0008809391577455626, Accuracy: 0.9234923492349235
Alpha: 0.0010492255761585217, Accuracy: 0.9302430243024302
Alpha: 0.0011583863507696791, Accuracy: 0.9324932493249325
Alpha: 0.0012001200120012, Accuracy: 0.9302430243024302
Alpha: 0.0012001200120012, Accuracy: 0.9302430243024302
Alpha: 0.0015430114440015434, Accuracy: 0.9324932493249325
Alpha: 0.0016413994596501849, Accuracy: 0.9342934293429342
Alpha: 0.0017281728172817278, Accuracy: 0.9342934293429342
Alpha: 0.0017472335468840998, Accuracy: 0.9342934293429342
Alpha: 0.0018001800180018005, Accuracy: 0.9342934293429342
Alpha: 0.001953314499671774, Accuracy: 0.9365436543654365
Alpha: 0.0024002400240024, Accuracy: 0.9374437443744374
Alpha: 0.0028002800280027998, Accuracy: 0.9383438343834384
Alpha: 0.003203261502620851, Accuracy: 0.9383438343834384
Alpha: 0.003477681101443477, Accuracy: 0.9315931593159316
Alpha: 0.0037388354220037395, Accuracy: 0.9293429342934293
Alpha: 0.0038403840384038388, Accuracy: 0.9293429342934293
Alpha: 0.0051026670693613526, Accuracy: 0.9261926192619262
Alpha: 0.005625562556255626, Accuracy: 0.9239423942394239
Alpha: 0.0077129141485577125, Accuracy: 0.8987398739873987
Alpha: 0.016269404718249607, Accuracy: 0.8793879387938794
Alpha: 0.0171358908042703, Accuracy: 0.8541854185418541
Alpha: 0.017286026758359663, Accuracy: 0.8541854185418541
Alpha: 0.026553222521829276, Accuracy: 0.8532853285328533

```

Eukόva 94 - Decision Trees (12)

K-Nearest Neighbors Classifier

Predicted Labels for the Test Set:

Eικόνα 95 - K-Nearest Neighbors

Eικόνα 96 - K-Nearest Neighbors

```
False False True False False False False False False False False False  
False False False False False False False False False False False False True  
True False False False False False False]  
Accuracy: 84.26%  
Precision: 54.55%
```

Eikόνα 97 - K-Nearest Neighbors

Random Forest Classifier

Predicted Labels for the Test Set:

Εικόνα 99 - Random Forest

```
True False False False False False False]  
Accuracy: 93.25%  
Precision: 97.10%
```

Εικόνα 100 - Random Forest

Συμπεράσματα

Μετά από όλες αυτές τις μεθόδους classification (Logistic Regression, Decision ee, KNN, Random Forest) τις οποίες εφαρμόσαμε στην προηγούμενη ενότητα για τα δεδομένα που μας δημοσίευσε η τηλεφωνική εταιρεία, καταλήξαμε στο συμπέρασμα πώς οι δύο καλύτερες μέθοδοι classification είναι το Random Forest και το Decision Tree. Οι τηλεφωνικές εταιρείες είναι πολύ σημαντικό να μπορούν να έχουν στην διάθεση τους ένα προβλεπτικό μοντέλο με το οποίο να μπορούν να ξεχωρίζουν ποιοι πελάτες θα είναι σε θέση να φύγουν από τα δικά τους πλάνα. Έτσι με τα δεδομένα της εταιρείας για την πραγματοποίηση της παραπάνω ανάλυσης που είχαμε στην διάθεση μας είδαμε ότι την σωστή προβλεπτική ικανότητα την είχαμε από την εφαρμογή της μεθόδου Random Forest. Η ακρίβεια του μοντέλου φαίνεται να είναι 95.05%, ενώ η ακρίβεια της κλάσης "True" (πελάτες που φεύγουν) είναι 100%. Αυτό σημαίνει ότι το μοντέλο έχει πολύ υψηλή ακρίβεια στο να αναγνωρίζει τους πελάτες που πρόκειται να φύγουν. Παρατηρήσαμε ότι η πρόβλεψη αποτελεί μια επαρκές πρόβλεψη και την μεγαλύτερη από όλες τις άλλες μεθόδους, το οποίο είναι και πολύ σημαντικό να μπορεί το μοντέλο μας να εφαρμόζεται σε μεγάλο βαθμό και με μεγάλη ακρίβεια σε ξένα δεδομένα. Η δεύτερη πιο καλή μέθοδος στην οπία καταλήξαμε, τα δέντρα απόφασης όπως προαναφέραμε, παρατηρήσαμε να παρουσιάζουν και αυτά την δεύτερη μεγαλύτερη ακρίβεια πρόβλεψης του μοντέλου σε ποσοστό 90% στο test dataset με 326 ατόμα να φεύγουν από την εταιρεία αλλά με 1896 να μένουν στην εταιρεία. Θεωρήσαμε ποιο σημαντικό για αυτή την ανάλυση να πάρουμε το μοντέλο το οποίο ικανοποιεί την μεγαλύτερη ακρίβεια πάνω στα σύνολο δεδομένων που διαθέτουμε και πάνω σε κάποιο άλλο μελλοντικό σύνολο δεδομένων και αυτό δεν είναι άλλο από το Random Forest.

Βιβλιογραφία

- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Kleinbaum, D. G., & Klein, M. (2002). *Logistic regression: a self-learning text*. Springer.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. CRC press.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
- Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1), 21-27.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
- Liaw, A., & Wiener, M. (2002). *Classification and regression by randomForest*. R news, 2(3), 18-22.
- Breiman, L. (2001). *Random forests*. Machine learning, 45(1), 5-32.