# Αντικείμενο της εργασίας

Μια επιχείρηση τηλεπικοινωνιών επιθυμεί να μειώσει τον αριθμό αποχώρησης των πελατών. Αρχικός στόχος είναι να ομαδοποιηθούν οι πελάτες με βάση τα χαρακτηριστικά τους. Με την κίνηση αυτή, θα μπορέσει να έχει κατηγοριοποιημένους στους πελάτες της και να αναγνωρίσει τα χαρακτηριστικά των πελατών που αποχωρούν. Δεύτερος στόχος της επιχείρησης είναι να κατασκευαστεί μοντέλο πρόβλεψης, που θα προβλέπει την αποχώρηση ή όχι του πελάτη.

# Μελέτη των δεδομένων

Τα δεδομένα που έχουμε στην διάθεση μας βρίσκονται αποθηκευμένα σε csv αρχείο με όνομα telco\_2023. Η κάθε σειρά του αρχείου αφορά έναν πελάτη της εταιρίας (που αποχώρησε ή όχι) και περιγράφονται τα εξής **19 χαρακτηριστικά** του:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Όνομα** | **Περιγραφή** | **Τιμές** |
| 1 | region | Γεωγραφική Περιοχή | {1, 2, 3, 4, 5} |
| 2 | marital | Οικογενειακή Κατάσταση | 1: Παντρεμένος 0: Ανύπαντρος |
| 3 | gender | Φύλο | 0: Άντρας 1: Γυναίκα |
| 4 | longmon | Δείκτης χρήσης Υπεραστικών κλήσεων | Πραγματικές τιμές\* |
| 5 | tollmon | Δείκτης χρήσης Υπεραστικών κλήσεων | Πραγματικές τιμές\* |
| 6 | equipmon | Δείκτης χρήσης Εξοπλισμού | Πραγματικές τιμές\* |
| 7 | cardmon | Δείκτης χρήσης Κάρτας Κλήσης | Πραγματικές τιμές\* |
| 8 | wiremon | Δείκτης χρήσης Ασύρματου Δικτύου | Πραγματικές τιμές\* |
| 9 | multline | Υπηρεσία Πολλαπλών Γραμμών | 0: ΟΧΙ 1: ΝΑΙ |
| 10 | voice | Υπηρεσία Φωνητικής | 0: ΟΧΙ 1: ΝΑΙ |
| 11 | pager | Υπηρεσία Paging | 0: ΟΧΙ 1: ΝΑΙ |
| 12 | internet | Υπηρεσία Internet | 0: ΟΧΙ 1: ΝΑΙ |
| 13 | callid | Υπηρεσία Αναγνώρισης Κλήσης | 0: ΟΧΙ 1: ΝΑΙ |
| 14 | callwait | Υπηρεσία Αναμονής Κλήσεων | 0: ΟΧΙ 1: ΝΑΙ |
| 15 | forward | Υπηρεσία Προώθησης Κλήσεων | 0: ΟΧΙ 1: ΝΑΙ |
| 16 | confer | Υπηρεσία Διάσκεψης | 0: ΟΧΙ 1: ΝΑΙ |
| 17 | ebill | Υπηρεσία Ηλεκτρονικής Πληρωμής | 0: ΟΧΙ 1: ΝΑΙ |
| 18 | custcat | Κατηγορία Πελάτη | 1: Βασικές Υπηρεσίες, 2: Ηλεκτρονικές Υπηρεσίες |
| 3: Επιπρόσθετες Υπηρεσίες, 4: Σύνολο Υπηρεσιών |
| 19 | churn | Αποχώρηση από την Εταιρεία | 0: ΟΧΙ 1: ΝΑΙ |

**Όλες οι μεταβλητές είναι αριθμητικές**, παρόλο που οι 14 είναι κατηγορικές χωρίς ιεραρχία μεταξύ τους. Οι υπόλοιπες 5 παίρνουν πραγματικές τιμές. Επιπλέον, **δεν υπάρχουν ελλιπείς τιμές** σε καμία από τις μεταβλητές/χαρακτηριστικά.

Στο παρακάτω διάγραμμα που αποτελείται από διαγράμματα διασποράς (scatter plot) φαίνονται οι συσχετίσεις μεταξύ των 5 μεταβλητών, που παίρνουν πραγματικές τιμές (longmon, tollmon, equipmon, cardmon, wiremon), με τον χρωματισμό να δείχνει αν αποχώρησαν από την εταιρία η όχι. *(Σημείωση: Τα διαγράμματα που βρίσκονται στην διαγώνιο δεν πρέπει να ληφθούν υπόψη καθώς δείχνουν την σχέση μιας μεταβλητής με τον εαυτό της.)*

Diagram, engineering drawing

Description automatically generated

Παρακάτω, βλέπουμε την συχνότητα εμφάνισης πελατών που αποχώρησαν στις μεταβλητές region, marital, gender, callid, callwait, custcat είναι οποίες παίρνουν αριθμητικές τιμές αλλά είναι κατηγορικές. Παρατηρούμε πως δεν υπάρχει σημαντική διαφοροποίηση ανά κατηγορία.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Από την εταιρία, λοιπόν, κρίθηκαν ως σημαντικότερα χαρακτηριστικά για την συσταδοποίηση των πελατών τα εξής: longmon, tollmon, equipmon, cardmon, wiremon, multline, voice, pager, internet, forward, confer, ebill.

Παρακάτω, βλέπουμε τα διαγράμματα των μεταβλητών που θεωρήθηκαν σημαντικά για την συσταδοποίηση των πελατών και την κατανομή τους ως προς την μεταβλητή churn.

Chart, bar chart

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generatedChart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

# Ζήτημα Α

Για την επίτευξη του πρώτου στόχου θα χρησιμοποιηθούν οι αλγόριθμοι συσταδοποίησης **k-mean**, **agglomerative clustering** και **DBSCAN**. Οι αλγόριθμοι συσταδοποίησης ομαδοποιούν αντικείμενα σε συστάδες (ομάδες) με βάση τα κοινά τους χαρακτηριστικά στο σύνολο δεδομένων μας.

## K-means

Ο πρώτος αλγόριθμος συσταδοποίησης που εφαρμόστηκε είναι ο **k-means**. Ο k-mean χωρίζει τους πελάτες σε k ομάδες με βάση την ομοιότητα τους. Για να προσδιορίσουμε τον αριθμό των ομάδων (k) εφαρμόζουμε την μέθοδο elbow.

Chart, line chart

Description automatically generated

Από το διάγραμμα, συμπεραίνουμε ότι η απότομη πτώση του SSE (άθροισμα τετραγώνων των σφαλμάτων) συμβαίνει όταν ο αριθμός των κλάσεων είναι 3, συνεπώς επιλέγω k=3.

Η μέθοδος για 3 συστάδες δίνει SSE= 680279.067 και τα κέντρα της καθεμίας βρίσκονται στα εξής σημεία:

|  |  |
| --- | --- |
| **Συστάδα** | **Κέντρο** |
| 0 | (16.354329, 32.16558442, 1.07229437, 23.1991342, 12.89891775, 0.44155844, 0.34199134, 0.28138528, 0.13419913, 0.84848485, 0.86147186, 0.12987013) |
| 1 | (11.00427136, 20.60175879, 43.30527638, 18.06532663, 41.21306533, 0.72361809, 0.81909548, 0.8040201, 0.81407035, 0.79396985, 0.78894472, 0.77386935) |
| 2 | (10.09719298, 3.05964912, 9.39359649, 8.46842105, 0.70675439, 0.40175439, 0.10877193, 0.06315789, 0.30701754, 0.24385965, 0.25614035, 0.32807018) |

Οι 3 αυτές συστάδες έχουν τα παρακάτω μεγέθη:

|  |  |
| --- | --- |
| **Συστάδα** | **Πλήθος πελατών σε συστάδα** |
| 0 | 231 |
| 1 | 199 |
| 2 | 570 |

Chart, bar chart

Description automatically generatedΟι συστάδες περιέχουν πελάτες που αποχώρησαν (ή όχι) με την σχετική συχνότητα που φαίνεται παρακάτω (έχει γίνει αποκοπή στο 2ο δεκαδικό ψηφίο):

|  |  |  |
| --- | --- | --- |
| **Συστάδα** | **Churn** | |
|  | 0 | 1 |
| 0 | 85.28 % | 14.71 % |
| 1 | 57.78 % | 42.21 % |
| 2 | 72.63 % | 27.36% |

Γραφικά, οι κλάσεις φαίνεται να είναι ευδιάκριτες μεταξύ τους μόνο σε ορισμένα από τα διαγράμματα διασποράς μεταξύ των συνεχών μεταβλητών.

Diagram

Description automatically generated

## Agglomerative clustering

Ο δεύτερος αλγόριθμος, με τον οποίο θα χωρίσουμε τους πελάτες της εταιρίας σε κλάσεις είναι αυτός της ιεραρχικής συσταδοποίησης (**agglomerative clustering**). Με τη μέθοδο αυτή δημιουργούμε κλάσεις ενώνοντας μικρότερες κλάσεις σε μια μεγαλύτερη. Το κάθε σημείο (παρατήρηση στο αρχείο δεδομένων) στην πρώτη φάση αποτελεί μια κλάση και ενώνεται με την κοντινότερη του. Αναζητώντας την κατάλληλη μέθοδο ανάμεσα στις single, complete και ward, η τελευταία φάνηκε να δίνει το καλύτερο αποτέλεσμα. Χρησιμοποιώντας την μέθοδο ward ο αλγόριθμος προσπαθεί να ελαχιστοποιήσει τη διακύμανση των αποστάσεων μεταξύ των κλάσεων που ενώνει. Για να μπορέσουμε να ορίσουμε τον βέλτιστο αριθμό κλάσεων σχεδιάζουμε το παρακάτω δενδρόγραμμα.

A picture containing chart

Description automatically generated

Με βάση το μεγαλύτερο άλμα που πραγματοποιείτε στο δενρόγραμμα, οι κλάσεις που φαίνεται να περιγράφουν καλύτερα τους πελάτες είναι 3.

Οι 3 αυτές συστάδες έχουν τα παρακάτω μεγέθη:

|  |  |
| --- | --- |
| **Συστάδα** | **Πλήθος πελατών σε συστάδα** |
| 0 | 493 |
| 1 | 202 |
| 2 | 305 |

Chart, bar chart

Description automatically generatedΟι συστάδες περιέχουν πελάτες που αποχώρησαν (ή όχι) με την σχετική συχνότητα που φαίνεται παρακάτω (έχει γίνει αποκοπή στο 2ο δεκαδικό ψηφίο):

|  |  |  |
| --- | --- | --- |
| **Συστάδα** | **Churn** | |
|  | 0 | 1 |
| 0 | 71.80 % | 28.19 % |
| 1 | 56.93 % | 43.06 % |
| 2 | 84.26 % | 15.73 % |

Γραφικά, οι κλάσεις φαίνεται να είναι ευδιάκριτες μεταξύ τους μόνο σε ορισμένα από τα διαγράμματα διασποράς μεταξύ των συνεχών μεταβλητών.

Calendar

Description automatically generated

## DBSCAN

Ο τρίτος και τελευταίος αλγόριθμος που θα χρησιμοποιήσουμε για την συσταδοποίηση των πελατών είναι ο **DBSCAN**. Ο DBSCAN δημιουργεί τις συστάδες ομαδοποιώντας περιοχές που είναι πυκνές σε σημεία και σημειώνοντας τα εξωτερικά σημεία σε αυτές ως απομακρυσμένα (outliers). Οι παράμετροι του αλγορίθμου που θα πρέπει να προσδιορίσουμε είναι τα eps και το ελάχιστο δείγμα (minPts). Τα eps ορίζουν το μέγεθος των κλάσεων και τα minPts τον ελάχιστο αριθμό σημείων που ορίζουν μία πυκνή περιοχή.

Για να μπορέσουμε να προσδιορίσουμε την τιμή του eps θα χρειαστούμε το k-distance graph. Για την κατασκευή του διαγράμματος δοκιμάστηκαν ως γείτονες οι αριθμοί από το 2 μέχρι και το 10, με το k-distance graph που παίρνουμε και για τους 8 είναι το παρακάτω.

Chart

Description automatically generated

Η απότομη αύξηση στο διάγραμμα παρατηρείται περίπου όταν το eps είναι 15. Συνεπώς, το επιλέγουμε ως τιμή της παραμέτρου.

Στη συνέχεια, με δοκιμές αναζητούμε, το ελάχιστο δείγμα (minPts) που δίνει το βέλτιστο αποτέλεσμα. Οι τιμές που εξετάστηκαν είναι το 5, το 6, το 7 και το 8. Το minPts επιλέχθηκε εξετάζοντας ποια συσταδοποίηση δίνει το καλύτερο αποτέλεσμα μέσω των διαγραμμάτων διασποράς των συνεχών μεταβλητών του δείγματος.

Diagram

Description automatically generatedDiagram

Description automatically generated

min\_samples=5

min\_samples=6

Diagram

Description automatically generatedDiagram

Description automatically generatedΑπό τα διαγράμματα διασποράς φαίνεται πως όταν το minPts είναι 7 τότε, δημιουργούνται κλάσεις που είναι πιο ευδιάκριτες μεταξύ τους και ίσως να έχουν και μεγαλύτερη «αξία» ως διαχωρισμός.

min\_samples=7

min\_samples=8

Εφόσον, έχουμε πλέον προσδιορίσει τις παραμέτρους μας, εφαρμόζουμε τον αλγόριθμο και δημιουργούνται 5 συστάδες που έχουν τα παρακάτω μεγέθη:

|  |  |
| --- | --- |
| **Συστάδα** | **Πλήθος πελατών σε συστάδα** |
| -1 | 95 |
| 0 | 500 |
| 1 | 60 |
| 2 | 169 |
| 3 | 176 |

Chart, bar chart

Description automatically generatedΟι συστάδες περιέχουν πελάτες που αποχώρησαν (ή όχι) με την σχετική συχνότητα που φαίνεται παρακάτω (έχει γίνει αποκοπή στο 2ο δεκαδικό ψηφίο):

|  |  |  |
| --- | --- | --- |
| **Συστάδα** | **Churn** | |
|  | 0 | 1 |
| -1 | 87.36 % | 12.63 % |
| 0 | 83.80 % | 16.20 % |
| 1 | 71.66 % | 28.33 % |
| 2 | 50.88 % | 49.11 % |
| 3 | 53.97 % | 46.03 % |

## Σύγκριση των συστάδων

Παρακάτω παρουσιάζονται τα χαρακτηριστικά **των συστάδων**:

### Text Description automatically generatedΣτον αλγόριθμο **k-means**:

Μέσοι όροι:

Text

Description automatically generatedΤυπικές αποκλίσεις:

Boxplots:

Diagram

Description automatically generated

### Στον αλγόριθμο agglomerative clustering:

Text

Description automatically generatedΜέσοι όροι:

Text

Description automatically generatedΤυπικές αποκλίσεις:

Boxplots:

Diagram

Description automatically generated

### Στον αλγόριθμο **DBSCAN**:

Text

Description automatically generatedΜέσοι όροι:

Text

Description automatically generatedΤυπική απόκλιση:

Boxplots:

Timeline

Description automatically generated with medium confidence

Έχοντας, πλέον, ολοκληρωμένη εικόνα των συστάδων που δημιούργησαν οι 3 αλγόριθμοί, παρατηρούμε ότι το μεγαλύτερο ποσοστό churn, σε συστάδα, είναι το 49.11% στην συστάδα 2, του αλγόριθμου DBSCAN. Για τον αλγόριθμο k-means η προβληματικότερη συστάδα ως προς το churn είναι η 1 με 42.21%, ενώ για τον αλγόριθμο agglomerative clustering είναι η συστάδα 1 με 43.06%.

# Ζήτημα Β

Δεύτερος στόχος της εταιρείας, είναι η κατασκευή μοντέλου που θα προβλέπει την αποχώρηση ενός πελάτη. Καλούμαστε, λοιπόν, να προβλέψουμε την τιμή της μεταβλητής churn, η οποία παίρνει τις τιμές 0 και 1 (0 όταν κάποιος δεν έχει αποχωρήσει). Συνεπώς, το πρόβλημα, ανάγεται σε ένα πρόβλημα δυαδικής κατηγοριοποίησης. Για την επίτευξη του παραπάνω στόχου γίνεται χρήση 3 αλγορίθμων κατηγοριοποίησης: των **δέντρων αποφάσεων**, του **k-nearest neighbors (knn)** και του **Naive Bayes.**

Για αρχή, επισημαίνουμε ότι **δεν έγινε χρήση όλων των χαρακτηριστικών** που μας παρέχονται στα δεδομένα για την πρόβλεψη. Οι 2 από τους 3 αλγόριθμους έδιναν καλύτερα αποτελέσματα όταν γινόταν χρήση μόνο των μεταβλητών που είχαν θεωρηθεί ως σημαντικές, στο πρώτο ζητούμενο. *(Σημείωση: τα δέντρα αποφάσεων ήταν εκείνα που έδιναν καλύτερη πρόβλεψη με το όλα τα χαρακτηριστικά που δίνονται)*

Χωρίζουμε τα δεδομένα μας σε δεδομένα εκπαίδευσης (train) και «δοκιμής» (test). Το 25% πηγαίνει στο test σύνολο και το 75% στο train.

Επιπλέον, θα χρησιμοποιηθεί η μέθοδος του cross validation για να αποφύγουμε την περίπτωση του overfitting και να εκπαιδεύσουμε το μοντέλο σε πολλά και διαφορετικά train-test splits.

Κάνοντας έναν αρχικό έλεγχο διαπιστώσαμε πως υπάρχει ανισορροπία νόμος στα δεδομένα μας η αναλογία μεταξύ πελατών που παρέμειναν στην εταιρεία και πελατών που έφυγαν είναι 2.65. αυτό καθίσταται τα δεδομένα μας **unbalanced**. Τα εργαλεία που θα χρησιμοποιηθούν είναι τέτοια ώστε, η ανισορροπία των δεδομένων να μην έχει μεγάλες επιπτώσεις στην προβλεπτική ικανότητα του μοντέλου. Ένα τέτοιο παράδειγμα είναι το StratifiedKFold που χωρίζει τα δεδομένα σε train και test διατηρώντας ισορροπία μεταξύ των κατηγοριών που θέλουμε να προβλέψουμε.

Σημαντικό για το μοντέλο μας είναι να προβλέπει σωστά την αποχώρηση ενός πελάτη για αυτό το λόγο θα χρησιμοποιηθεί ως μετρική των μοντέλων η ευαισθησία (**recall**).

## Decision Tree

Το πρώτο μοντέλο πρόβλεψης, που θα φτιάξουμε, χρησιμοποιεί **δέντρα αποφάσεων.** Τα δέντρα αποφάσεων είναι ένας αλγόριθμος κατηγοριοποίησης που χωρίζει τα δεδομένα σε μικρότερα σύνολα με βάση συγκεκριμένες συνθήκες (επί των μεταβλητών/χαρακτηριστικών) καταλήγοντας στην κατηγοριοποίηση της κάθε παρατήρησης.

Τα δεδομένα μας χωρίστηκαν σε 11 διαφορετικά train-test σύνολα για να εφαρμόσουμε cross validation. Το δέντρο αποφάσεων με τις καλύτερες επιδόσεις έχει recall=0.4806 (στρογγυλοποίηση στο 4ο δεκαδικό ψηφίο) και accuracy=0.664.

Πιο συγκεκριμένα, οι μετρικές ανάλογα με το label του churn παρουσιάζουν τα συγκεκριμένα αποτελέσματα ανά μετρική:

|  |  |  |
| --- | --- | --- |
|  | **0** | **1** |
| **Precision** | 0.76630435 | 0.37878788 |
| **Recall** | 0.77472527 | 0.36764706 |
| **F-score** | 0.7704918 | 0.37313433 |

Στο παρακάτω διάγραμμα αναπαρίσταται γραφικά το confusion matrix και βλέπουμε την κατανομή των true label με τα predicted label.

Chart, treemap chart

Description automatically generated

Το βέλτιστο μοντέλο που κατασκευάστηκε με χρήση των δέντρων αποφάσεων απεικονίζεται στο παρακάτω σχήμα.

Chart, scatter chart

Description automatically generated

## Knn

Ο δεύτερος αλγόριθμος που εφαρμόστηκε για την δημιουργία μοντέλου που προβλέπει την αποχώρηση των πελατών είναι ο **knn**. Εφαρμόζοντας τον, βρίσκουμε τα k πιο κοντινά σημεία κατηγοριοποιούμε το σημείο που εξετάζουμε με βάση αυτά.

Για τις τιμές του k δοκιμάσαμε τους ακεραίους από το 2 μέχρι και το 11 και διαπιστώθηκε ότι η καλύτερη απόδοση ήταν του μοντέλου με k=3. Στο μοντέλο knn με k=3 εφαρμόσαμε επίσης cross validation.

Το μοντέλο με k=3 έδωσε accuracy=0.74 και recall=0.4418 (στρογγυλοποίηση στο 4ο δεκαδικό ψηφίο)

Πιο συγκεκριμένα, οι μετρικές ανάλογα με το label του churn παρουσιάζουν τα συγκεκριμένα αποτελέσματα ανά μετρική:

|  |  |  |
| --- | --- | --- |
|  | **0** | **1** |
| **Precision** | 0.79104478 | 0.53061224 |
| **Recall** | 0.87362637 | 0.38235294 |
| **F-score** | 0.83028721 | 0.44444444 |

Στο παρακάτω διάγραμμα αναπαρίσταται γραφικά το confusion matrix και βλέπουμε την κατανομή των true label με τα predicted label.

Chart, treemap chart

Description automatically generated

## Naive Bayes

Ο τρίτος και τελευταίος αλγόριθμος που θα εφαρμόσουμε για να την κατασκευή μοντέλου πρόβλεψης είναι ο **Naive Bayes.** Εφαρμόζουμε και πάλι cross validation, αυτή την φορά χωρίζοντας τα δεδομένα σε 8 διαφορετικά train-test splits. επιπλέον δοκιμάστηκε η απόδοση του Naive Bayes με threshold 0.5, 0.4, 0.3, 0.2 και επιλέχθηκε το 0.4 ως το καλύτερο.

Το accuracy του μοντέλου αυτού είναι 0.74. Πιο συγκεκριμένα, οι μετρικές ανάλογα με το label του churn παρουσιάζουν τα συγκεκριμένα αποτελέσματα ανά μετρική:

|  |  |  |
| --- | --- | --- |
|  | **0** | **1** |
| **Precision** | 0.8253012 | 0.46428571 |
| **Recall** | 0.75274725 | 0.57352941 |
| **F-score** | 0.78735632 | 0.51315789 |

Στο παρακάτω διάγραμμα αναπαρίσταται γραφικά το confusion matrix και βλέπουμε την κατανομή των true label με τα predicted label.

Chart, treemap chart

Description automatically generated

## Βέλτιστο μοντέλο πρόβλεψης

Από τα παραπάνω αποτελέσματα και λόγω του ότι μας ενδιαφέρει να προβλέπουμε την αποχώρηση ενός πελάτη, το μοντέλο που δημιουργήθηκε με τον Naïve Bayes έχει την καλύτερη επίδοση.