**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

**Σχολή Χρηματοοικονομικής και Στατιστικής**

**Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης**

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

# ΜΕΘΟΔΟΙ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΑΣΦΑΛΙΣΗ ΥΓΕΙΑΣ

Μπουντούλης Χρήστος

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

Πειραιάς
Σεπτέμβριος 2022

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

**Σχολή Χρηματοοικονομικής και Στατιστικής**

**Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης**

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

# ΜΕΘΟΔΟΙ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΑΣΦΑΛΙΣΗ ΥΓΕΙΑΣ

Μπουντούλης Χρήστος

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

Πειραιάς
Σεπτέμβριος 2022

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

Τα μέλη της Επιτροπής ήταν:
- Μπερσίμης Σωτήριος (Αναπληρωτής καθηγητής) (Επιβλέπων)
- Χατζηκωνσταντινίδης Ευστάθιος (Αναπληρωτής καθηγητής)
- Ξένος Παναγιώτης (Επίκουρος καθηγητής)

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

**UNIVERSITY OF PIRAEUS**

**School of Finance and Statistics**

**Department of Statistics and Insurance Science**

POSTGRADUATE PROGRAM IN
**APPLIED STATISTICS**

# DATA ANALYTICS AND MACHINE LEARNING METHODS FOR HEALTH INSURANCE

By

Bountoulis Christos

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

Piraeus, Greece
September 2022

*Στην Οικογένειά μου*

# Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω θερμά τον καθηγητή και μέντορά μου κ. Σωτήριο Μπερσίμη για την επιστημονική του καθοδήγηση, τη συνεχή υποστήριξή του και την εμπιστοσύνη που μου έδειξε για την εκπόνηση της. Επίσης, θα ήθελα να τον ευχαριστήσω για όλες τις υποδείξεις και συμβουλές του καθώς και όλες τις γνώσεις που αποκόμισα καθ΄ όλη τη διάρκεια των σπουδών μου.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στη οικογένεια μου και τη σύζυγό μου για τη στήριξη, την κατανόηση και τη συμπαράστασή τους καθ' όλη τη διάρκεια των σπουδών μου.

# Περίληψη

Μπουντούλης Χρήστος

**ΜΕΘΟΔΟΙ ΑΝΑΛΥΤΙΚΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΑΣΦΑΛΙΣΗ ΥΓΕΙΑΣ**

Σεπτέμβριος 2022

Σήμερα ο όγκος των δεδομένων που έχουν στη διάθεσή τους οι εταιρείες του ασφαλιστικού κλάδου, αυξάνεται με ιλιγγιώδεις ρυθμούς οδηγώντας στη δημιουργία «δεξαμενών» τεράστιου όγκου ανομοιογενών δεδομένων. Η ανάληψη στρατηγικών αποφάσεων με αξιοποίηση δομημένων αλλά και αδόμητων πηγών δεδομένων σε πραγματικό χρόνο, αποτελεί μεγάλη πρόκληση για τον ασφαλιστικό κλάδο, καθώς και στρατηγικό πλεονέκτημα για τις εταιρείες που καταφέρνουν να διαχειριστούν και να ξεκλειδώσουν όλη αυτή την πληροφορία που κρύβεται στην καρδιά των δεδομένων, σε μια εποχή όπου τα περισσότερα προϊόντα είναι τυποποιημένα και δε διαφέρουν ιδιαίτερα από εταιρεία σε εταιρεία. Σε αυτή την εργασία, θα γίνει μία εκτενής περιγραφή σε προβλήματα που εφάπτονται του τομέα της Ασφάλισης Υγείας με τα οποία οι εταιρείες βρίσκονται πολύ συχνά αντιμέτωπες και στη συνέχεια θα παρουσιαστούν εφαρμογές που οδηγούν στην ανάληψη στρατηγικών αποφάσεων με τη χρήση μεθόδων αναλυτικής των δεδομένων και στατιστικής μηχανικής μάθησης.

# Abstract

Bountoulis Christos

**DATA ANALYTICS AND MACHINE LEARNING METHODS FOR HEALTH INSURANCE**

September 2022

In today's era, the volume of data available to companies in the insurance industry is growing rapidly, leading to the creation of huge volumes of heterogeneous data. Making strategic decisions using structured and unstructured data sources in real time is a major challenge for the insurance industry, as well as a strategic advantage for companies that manage and unlock all this information hidden at the heart of the data, in a time that most products are standardized and do not differ particularly from company to company. In this document, an extensive description of the health insurance problems that companies very often face will be carried out, and then applications leading to strategic decisions will be presented using data analytics and Statistical Machine Learning methods.

# Περιεχόμενα

# Κ Ε Φ Α Λ Α Ι Ο 1

## 1. Εισαγωγή

### 1.1. Ο ασφαλιστικός κλάδος και ο τεράστιος όγκος δεδομένων

Ο ασφαλιστικός κλάδος από τη φύση του είναι ένας κλάδος που διαχειρίζεται μεγάλο όγκο δεδομένων εδώ και πολλά χρόνια, με τη δραστηριότητα των ασφαλιστικών ιδρυμάτων να βασίζεται στην ανάλυση τους για την κατανόηση και την αποτελεσματική αξιολόγηση του κινδύνου, γεγονός που οδηγεί στο συμπέρασμα ότι τα δεδομένα αυτά είναι κυρίαρχη δύναμη στο συγκεκριμένο κλάδο, παίζοντας καθοριστικό ρόλο στον στρατηγικό σχεδιασμό και στη λήψη αποφάσεων των ιδρυμάτων.

Ωστόσο, τα τελευταία χρόνια ο όγκος των δεδομένων που έχουν στη διάθεσή τους τα ασφαλιστικά ιδρύματα αυξάνεται εκθετικά, λόγω της δημιουργίας πολλαπλών πηγών άντλησης πληροφορίας, όπως είναι τα εσωτερικά πληροφοριακά συστήματα που διαθέτουν, όσο και εξωτερικά κανάλια, όπως είναι τα μέσα κοινωνικής δικτύωσης, κινητά τηλέφωνα, δορυφορικές κεραίες, κάμερες, μικρόφωνα, λογισμικά φωνητικής καταγραφής, συσκευές με βιοαισθητήρες και άλλες διαφορετικές πηγές που διαμορφώνουν ένα τεράστιο σύνολο ανομοιογενών δεδομένων. Συγκεκριμένα, με βάση έρευνα από το International Data Group, προβλέπεται ότι θα υπάρχουν παγκοσμίως 163 ζεταμπάιτ δεδομένων μέχρι το 2025.

### 1.2. Η επιστήμη των δεδομένων και η Στατιστικη Μηχανική Μάθηση

Η επιστήμη των δεδομένων διαμορφώνεται ως η γνώση που προκύπτει από την τομή τριών κύριων γνωστικών πεδίων, (α) της επιστήμης της Στατιστικής (Statistical Science), (β) της επιστήμης της Πληροφορικής (Computer Science) και (γ) του εκάστοτε τομέα οικονομικής δραστηριότητας. Η συνταγή της επιτυχίας για την επίτευξη των εταιρικών στόχων προϋποθέτει την αξιοποίηση όλων των διαθέσιμων δεδομένων που είναι απαραίτητα για τη χάραξη επιτυχούς στρατηγικής και η επιστήμη των δεδομένων παίζει καθοριστικό ρόλο σε αυτό. Ένας πολύτιμος πυλώνας της επιστήμης των δεδομένων είναι η τομή των επιστημών της Στατιστικής και της Πληροφορικής, δηλαδή, η Στατιστική Μηχανική Μάθηση, η οποία αφορά στην δημιουργία μοντέλων πρόβλεψης ή προτύπων από ένα σύνολο δεδομένων, αντικαθιστώντας με πολύ μεγαλύτερη επιτυχία τον τρόπο με τον οποίο ο ανθρώπινος νους επιχειρεί να κατανοήσει το περιβάλλον του μέσω της παρατήρησης και της γνώσης που μπορεί να λάβει μέσω της ανάλυσης των δεδομένων. Στοιχεία από διάφορες πηγές αναλύονται και με τη χρήση κατάλληλων τεχνικών και πολύπλοκων αλγορίθμων μετασχηματίζουν τα δεδομένα σε γνώση, δημιουργώντας τις προϋποθέσεις για τη βελτιστοποίηση της εκμετάλλευσης όλων των διαθέσιμων πόρων, συμβάλλοντας έτσι στην κερδοφορία και την ανάπτυξη βέλτιστων πρακτικών σε τομείς όπως η διοίκηση κινδύνου, προωθητικές δραστηριότητες ασφαλιστικών προϊόντων, εξυπηρέτηση πελατών κ.α.

## 1.3. Η έννοια της Ασφάλισης και οι κλάδοι της Ασφαλιστικής Αγοράς

Σύμφωνα με τη βιβλιογραφία, η «Ασφάλιση» είναι μια σύμβαση, που αντιπροσωπεύεται από ένα συμβόλαιο στο οποίο ο ασφαλισμένος λαμβάνει οικονομική προστασία ή αποζημίωση έναντι ζημιών από μια ασφαλιστική εταιρεία. Διαμέσου της ασφαλιστικής σύμβασης, μεταφέρονται τυχαίοι και απρόβλεπτοι κίνδυνοι στους Ασφαλιστές έναντι ενός αντιτίμου το λεγόμενο και ως ασφάλιστρο, συμφωνώντας να αποζημιώσουν τους ασφαλισμένους για τυχαίες ζημίες. Πρακτικά, τα ασφαλιστήρια συμβόλαια χρησιμοποιούνται για την αντιστάθμιση οικονομικών ζημιών, με τους Ασφαλιστές να συγκεντρώνουν όλους τους κινδύνους των ασφαλισμένων και να αναλαμβάνουν τη διαχείρισή τους με σκοπό τη διαφύλαξη τους από την επέλευση ενός κινδύνου, τις συνέπειες του οποίου σε άλλη περίπτωση δε θα μπορούσαν να διαχειριστούν.

Σύμφωνα με τα άρθρα 4 & 5 του Ν. 4364/2016 οι ασφαλίσεις χωρίζονται σε 2 βασικές κατηγορίες:

- Κλάδοι Ασφαλίσεων κατά ζωής
- Κλάδοι Ασφαλίσεων κατά ζημιών

Οι ασφαλίσεις κατά ζωής ταξινομούνται στους ακόλουθους κλάδους:

1. **Ασφαλίσεις ζωής**: Επιβίωσης ή θανάτου, προσόδων, σωματικών βλαβών, αναπηρίας κ.α.
2. **Ασφαλίσεις γάμου και γεννήσεως**
3. **Ασφαλίσεις ζωής συνδεδεμένες με επενδύσεις**
4. **Διαρκής ασφάλιση Ασθένεια** (μη υποκείμενης σε ακύρωση από τον ασφαλιστή)
5. **Τοντίνες**: Δημιουργία ομάδων για την από κοινού συγκέντρωση κεφαλαίου και τη διανομή του είτε μεταξύ των επιζώντων, είτε μεταξύ των κληρονόμων των αποθανόντων
6. **Εργασίες κεφαλαιοποίησης**: Η επιχείρηση αναλαμβάνει ασφαλιστικές υποχρεώσεις για ορισμένο χρονικό διάστημα και για ορισμένο ποσό έναντι προκαθορισμένων καταβολών
7. **Διαχείριση συνολικών συνταξιοδοτικών κεφαλαίων ή οργανισμών**
8. **Ομαδικά προγράμματα πρόνοιας σύμφωνα με το κεφάλαιο Ι, τίτλος IV του βιβλίου IV του Γαλλικού κώδικα ασφαλίσεων**
9. **Εργασίες κοινωνικής ασφάλισης**

Οι ασφαλίσεις κατά ζημιών ταξινομούνται στους ακόλουθους κλάδους:

1. **Ατυχήματα**
2. **Ασθένειες**
3. **Χερσαία Οχήματα** (εκτός σιδηροδρομικών)
4. **Σιδηροδρομικά Οχήματα**
5. **Αεροσκάφη**
6. **Πλοία**
7. **Μεταφερόμενα Εμπορεύματα**: Καλύπτει κάθε ζημία την οποία υφίστανται τα μεταφερόμενα εμπορεύματα, περιλαμβανομένων αποσκευών και κάθε άλλου αγαθού, ανεξαρτήτων του μεταφορικού μέσου
8. **Πυρκαγιά και στοιχεία της φύσης**
9. **Λοιπές ζημίες αγαθών**: Περιλαμβάνει καλύψεις που υφίστανται τα αγαθά, εξαιρουμένων των περιπτώσεων 3,4,5, 7 και 8 (π.χ. χαλάζι, παγετός, κλοπή κ.α.)
10. **Αστική ευθύνη από χερσαία αυτοκίνητα οχήματα**

11. **Αστική ευθύνη από αεροσκάφη**
12. **Αστική ευθύνη από θαλάσσια λιμναία και ποτάμια σκάφη**
13. **Γενική αστική ευθύνη**
14. **Πιστώσεις**: Καλύπτει γενική αφερεγγυότητα, εξαγωγικές πιστώσεις, (εκτός εξαγωγικών πιστώσεων που γίνονται για λογαριασμό ή με την υποστήριξη του Κράτους), πωλήσεις με δόσεις, ενυπόθηκες πιστώσεις, αγροτικές πιστώσεις.
15. **Εγγυήσεις**: Άμεσες εγγυήσεις, έμμεσες εγγυήσεις
16. **Διάφορες χρηματικές απώλειες**: Κάλυψη κινδύνων όπως: απώλεια επαγγελματικής απασχόλησης, απώλεια κερδών, τρέχοντα γενικά έξοδα, γενική ανεπάρκεια εισοδήματος απρόβλεπτες εμπορικές δαπάνες κ.α.
17. **Νομική προστασία**: Περιλαμβάνει την ανάληψη δικαστικών εξόδων και την παροχή νομικής προστασίας
18. **Βοήθεια**: Περιλαμβάνει την ανάληψη της υποχρέωσης άμεσης παροχής βοήθειας, στις περιπτώσεις και με τους όρους που προβλέπει σύμβαση, σε χρήμα ή σε είδος, έναντι προηγούμενης καταβολής ασφαλίστρου, προς πρόσωπα, που περιέρχονται σε δυσχερή θέση κατά τη διάρκεια μετακινήσεων ή απουσίας από την κατοικία ή από τον τόπο συνήθους διαμονής τους είτε υπό άλλες περιστάσεις ανεξάρτητα από μετακίνηση ή απουσία. Η σε είδος παροχή βοήθειας είναι δυνατόν να συνίσταται και στην χρησιμοποίηση του προσωπικού και του εξοπλισμού που ανήκουν σε αυτόν που παρέχει την βοήθεια. Δεν συνιστούν υπηρεσίες βοήθειας οι υπηρεσίες συντήρησης ή διατήρησης, η εξυπηρέτηση μετά την πώληση, ούτε η απλή υπόδειξη ή πρόβλεψη παροχής βοήθειας ως μεσολάβηση

Σύμφωνα με στατιστικά που δημοσιεύτηκαν στην επίσημη ιστοσελίδα του Οργανισμού Οικονομικής Συνεργασίας και Ανάπτυξης (OECD - The Organisation for Economic Co-operation and Development) το σύνολο των μικτών ασφαλίστρων στο σύνολο των 38 κρατών μελών του Οργανισμού είναι ίσο με 5,225 τρισεκατομμύρια δολάρια, με το 48,24% (2,52 τρις $) να ανήκει στους κλάδους ζωής και το υπόλοιπο 51,76% (2,705 τρις $) στους κλάδους κατά ζημιών.

Στην παρούσα εργασία θα γίνει εκτενής αναφορά στις εφαρμογές της επιστήμης των δεδομένων στην ασφάλιση υγείας, που είναι μέρος του κλάδου ζωής.

## 1.4. Η σημαντικότητα της επιστήμης των δεδομένων στη ασφάλιση

Η συμβολή της επιστήμης των δεδομένων στην ασφάλιση είναι η ίδια με τις άλλες βιομηχανίες, καθώς οι ασφαλιστικές εταιρείες είναι επιχειρήσεις των οποίων φυσικά ο στόχος είναι η αύξηση της αποδοτικότητας, η βελτιστοποίηση των στρατηγικών μάρκετινγκ, η βελτίωση της κερδοφορίας και η μείωση του κόστους. Προβλήματα, στα οποία έρχεται να δώσει λύσεις η επιστήμη των δεδομένων, μερικά εκ των οποίων παρουσιάζονται και στην παρούσα διπλωματική εργασία είναι τα εξής:

**Η Διαχείριση Απαιτήσεων (Claim Management)**: Οι ασφαλιστικές απαιτήσεις, είναι ένας από τους σημαντικότερους παράγοντες του κύκλου εργασιών των ασφαλιστικών εταιρειών καθώς αποτελούν τον κύριο πυλώνα δαπανών τους, που μάλιστα έχει και το υψηλότερο ποσοστό απόπειρας ασφαλιστικής απάτης. Δεδομένου του γεγονότος αυτού, η έγκυρη πρόβλεψη των μελλοντικών απαιτήσεων αποτελεί μοχλό βελτιστοποίησης των εταιρειών και κατ’ επέκταση, έναν από τους πρώτους τομείς που αναζητούν λύσεις οι ασφαλιστικές εταιρείες

μέσω της της τεχνητής νοημοσύνης και της στατιστικής μηχανικής μάθησης. Με αυτόν τον τρόπο, οι ασφαλιστικές εταιρείες τιμολογούν σε ανταγωνιστικές τιμές τα ασφάλιστρά τους, τα οποία δεν πρέπει να είναι ούτε πολύ υψηλά, αλλά ούτε και πολύ χαμηλά, αυξάνοντας έτσι το περιθώριο κέρδους τους και παραμένοντας ένα βήμα μπροστά από τον ανταγωνισμό.

**Η Ανίχνευση της Απάτης (Fraud Detection)**: Η ασφαλιστική απάτη προκαλεί τεράστια οικονομική ζημία στις ασφαλιστικές εταιρείες κάθε χρόνο αποτελώντας αδιαμφισβήτητα μια από τις σημαντικότερες προκλήσεις που έχουν να αντιμετωπίσουν οι ασφαλιστικές εταιρείες στα πλαίσια του ψηφιακού τους μετασχηματισμού. Η αναλυτική των δεδομένων, η μηχανική μάθηση και η τεχνητή νοημοσύνη καταστούν δυνατή την ανίχνευση δραστηριοτήτων ασφαλιστικής απάτης και η χρήση αυτών από τις εταιρείες μπορούν να οδηγήσουν στην αποτελεσματική ανίχνευση της. Αυτά τα μοντέλα βασίζονται σε ιστορικά καταγεγραμμένα και επιβεβαιωμένα δεδομένα απόπειρας ασφαλιστικής απάτης τα οποία εντοπίζουν συσχετίσεις μεταξύ ύποπτων ιστορικά δραστηριοτήτων και βοηθούν στην άμεση αναγνώριση της απόπειρας απάτης μελλοντικά

**Η κατηγοριοποίηση πελατών (Customer Segmentation)**: Οι ασφαλιστικές εταιρείες δαπανούν ένα μεγάλο μέρος του προϋπολογισμού τους στον τομέα του μάρκετινγκ, για τη δημιουργία καμπανιών, τη διατήρησή των υπαρχόντων πελατών, καθώς και την προσέλκυση νέων. Για να επιτευχθεί αυτό όμως με όσο το δυνατόν λιγότερο κόστος και με περισσότερη στόχευση, βασική προϋπόθεση είναι η αναγνώριση των διαφόρων προφίλ πελατών, μέσω της κατανόησης των συμπεριφορικών χαρακτηριστικών τους, των δημογραφικών τους στοιχείων, της πιστοληπτικής τους ικανότητας, της συνέπειας τους κ.λπ.. Με τη χρήση μοντέλων μηχανικής μάθησης και τεχνητής νοημοσύνης τα οποία θα βασίζονται σε ιστορικά δεδομένα πελατών μπορεί να πραγματοποιηθεί μια πιο ουσιαστική κατηγοριοποίηση των πελατών, βασισμένη σε συσχετίσεις και πρότυπα που οι παραδοσιακές τεχνικές αγνοούν. Η ομαδοποίηση της πελατειακής βάσης μιας ασφαλιστικής εταιρείας και η ανάλυση της απόδοσης αυτών των ομαδοποιήσεων μπορεί να βελτιώσει τις στρατηγικές μάρκετινγκ, τις πωλήσεις και την εξυπηρέτηση πελατών. Επίσης, η ομαδοποίηση των πελατών ενισχύει την κατανόηση για τον τρόπο με τον οποίο οι πελάτες από διάφορες ομάδες αλληλοεπιδρούν με την εταιρεία.

**Η πρόβλεψη της Διασταυρούμενης πώλησης και της πώλησης της πιο ακριβής εκδοχής ενός υπάρχοντος προϊόντος (Cross-selling/Up-selling):** Για τις ασφαλιστικές εταιρίες, μοχλός ανάπτυξης είναι η πώληση προϊόντων, ωστόσο ο προτιμώμενος τρόπος είναι μέσω του υφιστάμενου πελατολογίου, καθώς η εύρεση νέων πελατών είναι αρκετά πιο δαπανηρή διαδικασία. Επομένως, οι εταιρίες βασίζονται στα δεδομένα που έχουν στην κατοχή τους, με σκοπό να κατανοήσουν τη συμπεριφορά των πελατών, να κάνουν πιο στοχευμένες προωθήσεις προϊόντων και να αποκομίσουν το μέγιστο δυνατό κεφάλαιο που ένας πελάτης μπορεί να διαθέσει. Η αναλυτική των δεδομένων και η μηχανική μάθηση αποτελούν εξαιρετικά εργαλεία στη διαδικασία εντοπισμού των υφιστάμενων πελατών, οι οποίοι θα αποκρίνονταν θετικά, με υψηλή πιθανότητα, στην πρόταση για αγορά ενός άλλου προϊόντος ή της πιο ακριβής εκδοχής του προϊόντος για το οποίο ενδιαφέρονται. Το αποτέλεσμα της εφαρμογής των μεθόδων αυτών είναι η αύξηση των εσόδων, η επιτάχυνση της ταχύτητας της διαδικασίας και η μείωση των δαπανών που συνδέονται κύρια με τους απαιτούμενους ανθρώπινους πόρους και τα έξοδα επικοινωνίας.

**Η Διατήρηση των πελατών:** Η διατήρηση της υπάρχουσας πελατείας είναι το πρωταρχικό μέλημα κάθε επιχείρησης, καθώς είναι λιγότερο κοστοβόρο από την αναζήτηση νέων. Με τη

χρήση της Αναλυτικής των δεδομένων και της μηχανικής μάθησης, μπορεί να διενεργηθεί η πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο και να καθοριστεί ποιος τύπος προωθητικής προσφοράς μπορεί να επιτύχει το καλύτερο αποτέλεσμα για την υψηλότερη αφοσίωση και διατήρηση, προσωπικά για κάθε πελάτη. Με βάση αυτό, το σύστημα εξατομικεύεται αυτόματα η προωθητική προσφορά χωρίς την ανάγκη ανθρώπινης παρέμβασης.

**Η Βελτιστοποίηση Τιμής:** Η διαδικασία βελτιστοποίησης τιμών των ασφαλιστικών προϊόντων είναι μια περίπλοκη διεργασία που απαιτεί συνδυασμούς διαφόρων μεθόδων και αλγορίθμων που δίνουν στις εταιρίες την δυνατότητα να προσαρμόσουν ρεαλιστικά και δυναμικά τα ασφάλιστρα τους. Με τη χρήση στατιστικής μηχανικής μάθησης, στοιχεία κόστους, εξόδων, αξιώσεων, κινδύνων και κερδοφορίας αναλύονται με σκοπό να εκτιμηθεί η μελλοντική τους συμπεριφορά.

**Η Προσωποποιημένη Προώθηση:** Οι πελάτες προτιμούν να λαμβάνουν εξατομικευμένες υπηρεσίες που ταιριάζουν με τις ανάγκες και τον τρόπο ζωής τους. Αναλύοντας τα χαρακτηριστικά των πελατών (Ατομικά, Δημογραφικά χαρακτηριστικά κ.λ.π) καθώς και ιστορικά τους στοιχεία με τη χρήση της μηχανικής μάθησης μπορούν να δημιουργηθούν εξατομικευμένες προτάσεις ικανοποιώντας τις προσδοκίες τους και βελτιστοποιώντας ταυτόχρονα την αποδοτικότητα του εταιρικού marketing.

**Η Εκτίμηση Κινδύνου:** Η ανάλυση δεδομένων και η στατιστική μηχανική μάθηση μπορούν να βοηθήσουν στην αξιολόγηση της κατηγορίας κινδύνου που ανήκει κάθε υποψήφιος πελάτης. Αυτό ολοκληρώνεται επίσης μέσω τεχνικών ομαδοποίησης, λαμβάνοντας υπόψη όχι μόνο μεμονωμένες βαθμολογίες για διάφορους παράγοντες κινδύνου, αλλά και τη συσχέτιση μεταξύ αυτών.

# ΚΕΦΑΛΑΙΟ 2

# 2. Βιβλιογραφική ανασκόπηση σε εφαρμογές στην Ασφάλιση Υγείας

## 2.1. Εισαγωγή

Σήμερα η πλειοψηφία των ασφαλιστικών εταιρειών οδηγείται σε επανασχεδιασμό της στρατηγικής τους, επανεξετάζοντας τα επιχειρηματικά τους μοντέλα με σκοπό τη μεταπήδηση σε ένα συμβατό, ασφαλές και ψηφιακά ενεργοποιημένο μοντέλο λειτουργίας. Οι αιτίες είναι η εκθετικά αυξανόμενη πληροφορία και οι συνεχώς αυξανόμενες προκλήσεις, σε συνδυασμό με το μεγάλο ανταγωνισμό που επικρατεί στην ασφαλιστική αγορά. Επομένως, δημιουργείται η ανάγκη στους ασφαλιστικούς φορείς να επενδύσουν σε νέες τεχνολογίες, και να αναζητούν νέες λύσεις στα προβλήματά τους μέσω της ανάλυσης των δεδομένων που έχουν, ή που θα μπορούσαν να έχουν στην κατοχή τους, με σκοπό τη λήψη των επιχειρηματικών αποφάσεων τους με μεγαλύτερη ακρίβεια. Όπως είπε σε ένα απόσπασμα της ομιλίας του ο Peter Sondergaard (2011), ο οποίος είναι ανώτερος αντιπρόεδρος και παγκόσμιος επικεφαλής της έρευνας στην Gartner, Inc, *«Η πληροφορία είναι το πετρέλαιο του 21ου αιώνα και η ανάλυση η μηχανή εσωτερικής καύσης».*

Έπειτα από μία διερευνητική ανασκόπηση στη βιβλιογραφία, εντοπίστηκαν κάποιες από τις πιο συνήθεις εφαρμογές τη αναλυτικής των δεδομένων και της στατιστικής μηχανικής μάθησης στην ασφάλιση υγείας και γενικότερα στον τομέα της ασφάλισης. Όλες αναφέρθηκαν συνοπτικά στο προηγούμενο κεφάλαιο και περιγράφονται αναλυτικά στις επόμενες ενότητες.

## 2.2. Πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστικό συμβόλαιο (Churn prediction)

Στις μέρες μας, λόγω της καθιέρωσης του διαδικτύου ως βασικό μέσω για έρευνα αγοράς, είναι πιο εύκολο για κάθε πελάτη να συγκρίνει τα συμβόλαιά του με τα υπόλοιπα της ασφαλιστικής αγοράς, να λάβει νέες προσφορές και να μεταπηδήσει από μία εταιρία σε μία άλλη. Η προτεραιότητα μιας επιχείρησης, ανεξαρτήτως επιχειρηματικού κλάδου είναι να διατηρήσει τους πελάτες της και να χτίσει μία πιστή σχέση μαζί τους, αποφεύγοντας έτσι το κόστος απόκτησης νέων πελατών. Σύμφωνα με τους [1]Independent Insurance agents of Dallas (IIAD), το μέσο ποσοστό διατήρησης στον ασφαλιστικό κλάδο είναι 84%, ενώ έχει παρατηρηθεί ότι υπάρχει πολύ ισχυρή συσχέτιση των κερδών μιας εταιρείας με το υψηλό ποσοστό διατήρησης πελατών. Επιπλέον, ο ασφαλιστικό κλάδος, έχει το υψηλότερο κόστος απόκτησης πελατών από κάθε άλλο κλάδο, με το κόστος απόκτησης πελατών να είναι 7 (επτά) έως 9 (εννιά) φορές μεγαλύτερο από το αντίστοιχο για τη διατήρηση ενός πελάτη. Έχουν πραγματοποιηθεί εκατοντάδες δημοσιεύσεις με εφαρμογές της στατιστικής, της στατιστικής μηχανικής μάθησης και της αναλυτικής των δεδομένων στη πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από ένα ασφαλιστήριο συμβόλαιο και κάποιες από αυτές περιγράφονται παρακάτω.

***Πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από την Ασφαλιστική εταιρία με μια προσέγγιση πολλαπλών ασφαλιστικών προϊόντων***

Αυτή η μελέτη του *Mauricio Henao Madrigal* et al. *(2020)* βασίστηκε σε δεδομένα από ασφαλιστικές εταιρίες τις Λατινικής Αμερικής σε πελάτες που είχαν στην κατοχή τους πάνω από ένα ασφαλιστήρια συμβόλαια και οι συγγραφείς προσέγγισαν το πρόβλημα της αποχώρησης των πελατών από την εκάστοτε ασφαλιστική εταιρεία αναπτύσσοντας μοντέλα Ανάλυσης Επιβίωσης, με σκοπό τον εντοπισμό των στατιστικά σημαντικών μεταβλητών που σχετίζονται με αυτή την διάρκεια «ζωής» των ασφαλισμένων και την εκτίμηση του χρόνου παραμονής ενός πελάτη στην εταιρεία μετά την πρώτη ακύρωση ενός ασφαλιστικού συμβολαίου. Συγκεκριμένα, η έρευνα εστιάστηκε σε ένα χαρτοφυλάκιο 11 (έντεκα) προϊόντων, μεταξύ των οποίων ήταν, 1 (ένα) προϊόν ασφάλισης οχημάτων, 4 (τέσσερις) τύποι ασφάλισης ζωής, 2 (δύο) τύποι προϊόντων ασφάλισης υγείας, 1 (ένα) προϊόν σε ασφάλισης κατοικίας και 1 (ένα) σε ασφάλιση πυρκαγιάς.

Οι ερευνητές, εφάρμοσαν το μη παραμετρικό κριτήριο ανάλυσης επιβίωσης των Kaplan-Meier, για να εκτιμήσουν τον αναμενόμενο χρόνο παραμονής των πελατών μετά την πρώτη ακύρωση ενός συμβολαίου, αγνοώντας τις συμμεταβλητές που μπορεί να επηρεάζουν το μοντέλο, με σκοπό να έχουν μια εικόνα της κατανομής των χρόνων επιβίωσης. **Ο Kaplan-Meier εκτιμητής και η αθροιστική συνάρτηση κινδύνου Hazard δίνονται από τους τύπους:**

$$\hat{S}(t) = \prod_{j/t_j \le t} \frac{n_{j-} d_j}{n_j} \quad \& \; H(t) = \; -\ln \mathrm{S}(t)$$

Όπου, $n_j$ είναι ο αριθμός των πελατών που είναι σε κίνδυνο τον χρόνο $t_j$ και $d_j$ είναι ο αριθμός που αποχώρησαν στον χρόνο $t_j$.

Από την παραπάνω ανάλυση διαπιστώθηκε ότι ο ενδιάμεσος χρόνος αποχώρησης των πελατών μετά την πρώτη ακύρωση, είναι τα δύο έτη, αναδεικνύοντας τη σημασία παρακολούθησης και λήψης μέτρων από τις ασφαλιστικές εταιρίες, προκειμένου να διατηρήσουν τη πίστη των πελατών τους.

Στη συνέχεια, εφάρμοσαν **το ημιπαραμετρικό μοντέλο αναλογικού κινδύνου του Cox**,

$$\lambda(t \backslash z) = \lambda_0(t) \exp\,(Z'\beta),$$

όπου $\lambda_0(t)$ είναι μία απροσδιόριστη συνάρτηση κινδύνου, για να εκτιμήσουν τον αναμενόμενο χρόνο παραμονής του πελάτη στην εταιρία μετά την πρώτη ακύρωση προϊόντος, λόγω του γεγονότος ότι τα δεδομένα ήταν ανομοιογενή και οι χρόνοι επιβίωσης των πελατών στην εταιρεία εξαρτώνται και από διάφορες άλλες ερμηνευτικές μεταβλητές, όπως το φύλο, η ηλικία, το κανάλι πώλησης, η περιοχή διαμονής του πελάτη, ο αριθμός ασφαλιστικών απαιτήσεων του πελάτη, α αριθμός ασφαλιστικών προϊόντων και άλλες. Συγκεκριμένα, ένα από τα πορίσματα ήταν ότι στην ασφάλιση υγείας οι πελάτες κατά την πρώτη ακύρωση έχουν μεγαλύτερο κίνδυνο αποχώρησης από την εταιρεία σε σχέση με αυτούς που ακυρώνουν άλλα προϊόντα.

***Πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από την Ασφαλιστική εταιρία που δραστηριοποιείται σε ασφάλειες οχημάτων***

Αυτή η έρευνα των *Maria Spiteri και Azzopardi (2021)* βασίστηκε σε δεδομένα από μία εταιρία ασφάλισης αυτοκινήτου της Μάλτας και είχε σκοπό την εύρεση των παραγόντων που σχετίζονται με την αποχώρηση των πελατών από την εταιρία. Συγκεκριμένα, χρησιμοποίησαν στατιστικά μοντέλα ταξινόμησης με σκοπό την κατηγοριοποίηση των πελατών σε αυτούς υπάρχει κίνδυνος αποχώρησης και σε αυτούς που δεν υπάρχει κίνδυνος αποχώρησης και στη

συνέχεια διενέργησαν ανάλυση επιβίωσης με σκοπό την εύρεση του χρόνου μέχρι την αποχώρηση από το ασφαλιστήριο συμβόλαιο. Για την ταξινόμηση χρησιμοποιήθηκαν οι εξής αλγόριθμοι μηχανικής μάθησης.

- Δέντρα αποφάσεων (Decision trees)
- Λογιστική παλινδρόμηση (Logistic regression)
- Random forest
- Naïve Bayes και
- Support Vector Machines (SVM)

Έπειτα από χρήση και των τεσσάρων αλγορίθμων κατέληξαν στο συμπέρασμα, ότι ο καλύτερος αλγόριθμος για την ταξινόμηση των πελατών σε αυτούς υπάρχει κίνδυνος αποχώρησης και σε αυτούς που δεν υπάρχει κίνδυνος αποχώρησης από το ασφαλιστήριο συμβόλαιο ήταν ο Random Forest, ο οποίος είχε 89.9% ποσοστό συνολικής ορθής ταξινόμησης. Ο αλγόριθμος αυτός παρουσιάζεται στο επόμενο κεφάλαιο.

## 2.3. Ανίχνευση της ασφαλιστικής απάτης (Fraud detection)

Ένα από τα πιο σημαντικά προβλήματα που αντιμετωπίζουν οι ασφαλιστικές φορείς είναι η ασφαλιστική απάτη η οποία προκαλεί σημαντικές ζημίες. Σύμφωνα με το [2]«National Health Care Anti-Fraud Association» περίπου το 3% ($56 δισεκατομμύρια) από τα $1.87 τρισεκατομμύρια που δαπανούνται για την υγειονομική περίθαλψη στην Ευρωπαϊκή Ένωση κάθε χρόνο, αποτελούν ασφαλιστική απάτη. Αντίστοιχα, το 3% ($68 δισεκατομμύρια) από τα $2.27 τρισεκατομμύρια που δαπανούνται κάθε χρόνο, χάνονται λόγω απάτης, σπατάλης και κατάχρησης στο σύστημα υγείας των Η.Π.Α., ωστόσο οι ειδικού ισχυρίζονται, ότι το πραγματικό ποσοστό ασφαλιστικής απάτης διαφέρει σημαντικά από αυτό και ότι η πραγματική έκταση των δαπανών παραμένει ακόμα ανερμήνευτη. Ασφαλιστική απάτη στον κλάδο της Υγείας, μπορεί να πραγματοποιηθεί, από τον ασφαλιστικό σύμβουλο, τον ασθενή, τον γιατρό του κέντρο υγειονομικής περίθαλψης, από το κέντρο υγειονομικής περίθαλψης, από τρίτα μέρη και φυσικά από όλους τους συνδυασμούς των προηγούμενων. **Μερικές από τους πιο συνήθεις τύπους ασφαλιστικής απάτης, σπατάλης και κατάχρησης είναι:**

- Η χρέωση για υπηρεσίες που δεν παρασχέθηκαν στον ασθενή
- Ο ασυνήθιστα υψηλός αριθμός τιμολογίων για έναν συγκεκριμένο ασφαλισμένο σε σύντομο χρονικό διάστημα
- Η κατευθυνόμενη συνταγογράφηση
- Ο υπερβολικός αριθμός αιτήσεων για φάρμακα σε μια συγκεκριμένη περίοδο
- Η διπλή υποβολή αίτησης για την ίδια υπηρεσία
- Η παραπλανητική περιγραφή της παρεχόμενης υπηρεσίας
- Η χρέωση για μια πιο σύνθετη ή ακριβή υπηρεσία σε σχέση με αυτή που παρασχέθηκε στην πραγματικότητα
- Η υπερχρέωση των υπηρεσιών και επιλογή των πιο ακριβών υλικών/μεθόδων/υπηρεσιών
- Η παροχή υγειονομικής φροντίδας σε περιπτώσεις που δε κρίνεται απαραίτητο, όπως η προτροπή στον ασθενή να προβεί σε μη απαραίτητες εξετάσεις ή χειρουργεία με σκοπό την άντληση του μεγαλύτερου δυνατού χρηματικού ποσού από αυτόν, εφόσον αυτό καλύπτεται από τον εκάστοτε ασφαλιστικό φορέα

Να σημειωθεί ότι πολλές από τις παραπάνω ασφαλιστικές απάτες, μπορεί να συμβούν ταυτόχρονα ανά περιστατικό. Ωστόσο, οι ζημίες αυτές επηρεάζουν άμεσα όλους τους

ασφαλισμένους, καθώς οι ασφαλιστικές εταιρίες για να αντισταθμίσουν τον κίνδυνο οδηγούνται σε αυξήσεις των ασφαλίστρων. Επομένως, είναι αισθητή η σημαντικότητα της επίβλεψης και πρόληψης της ασφαλιστικής απάτης, με τους επιστήμονες να έχουν πραγματοποιήσει εκατοντάδες δημοσιεύσεις με προτεινόμενες εφαρμογές της αναλυτικής δεδομένων και της στατιστικής στη πρόβλεψη της ασφαλιστικής απάτης, όπου κάποιες από αυτές περιγράφονται παρακάτω.

***Μια προσέγγιση Ανίχνευσης απάτης με τεχνικές εξόρυξης δεδομένων στον τομέα της Ασφάλισης Υγείας***

Αυτή η μελέτη των *Melih Kirlidog & Cuneyt Asuk (2012),* βασίστηκε σε δεδομένα από μία Τουρκική ασφαλιστική εταιρία και η βάση δεδομένων που χρησιμοποιήθηκε περιλαμβάνει αναλυτικές πληροφορίες για τους πελάτες, τις ασφαλιστικές απαιτήσεις, καθώς και τους επιχειρηματικούς συνέταιρους, όπως νοσοκομεία, φαρμακευτικές εταιρίες, συνεργαζόμενα κέντρα υγείας, ελεύθερους επαγγελματίες γιατρούς κ.λ.π. Η ανάλυση ανίχνευσης ανωμαλιών πραγματοποιήθηκε με μία μη εποπτευόμενη μέθοδο της στατιστικής μηχανικής μάθησης που βασίζεται σε έναν αλγόριθμο συσταδοποίησης μηχανικής διανυσματικής υποστήριξης (SVM), ο οποίος είναι ένα εργαλείο πρόβλεψης, ταξινόμησης και παλινδρόμησης που χρησιμοποιεί τη θεωρία της μηχανικής μάθησης για τη μεγιστοποίησης της προγνωστικής ακρίβειας, αποφεύγοντας τα υπερβολικά ταιριαστά δεδομένα. Σκοπός ήταν να εντοπιστούν τυχόν ανωμαλίες ως προς το ύψος των αποζημιώσεων, οι οποίες όπως προέκυψε, αποτελούν λιγότερο από το 1% των συνολικών αποζημιώσεων. Ωστόσο, εντοπίστηκαν σημαντικές διαφοροποιήσεις στο ποσοστό των ανωμαλιών που εντοπίστηκαν ανά πάροχο και προέκυψε ότι το μεγαλύτερο ποσοστό παρατηρήθηκε στα νοσοκομεία, τα φαρμακεία, πολυκλινικές και ελεύθερους επαγγελματίες γιατρούς, όπου οι μέσες αποζημιώσεις τους ήταν σημαντικά υψηλότερες από τις υπόλοιπες περιπτώσεις. Οι συγγραφείς προτείνουν τις τεχνικές εξόρυξης γνώσεις, όπως δημιουργία κανόνων αποφάσεων, και κατηγοριοποίηση με ή χωρίς τη γνώση των περιπτώσεων απάτης για τη διενέργεια ανάλυσης ανίχνευσης ανωμαλιών.

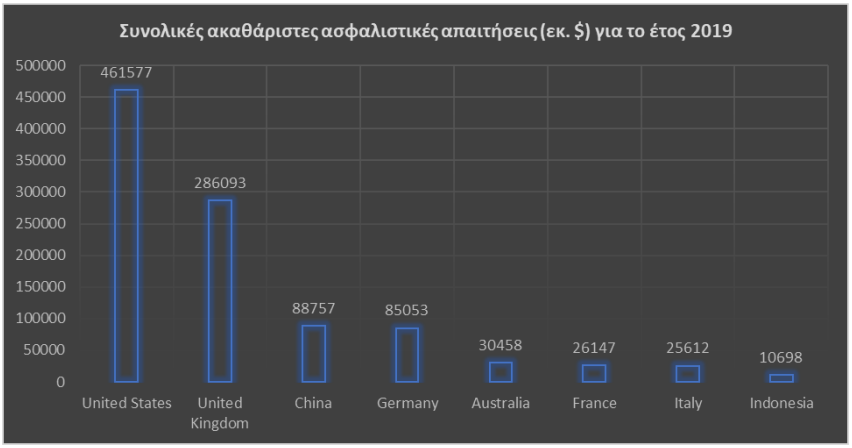
***Μαρκοβιανό μοντέλο με ενσωμάτωση μηχανικής μάθησης για ανίχνευση απάτης στον τομέα της Ασφάλισης Υγείας***

Αυτή η έρευνα των Rohan Yashraj και άλλων *(2021)* βασίστηκε σε δεδομένα από μία ασφαλιστική εταιρία της Ινδίας και είχε ως σκοπό την εφαρμογή στατιστικών μεθόδων για την ανίχνευση περιπτώσεων ασφαλιστικής απάτης στην ασφάλιση υγείας. Συγκεκριμένα, εφάρμοσαν ένα μοντέλο εξάρτησης του Markov και ένα επιπλέον βελτιωμένο Μαρκοβιανό μοντέλο χρησιμοποιώντας τη μέθοδο της ενίσχυσης κλίσης (gradient boosting method), τα οποία είναι μαθηματικά μοντέλα που ανήκουν σε μια μεγάλη οικογένεια στοχαστικών – πιθανοθεωρητικών μοντέλων. Το σύνολο δεδομένων που χρησιμοποίησαν ήταν 382,587 αιτήματα αποζημίωσης από τις οποίες οι 38,082, δηλαδή περίπου το 9.95% αποτελούσαν περιπτώσεις επιβεβαιωμένης ασφαλιστικής απάτης. Από το μοντέλο εξάρτησης του Markov, προέκυψε 94.07% ακρίβεια ταξινόμησης των αποζημιώσεων ασφαλιστικής απάτης, με το F1-score το οποίο έχει ιδιαίτερη βαρύτητα σε περιπτώσεις που υπάρχουν άνισες κατανομές κλάσεων να είναι ίσο με 0.6683. Όσον αφορά το Μαρκοβιανό μοντέλο με τη μέθοδο της ενίσχυσης κλίσης (gradient boosting method), εμφανίστηκε αρκετά βελτιωμένο με την ακρίβεια ταξινόμησης να είναι ίση με 97.10% και το F1-score να είναι ίσο με 0.8546, ενώ το ποσοστό λανθασμένων θετικών περιπτώσεων παρατηρήθηκε αρκετά μικρότερο από το αντίστοιχο του απλού μοντέλου εξάρτησης του Markov.

## 2.4. Πρόβλεψη των ασφαλιστικών απαιτήσεων των πελατών (Health insurance claim prediction)

Λόγω της κλιμάκωσης του κόστους υγειονομικής περίθαλψης, η πρόβλεψη του κόστους που θα υποστούν οι ασθενείς αποτελεί σημαντικό παράγοντα τόσο για τις ασφαλιστικές εταιρίες όσο και για τους παρόχους υγειονομικής περίθαλψης. Υπάρχει πληθώρα μαθηματικών πρακτικών που χρησιμοποιούν οι αναλογιστές για να προβλέψουν το ποσό των ετήσιων αναμενόμενων αποζημιώσεων στον τομέα της ιατροφαρμακευτικής περίθαλψης σε μια ασφαλιστική εταιρεία. Το ποσό αυτό πρέπει να περιλαμβάνεται στους ετήσιους οικονομικούς προϋπολογισμούς και μια λανθασμένη εκτίμηση έχει γενικά αρνητικές επιπτώσεις στη συνολική απόδοση της επιχείρησης.

Το ποσό των συνολικών ακαθάριστων ασφαλιστικών απαιτήσεων στον κλάδο της υγείας είναι το υψηλότερο στον ασφαλιστικό κλάδο παγκοσμίως επομένως η πρόβλεψη των ασφαλιστικών απαιτήσεων κρίνεται απαραίτητη και συμβάλλει άμεσα στη βιωσιμότητα και την ανάπτυξη των ασφαλιστικών εταιρειών. Ακολουθεί, ένα ενδεικτικό ραβδόγραμμα (Σχήμα 2.1) με τις συνολικές ακαθάριστες ασφαλιστικές απαιτήσεις σε οκτώ (8) από τις μεγαλύτερες χώρες στον χάρτη του ασφαλιστικού κλάδου, σύμφωνα με δεδομένα από τον Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (OECD).




**Σχήμα 2.1:** Ραβδόγραμμα με τις συνολικές ακαθάριστες ασφαλιστικές απαιτήσεις σε οκτώ (8) από τις μεγαλύτερες χώρες στον χάρτη του ασφαλιστικού κλάδου

Βασική αιτία αύξηση του κόστους είναι τα λάθη που γίνονται κατά την διεκπεραίωση των ασφαλιστικών απαιτήσεων και σε περιπτώσεις σφαλμάτων πληρωμής η εκ νέου διεκπεραίωση των απαιτήσεων αυξάνει το διοικητικό κόστος. Οι ερευνητές διεθνώς έχουν δημοσιεύσει πολλές εφαρμογές που αναφέρονται στην πρόβλεψή των ασφαλιστικών απαιτήσεων παρέχοντας και δύο (2) από αυτές περιγράφονται παρακάτω.

### *Πρόβλεψη Ασφαλιστικών απαιτήσεων στην Ασφάλιση υγείας με τη χρήση τεχνητών νευρωνικών δικτύων*

Στη μελέτη των *Sam Goundar* και άλλων *(2020)* παρουσιάζεται η ανάπτυξη μοντέλου τεχνητών νευρωνικών δικτύων τα οποία σύμφωνα με τους *Chapko et al. (2011) και El-said et al. (2013)* προτείνονται ως κατάλληλα για την πρόβλεψη των αναμενόμενων ετήσιων ιατρικών ισχυρισμών. Συγκεκριμένα, χρησιμοποιήθηκαν δεδομένα από τις ιατρικές απαιτήσεις που έχει καταβάλει τα τελευταία 12 χρόνια η εταιρία BSP Life (Fiji) Ltd, η οποία παρέχει ασφάλεια υγείας και ζωής στα νησιά Φίτζι. Το ύψος των ασφαλιστικών απαιτήσεων που καταβάλλονται

σε διάστημα ενός έτους αυξάνει άμεσα τις συνολικές δαπάνες τις εταιρίας κατά πολλά εκατομμύρια δολάρια επηρεάζοντας έτσι το περιθώριο κέρδους της και σκοπός είναι η βελτιστοποίηση της πρόβλεψης των ασφαλιστικών απαιτήσεων αντικαθιστώντας τα παραδοσιακά μοντέλα με νέα πιο σύγχρονα και ακριβή μοντέλα πρόβλεψης όπως είναι αυτά των νευρωνικών δικτύων. Με αυτό τον τρόπο μπορεί να πραγματοποιηθεί με μεγαλύτερη ακρίβεια η προετοιμασία των ετήσιων οικονομικών προϋπολογισμών καθώς και των ασφαλιστήριων συμβολαίων, αυξάνοντας έτσι το περιθώριο κέρδους. Εστιάζοντας, στη δημιουργία ενός πολυστρωματικού νευρωνικού δικτύου τροφοδοσίας προς τα εμπρός με αλγόριθμο οπισθοδιάδωσης που βασίζεται στη μέθοδο *gradient descent*. Το δίκτυο εκπαιδεύτηκε χρησιμοποιώντας δεδομένα ιατρικών αξιώσεων τα τελευταία 12 χρόνια και αφού έγινε η βελτιστοποίηση των παραμέτρων του, δημιουργήθηκε το τελικό μοντέλο το οποίο σημείωσε ελάχιστο μέσο απόλυτο ποσοστό σφάλματος (MAPE) μειωμένο κατά 11.5% σε σχέση με το παραδοσιακό μοντέλο πρόβλεψης, πετυχαίνοντας 90.38% ακρίβεια στο σύνολο δεδομένων εκπαίδευσης και 93.58% στο σύνολο δεδομένων δοκιμής.

***Ανάπτυξη μοντέλου πρόβλεψης ιατρικού κόστους βάσει της στατιστικής μηχανικής μάθησης με χρήση δεδομένων αξιώσεων ασφάλισης υγείας***

Η Ιαπωνική κυβέρνηση ζήτησε από τις ασφαλιστικές εταιρείες να αναδιαμορφώσουν τη διαχείριση της υγείας του πληθυσμού κατασκευάζοντας ένα μοντέλο με σκοπό την καλύτερη πρόβλεψη του αναμενόμενου ύψους ιατρικού κόστους των ασθενών, αντικαθιστώντας τα παραδοσιακά γραμμικά μοντέλα πρόβλεψης με νέα μοντέλα βασισμένα σε τεχνικές στατιστικής μηχανικής μάθησης. Στη μελέτη των Takeshima T. και άλλων (2018), χρησιμοποιήθηκαν δεδομένα αποζημιώσεων έξι ασφαλιστικών εταιρειών υγείας (1.009.167 ασφαλισμένοι, μέση ηλικία: 34,3, γυναίκες: 54%) για το έτος 2016 και κατασκευάστηκε ένα μοντέλο παλινδρόμησης με τη χρήση της μεθόδου LASSO, η οποία είναι μια μέθοδος που παράγει αραιά μοντέλα, δηλαδή μοντέλα τα οποία περιέχουν μόνο ένα υποσύνολο παραγόντων από τους αρχικά διαθέσιμους, μειώνοντας παράλληλα τη διασπορά των σφαλμάτων. Από τους στατιστικά σημαντικούς παράγοντες του μοντέλου που δημιουργήθηκε, το 58% ήταν μεμονωμένες ασθένειες και το 24%, το 14% και το 4% ήταν όροι αλληλεπίδρασης δύο, τριών και τεσσάρων ασθενειών αντίστοιχα. Συγκεκριμένα, μεταξύ των παραγόντων, η δυσκοιλιότητα είχε τον υψηλότερο αντίκτυπο στο ιατρικό κόστος ακολουθούμενη από την αϋπνία, το βρογχικό άσθμα, το διαβήτη και τον πόνο στη μέση. Μεταξύ των όρων αλληλεπίδρασης των δύο παραγόντων, η αϋπνία και το βρογχικό άσθμα είχαν τον υψηλότερο αντίκτυπο. Τέλος, η συνολική προβλεπτική ικανότητα του μοντέλου πρόβλεψης ήταν ίση με 42%, υψηλότερη κατά 17% από αυτή των παραδοσιακών γραμμικών μοντέλων παλινδρόμησης.

***Πρόβλεψη ημερών νοσηλείας στο νοσοκομείο με χρήση ασφαλιστικών απαιτήσεων στον τομέα της υγείας***

Οι διαχειριστές υγειονομικής περίθαλψης σε όλο τον κόσμο προσπαθούν να μειώσουν το κόστος της υγειονομικής περίθαλψης βελτιώνοντας παράλληλα την ποιότητα των παρεχόμενων υπηρεσιών υγείας. Η νοσηλεία είναι η πιο δαπανηρή συνιστώσα στον τομέα της υγείας, επομένως, η έγκαιρη αναγνώριση εκείνων που διατρέχουν υψηλότερο κίνδυνο νοσηλείας θα βοηθούσε τους διαχειριστές της υγειονομικής περίθαλψης και τις ασφαλιστικές εταιρίες που δραστηριοποιούνται στον τομέα της υγείας να αναπτύξουν καλύτερα σχέδια και στρατηγικές. Σε αυτή τη μελέτη, των Xang Xie και άλλων *(2015)* χρησιμοποιήθηκαν δεδομένα εισαγωγών σε νοσοκομεία και ασφαλιστικών απαιτήσεων από ένα σύνολο 247,075 ατόμων σε διάστημα τριών ετών με σκοπό τη διενέργεια πρόβλεψης του αριθμού των ημερών νοσηλείας

στο νοσοκομείο κατά το τελευταίο έτος. Για την πρόβλεψη χρησιμοποιήθηκε ένας αλγόριθμος παλινδρόμησης βασισμένος σε δέντρα αποφάσεων ο οποίος φάνηκε να είναι αποδοτικός τόσο στον γενικό πληθυσμό όσο και σε υποπληθυσμούς, πετυχαίνοντας ακρίβεια πρόβλεψης στο γενικό πληθυσμό ύψους 84.3%. Ωστόσο, σε άτομα ηλικίας άνω των 63 ετών φάνηκε να είναι ακόμα πιο αποδοτικός σε σχέση με το γενικό σύνολο.

## 2.5. Βελτιστοποίηση τιμής των ασφαλιστικών προϊόντων (Price optimization)

Καθώς το επίπεδο του ανταγωνισμού αυξάνεται, η βελτιστοποίηση τιμολόγησης αποκτά κεντρικό ρόλο στις περισσότερες ώριμες ασφαλιστικές αγορές υποχρεώνοντας τους ασφαλιστές να βελτιστοποιήσουν την αξιολόγησή τους και να λάβουν πιο σοβαρά υπόψη τη συμπεριφορά των πελατών τους.

Δεν υπάρχει καθολικά αποδεκτός ορισμός της βελτιστοποίησης τιμών. Ωστόσο, οι ρυθμιστικές αρχές των ασφαλιστικών αρχών το περιγράφουν γενικά ως τη χρήση εξελιγμένων εργαλείων εξόρυξης δεδομένων και τεχνικών μοντελοποίησης από έναν ασφαλιστή κατά τη διαδικασία καθορισμού των επιτοκίων για να διαφοροποιούνται τα ποσοστά με βάση άλλους παράγοντες εκτός από τον κίνδυνο απώλειας ενός ατόμου. Επομένως, η διαδικασία βελτιστοποίησης τιμών των ασφαλιστικών προϊόντων είναι μια περίπλοκη διεργασία που απαιτεί συνδυασμούς διαφόρων μεθόδων και αλγορίθμων. Ο στόχος της βελτιστοποίησης τιμών είναι να χρεώσει ένα ασφαλισμένο άτομο το υψηλότερο ποσό που θα ανεχθεί πριν αγοράσει εναλλακτική κάλυψη ή δεν ανανεώσει ένα ασφαλιστήριο συμβόλαιο. Στοιχεία κόστους, εξόδων, αξιώσεων, κινδύνων, κερδοφορίας αναλύονται και εκτιμάται η μελλοντική τους συμπεριφορά. Αναπτύσσονται έτσι αλγόριθμοι που δίνουν στις εταιρίες την δυνατότητα να προσαρμόσουν ρεαλιστικά και δυναμικά τα ασφάλιστρα τους.

### *Μέθοδοι Στατιστικής Μηχανικής Μάθησης για τη βελτιστοποίηση της τιμολόγησης. Μια σύγκριση με τα τυπικά Γενικευμένα Γραμμικά Μοντέλα (GLM)*

Στα πλαίσια του συνεχώς αυξανόμενου ανταγωνισμού, οι ασφαλιστικές εταιρίες, αναζητούν συνεχώς τις βέλτιστες λύσεις για την τιμολόγηση των ασφαλιστικών συμβολαίων. Στη μελέτη των *Giorgio Spedicato και άλλων (2021)* διερευνάται η δυνατότητα νέων τεχνικών στατιστικής μηχανικής μάθησης, όπως τα ενισχυμένα μοντέλα δέντρων αποφάσεων Gradient Boosting (GB) και Extreme Gradient Boosting (XGBoost) για τη βελτιστοποίηση του προτεινόμενου ασφαλίστρου στους υποψήφιους ασφαλισμένους. Από τη διερεύνηση, προκύπτει ότι τα μοντέλα μηχανικής εκμάθησης μπορούν να προσφέρουν υψηλότερη ακρίβεια σε σχέση με τα γενικευμένα γραμμικά μοντέλα (GLM), ωστόσο ακόμη διερευνάται αν το προγνωστικό κέρδος απόδοσης των αλγορίθμων, είναι αρκετό ώστε να οδηγήσει στην ευρεία υιοθέτησή τους, καθώς οι προσεγγίσεις τους όσον αφορά τον βελτιστοποιημένο όγκο ασφαλιστηρίων, είχαν σχεδόν παρόμοια αποτελέσματα. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι σημαντικοί παράγοντες του μοντέλου, είχαν σχεδόν γραμμική συμπεριφορά στη λογαριθμική κλίμακα. Ως εκ τούτου, τα μοντέλα στατιστικής μηχανικής μάθησης έχουν σαφές πλεονέκτημα από τα γενικευμένα γραμμικά μοντέλα, όταν χρησιμοποιούνται σε περιπτώσεις marketing σε υποψήφιους πελάτες, παρά κατά τη βελτιστοποίηση των ασφαλίστρων, ενώ τα γενικευμένα γραμμικά μοντέλα έχουν σαφές πλεονέκτημα ως προς τους υπολογιστικούς πόρους, απαιτώντας πολύ μικρό υπολογιστικό χρόνο σε σχέση με τα μοντέλα μηχανικής μάθησης.

Στις προηγούμενες ενότητες έγινε αναφορά σε σημαντικά προβλήματα που αντιμετωπίζουν οι ασφαλιστικές εταιρίες, ώστε να παραμείνουν ανταγωνιστικές στο σύγχρονο ασφαλιστικό χάρτη, το μέγεθος των οποίων καλούνται να εκτιμήσουν, όπως είναι η οι πελάτες που πρόκειται να αποχωρήσουν από ένα ασφαλιστήριο συμβόλαιο, ο εντοπισμός της ασφαλιστικής απάτης, η σωστή διαχείριση των ασφαλιστικών απαιτήσεων των πελατών, καθώς και η βελτιστοποίηση της τιμής των ασφαλιστικών προϊόντων. Η σωστή εκτίμηση όλων αυτών των προβλημάτων, συνεισφέρει αθροιστικά στην εκτίμηση του κινδύνου που αντιμετωπίζουν οι ασφαλιστικές εταιρίες, οδηγώντας σε πιο ακριβείς προβλέψεις του μέλλοντος. Αυτό συνεπάγεται της αύξησης του περιθωρίου κέρδους και της μείωσης της οικονομικής ζημίας που αποτελεί το μείζον ζήτημα κάθε επιχείρησης.

# ΚΕΦΑΛΑΙΟ 3

## 3. Μηχανική Μάθηση

### 3.1. Εισαγωγή

Η Μηχανική μάθηση (Machine Learning) είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνική νοημοσύνη. Συγκεκριμένα, διερευνά τη μελέτη και τη κατασκευή αλγορίθμων που μπορούν να λαμβάνουν τεράστιες ποσότητες δεδομένων και να μαθαίνουν από αυτά, χρησιμοποιώντας την επιστήμη της στατιστικής και των πιθανοτήτων για τη λήψη αποφάσεων με μια εύλογη ακρίβεια, χωρίς να είναι απαραίτητη η ανθρώπινη παρέμβαση.

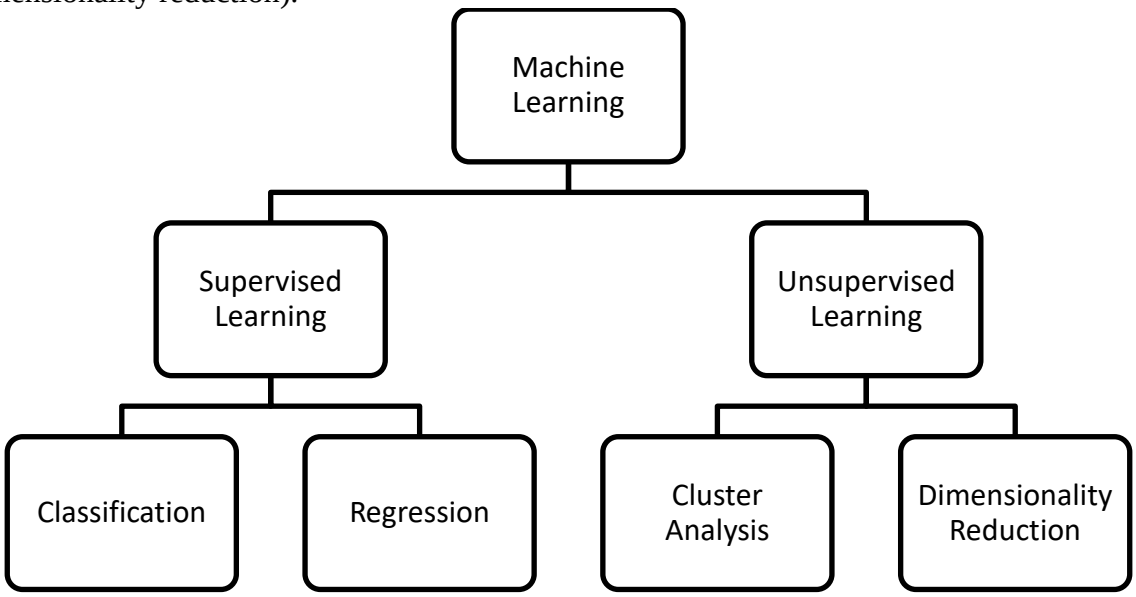
Η μηχανική μάθηση διακρίνεται συνήθως σε 2 μεγάλες κατηγορίες οι οποίες θα αναλυθούν εκτενώς στις παρακάτω ενότητες. Η πρώτη κατηγορία είναι η εποπτευόμενη μηχανική μάθηση (Supervised machine learning), κατά την οποία γίνεται η εκμάθηση των προτύπων και συσχετίσεων μεταξύ ενός συνόλου μεταβλητών εισόδου και μιας μεταβλητής εξόδου, με σκοπό τη δημιουργία προβλέψεων σε νέα άγνωστα σύνολα δεδομένων. Η δεύτερη κατηγορία είναι η μη εποπτευόμενη μηχανική μάθηση (Unsupervised machine learning), κατά την οποία οι αλγόριθμοι δε χρειάζονται δεδομένα για εκπαίδευση και προσπαθούν από μόνοι τους να ανακαλύψουν συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα χωρίς να είναι γνωστό αν υπάρχουν και ποια είναι.

### 3.2. Στατιστική και μηχανική μάθηση

Η διαφορά μεταξύ στατιστικής και μηχανικής μάθησης έχει αποτελέσει αντικείμενο μακροχρόνιας συζήτησης. Σύμφωνα με το Τμήμα Στατιστικής του Πανεπιστημίου της Καλιφόρνια, Irvine στατιστική ορίζεται ως «η επιστήμη που ασχολείται με την ανάπτυξη και τη μελέτη μεθόδων συλλογής, ανάλυσης, ερμηνείας και παρουσίασης εμπειρικών δεδομένων». Οι δυο κύριες στατιστικές μέθοδοι είναι η περιγραφική στατιστική, η οποία χρησιμοποιείται για τη διερευνητική ανάλυση των δεδομένων και η επαγωγική στατιστική, η οποία είναι η διαδικασία γενίκευσης των συμπερασμάτων που προέκυψαν κατά την ανάλυση ενός δείγματος, σε όλο τον πληθυσμό. Οι τομείς της μηχανικής μάθησης και της στατιστικής φαίνεται να είναι άρρηκτα συνδεδεμένοι μεταξύ τους, με όλο και περισσότερους στατιστικούς να κάνουν χρήση της μηχανικής μάθησης με αποτέλεσμα τη δημιουργία ενός ακόμη πεδίου, αυτού της στατιστικής μάθησης. Παρά τις σημαντικές τους ομοιότητες και το γεγονός ότι έχουν κοινό στόχο, σημαντική διαφορά ανάμεσα στους δύο τομείς αποτελεί ο τρόπος επεξεργασίας των δεδομένων, αφού σε αντίθεση με τη στατιστική που απαιτεί κατανόηση της συλλογής των δεδομένων και επιλογή των κατάλληλων παραμέτρων για τη δημιουργία προγνωστικών, η μηχανική μάθηση βασίζεται στον όγκο των δεδομένων και εφαρμόζει κύρια αλγοριθμικές διαδικασίες . Συγκεκριμένα, οι αλγόριθμοι που εφαρμόζονται χρησιμοποιούν το σύνολο των δεδομένων και καταλήγουν στην επιλογή των παραμέτρων που θα οδηγήσουν σε μια επιτυχημένη πρόβλεψη, η ακρίβεια της οποίας αυξάνεται καθώς αυξάνεται και ο αριθμός των διαθέσιμων δεδομένων. Ουσιαστικά, στόχος της μηχανικής μάθησης εκτός της ανάλυσης των δεδομένων με χρήση στατιστικών τεχνικών είναι η βελτιστοποίηση των αποτελεσμάτων, ενώ η στατιστική εστιάζει κύρια στην ανάλυση και στα συμπεράσματα που προκύπτουν από την ανάλυση.

## 3.3. Είδη μηχανικής μάθησης

Τα πιο διαδεδομένα είδη μηχανικής μάθησης, είναι, η εποπτευόμενη μάθηση (Supervised learning) και η μη εποπτευόμενη μάθηση (Unsupervised learning). Η εποπτευόμενη μηχανική μάθηση χωρίζεται σε προβλήματα ταξινόμησης (classification) και προβλήματα παλινδρόμησης (regression), ενώ η μη εποπτευόμενη μηχανική μάθηση χωρίζεται σε προβλήματα κατηγοριοποίησης (cluster analysis) και προβλήματα μείωσης διαστάσεων (dimensionality reduction).




**Σχήμα 3.1**: Είδη μηχανικής μάθησης

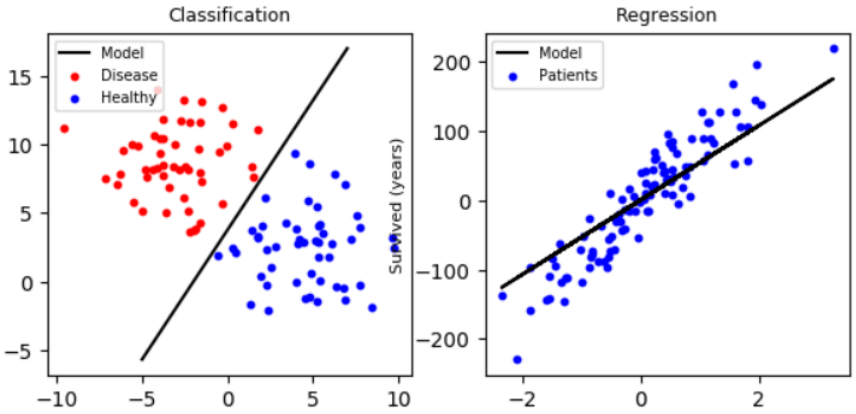
Στην παρούσα εργασία θα χρησιμοποιηθεί η διαδικασία της εποπτευόμενης μάθησης για τη δημιουργία προγνωστικών.

### 3.3.1. Εποπτευόμενη μάθηση (Supervised learning)

Κατά την εποπτευόμενη μάθηση, οι αλγόριθμοι δέχονται σαν είσοδο ($X_1$, $X_2$, $X_3$, …., $X_v$) πειραματικά δεδομένα τα οποία εκπαιδεύονται με σκοπό την εύρεση προτύπων και σαν έξοδο (Y) δέχονται τιμές σε περίπτωση προβλημάτων παλινδρόμησης ή πιθανότητες σε περίπτωση προβλημάτων ταξινόμησης, αναλόγως το είδος του προβλήματος. Μόλις ολοκληρωθεί η εκπαίδευση των δεδομένων, ο αλγόριθμος θα εφαρμόσει όσα έμαθε σε νέα δεδομένα με σκοπό την εξαγωγή προβλέψεων.

Η εποπτευόμενη μάθηση χρησιμοποιείται στις ακόλουθες περιπτώσεις προβλημάτων:

- σε προβλήματα παλινδρόμησης (regression) που στοχεύουν στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών, όπως είναι το παράδειγμα της πρόβλεψη του ύψους των ασφαλιστικών απαιτήσεων των πελατών μιας ασφαλιστικής εταιρίας, ή της κατηγοριοποίησης των ανθρώπων σε υγιείς ή ασθενείς
- σε προβλήματα ταξινόμησης (classification) που στοχεύουν στην δημιουργία μοντέλων πρόβλεψης διακριτών κατηγοριών-τάξεων, όπως είναι το παράδειγμα της κατηγοριοποίησης των ασφαλιστικών απαιτήσεων ως απάτη ή ειλικρινής δήλωση, ή της πρόβλεψης του αριθμού των ανθρώπων που πρόκειται να προσβληθούν από μια ασθένεια.



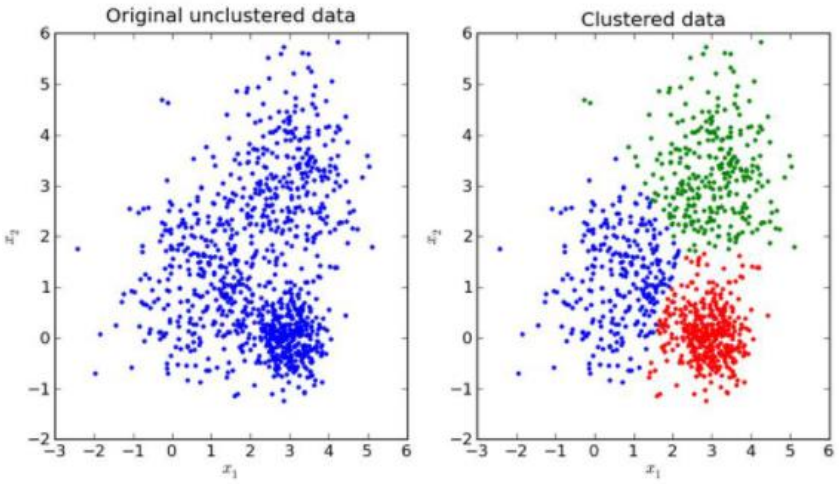

**Σχήμα 3.2**: Διαγραμματική απεικόνιση των προβλημάτων της ταξινόμησης και της παλινδρόμησης στην εποπτευόμενη μηχανική μάθηση

### 3.3.2. Μη Εποπτευόμενη μάθηση (Unsupervised learning)

Η μη εποπτευόμενη μάθηση, είναι μια τεχνική μηχανικής μάθησης η οποία προσπαθεί να ανακαλύψει πρότυπα και συσχετίσεις χωρίς να της παρέχεται η εμπειρία από τη χρήση προηγούμενων δεδομένων. Συγκεκριμένα, οι αλγόριθμοι στη μη εποπτευόμενη μάθηση καλούνται να ανακαλύψουν τη δομή των δεδομένων εισόδου και στη συνέχεια να κατασκευάσουν μοντέλα συσχέτισης βάσει του βαθμού συσχέτισης και των κοινών χαρακτηριστικών των δεδομένων, χωρίς να γνωρίζουν τις επιθυμητές εξόδους.

Η μη εποπτευόμενη μάθηση χρησιμοποιείται στις ακόλουθες περιπτώσεις προβλημάτων:

- σε προβλήματα κατηγοριοποίησης (Cluster analysis), των οποίων οι αλγόριθμοι στοχεύουν στη δημιουργία κλάσεων βάσει των κοινών χαρακτηριστικών του συνόλου δεδομένων, χωρίς να γνωρίζουν τις επιθυμητές κλάσεις που θα δημιουργηθούν
- σε προβλήματα μείωσης των διαστάσεων (Dimensionality reduction), των οποίων οι αλγόριθμοι στοχεύουν στη μείωση των διαστάσεων του συνόλου των δεδομένων, δημιουργώντας παράγοντες που προκύπτουν από σύνολα χαρακτηριστικών που έχουν υψηλές συσχετίσεις μεταξύ τους




**Σχήμα 3.3**: Διαγραμματική απεικόνιση της εφαρμογής της κατηγοριοποίησης σε ένα σύνολο δεδομένων

## 3.4. Αλγόριθμοι μηχανικής μάθησης

Στη βιβλιογραφία υπάρχουν πολλοί αλγόριθμοι μηχανικής μάθησης και κάθε αλγόριθμος έχει δημιουργηθεί με σκοπό να εξυπηρετήσει ένα διαφορετικό τύπου προβλήματος. Αυτό σημαίνει ότι η επιλογή και η απόδοση του εκάστοτε αλγορίθμου εξαρτάται από το σύνολο των δεδομένων που διαθέτουμε, το είδος τους, καθώς και τη φύση του προβλήματος που καλούμαστε να αντιμετωπίσουμε. Παρακάτω παρουσιάζονται οι βασικοί αλγόριθμοι της εποπτευόμενης και μη εποπτευόμενης μηχανικής μάθησης, σύμφωνα με τη βιβλιογραφία.

### 3.4.1. Τεχνικές παλινδρόμησης (Regression methods)

Η παλινδρόμηση είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης που χρησιμοποιείται για την εκτίμηση και αξιολόγηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής (Y) που απαραιτήτως πρέπει να λαμβάνει συνεχείς αριθμητικές τιμές και μίας ή περισσοτέρων ανεξάρτητων μεταβλητών ($X_1$, $X_2$, …., $X_v$). Ο στόχος της παλινδρόμησης δεν είναι μόνο να εκφράσει τη σχέση μεταξύ αυτών των μεταβλητών, αλλά και να προβλέψει τις τιμές τις εξαρτημένης μεταβλητής βάσει των τιμών των ανεξάρτητων μεταβλητών. Γενικά η παλινδρόμηση περιγράφεται με τη μορφή εξίσωσης ως εξής:

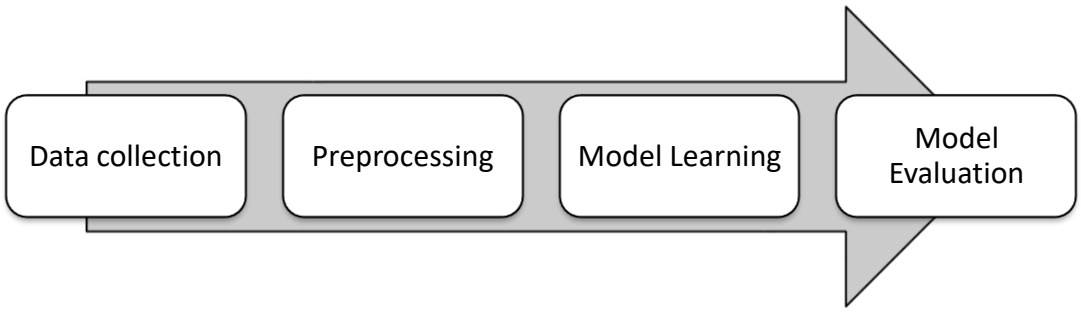
$$y = f(X, \beta) + \varepsilon$$

όπου $y$ είναι η εξαρτημένη μεταβλητής και $X$ είναι το διάνυσμα των ανεξαρτήτων μεταβλητών, $\beta$ είναι το διάνυσμα των άγνωστων παραμέτρων συσχέτισης, $\varepsilon$ το τυχαίο σφάλμα που προκύπτει κατά την πρόβλεψη από την ύπαρξη μη ελεγχόμενων "τυχαίων" παραγόντων και f είναι η συνάρτηση παλινδρόμησης. Για την επιτυχία του μοντέλου είναι απαραίτητη η ελαχιστοποίηση του σφάλματος μεταξύ πραγματικής και προβλεπόμενης τιμής της εξαρτημένης μεταβλητής, όπως περιγράφεται παρακάτω.

$$d = y - f(X, \beta).$$

Η ελαχιστοποίηση των σφαλμάτων είναι μια διαδικασία με περίπλοκους υπολογισμούς και η συνάρτησης παλινδρόμησης που χρησιμοποιείται διαφέρει ανάλογα με το είδος του προβλήματος που πρέπει να αντιμετωπιστεί. Στη μελέτη περιπτώσεων που ακολουθεί σε αυτή την εργασία, οι αλγόριθμοι παλινδρόμησης που χρησιμοποιήθηκαν είναι οι εξής:

- Γραμμική παλινδρόμηση (Linear Regression)
- Παλινδρόμηση ραχοειδής (Ridge)
- Παλινδρόμηση LASSO
- Παλινδρόμηση με Δέντρα Απόφασης (Decision Trees regression)
- Παλινδρόμηση με Τυχαία Δάση (Random Forest regression)
- Παλινδρόμηση Gradient Boosting




**Σχήμα 3.4**: Διαγραμματική απεικόνιση της διαδικασίας που ακολουθείται για τη διενέργεια ενός μοντέλου παλινδρόμησης στη μηχανική μάθηση

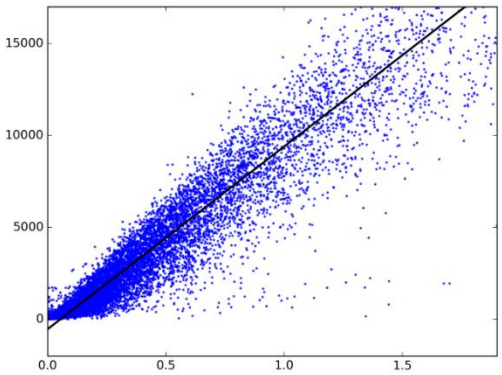
### 3.4.1.1. Γραμμική παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι το πιο γνωστό μοντέλο στην ανάλυση παλινδρόμησης, στην οποία η σχέση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών εκφράζεται μέσω μιας γραμμικής συνάρτησης. Για την επίλυση αυτής της συνάρτησης, ευρέως χρησιμοποιούμενη είναι η μέθοδος των ελαχίστων τετραγώνων, η οποία στοχεύει στην ελαχιστοποίηση του αθροίσματος των τετραγώνων των κατακόρυφων αποστάσεων των σημείων ($x_i$, $y_i$) από την ευθεία $Y_i = a + \beta X_i$ (στην περίπτωση της απλής παλινδρόμησης).
Η εξίσωση που πρέπει να ελαχιστοποιηθεί είναι η: $\sum_{i=1}^{\nu} e_i^2 = \sum_{i=1}^{\nu}(y_i - \alpha \pm \beta x_i)^2$ και οι τιμές α και β που ελαχιστοποιούν την παραπάνω εξίσωση, ονομάζονται αμερόληπτες εκτιμήτριες ελαχίστων τετραγώνων και υπολογίζονται από τις παρακάτω σχέσεις:

$$\hat{\beta} = \frac{\nu \sum_{i=1}^{\nu} x_i y_i - (\sum_{i=1}^{\nu} x_i)(\sum_{i=1}^{\nu} y_i)}{\nu \sum_{i=1}^{\nu} x_i^2 - (\sum_{i=1}^{\nu} x_i)^2}, \qquad \text{όπου } \bar{y} = \frac{1}{\nu}\left(\sum_{i=1}^{\nu} y_i\right), \bar{x} = \frac{1}{\nu}\left(\sum_{i=1}^{\nu} x_i\right)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Στην περίπτωση που υπάρχει μόνο μια ανεξάρτητη μεταβλητή τότε αναφέρεται ως απλή γραμμική παλινδρόμηση (Simple linear regression), ενώ όταν υπάρχει πάνω από μια ανεξάρτητες μεταβλητές, αναφέρεται ως πολλαπλή γραμμική παλινδρόμηση (multiple linear regression).




**Σχήμα 3.5**: Διαγραμματική απεικόνιση της προσαρμογής της ευθείας των ελαχίστων τετραγώνων σε ένα παράδειγμα απλής γραμμικής παλινδρόμησης

Με μια σύντομη αναφορά, οι προϋποθέσεις αξιοπιστίας ενός μοντέλου γραμμικής παλινδρόμησης είναι οι εξής:

- Η εξαρτημένη μεταβλητή $y_i$ να είναι ένας γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών
- Κανονικότητα των σφαλμάτων
- Ομοσκεδαστικότητα των σφαλμάτων
- Ανεξαρτησία των σφαλμάτων
- Απουσία πολυσυγγραμμικότητας (στην περίπτωση της πολλαπλής παλινδρόμησης)

### 3.4.1.2. Παλινδρόμηση κορυφογραμμής (Ridge)

Η παλινδρόμηση Ridge είναι μια τεχνική για την ανάλυση δεδομένων πολλαπλής παλινδρόμησης που πάσχουν από πολυσυγγραμμικότητα, δηλαδή από μεταβλητές οι οποίες παρουσιάζουν υψηλές συσχετίσεις μεταξύ τους. Όταν εμφανίζεται πολυσυγγραμμικότητα, μπορεί οι εκτιμήτριες ελαχίστων τετραγώνων να είναι αμερόληπτες, αλλά οι διασπορές τους είναι μεγάλες, και δε κρίνονται κατάλληλες. Προσθέτοντας έναν βαθμό μεροληψίας στις

εκτιμήσεις της παλινδρόμησης, η παλινδρόμηση κορυφογραμμής μειώνει την διασπορά. Σκοπός είναι να δοθούν πιο αξιόπιστες εκτιμήσεις αποφεύγοντας το overfitting χωρίς όμως να αφαιρεθούν μεταβλητές από το μοντέλο.

Η εξίσωση της παλινδρόμησης Ridge που πρέπει να ελαχιστοποιηθεί περιγράφεται από την παρακάτω εξίσωση:

$$\sum_{i=1}^{\nu} e_i^2 = \sum_{i=1}^{\nu} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=0}^{p} \beta_j^2$$

Ουσιαστικά στην παλινδρόμηση Ridge προστίθεται μια ποινή (penalty) μέσω μιας παραμέτρου που συμβολίζεται με λ, με τον επιπλέον περιορισμό

$$\sum_{j=0}^{p} \beta_j^2 < c, \text{ όπου } c > 0$$

Επομένως, η παλινδρόμηση Ridge, συρρικνώνει τους συντελεστές μειώνοντας έτσι τη διακύμανση των σφαλμάτων και αποφεύγοντας την πολυσυγγραμικότητα.

Όπως φαίνεται και από την εξίσωση της παλινδρόμησης Ridge, όταν το λ→0, τότε η συνάρτηση της είναι όμοια με τη συνάρτηση της γραμμικής παλινδρόμησης.

### 3.4.1.3. Παλινδρόμηση LASSO

Η μέθοδος LASSO (Least Absolute Shrinkage and Selection Operator) που προτάθηκε από τον Tibshirani (1996,1997) είναι επίσης μια εκτίμηση των ποινικοποιημένων ελαχίστων τετραγώνων. Η χρήση της συνίσταται κυρίως σε περιπτώσεις όπου ο αριθμός των επεξηγηματικών μεταβλητών είναι μεγαλύτερος του αριθμού των διαθέσιμων παρατηρήσεων (p > N). Το σημαντικότερο πλεονέκτημα της μεθόδου αυτής, είναι η ικανότητά της να παράγει αραιά μοντέλα, δηλαδή μοντέλα τα οποία περιέχουν μόνο ένα υποσύνολο παραγόντων από τους αρχικά διαθέσιμους. Οι εκτιμήτριες ελαχίστων τετραγώνων παρόλο που είναι αμερόληπτες, έχουν μεγάλη διασπορά. Επομένως, θέτοντας περιορισμό στις παραμέτρους εισάγουμε ένα μικρό ποσό μεροληψίας αλλά επιτυγχάνουμε σημαντική μείωση στη διασπορά τους. Έτσι το μέσο τετραγωνικό σφάλμα των εκτιμητριών, συνήθως μειώνεται, με αποτέλεσμα να οδηγούμαστε σε καλύτερες προβλέψεις για τη μεταβλητή απόκρισης.

Η συνάρτηση απώλειας της LASSO που πρέπει να ελαχιστοποιηθεί περιγράφεται από την παρακάτω εξίσωση (The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie et. Al. (2008), p. 68-69):

$$\sum_{i=1}^{\nu} e_i^2 = \sum_{i=1}^{\nu} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=0}^{p} |\beta_j|$$

Ουσιαστικά στην παλινδρόμηση LASSO προστίθεται μια ποινή μέσω της παραμέτρου συντονισμού λ, όπως και στην περίπτωση της παλινδρόμησης Ridge, η διαφορά όμως έγκειται στο γεγονός ότι όπως αναφέρθηκε και παραπάνω, η παλινδρόμηση LASSO βοηθάει και στην επιλογή των στατιστικά σημαντικών μεταβλητών, βελτιώνοντας έτσι το προγνωστικό μοντέλο και κάνοντας την ερμηνεία του μοντέλου πιο απλή. Η συνάρτηση ποινής υπόκειται στον ακόλουθο περιορισμό

$$\sum_{j=0}^{p} |\beta_j| < t, \text{ όπου } t > 0$$

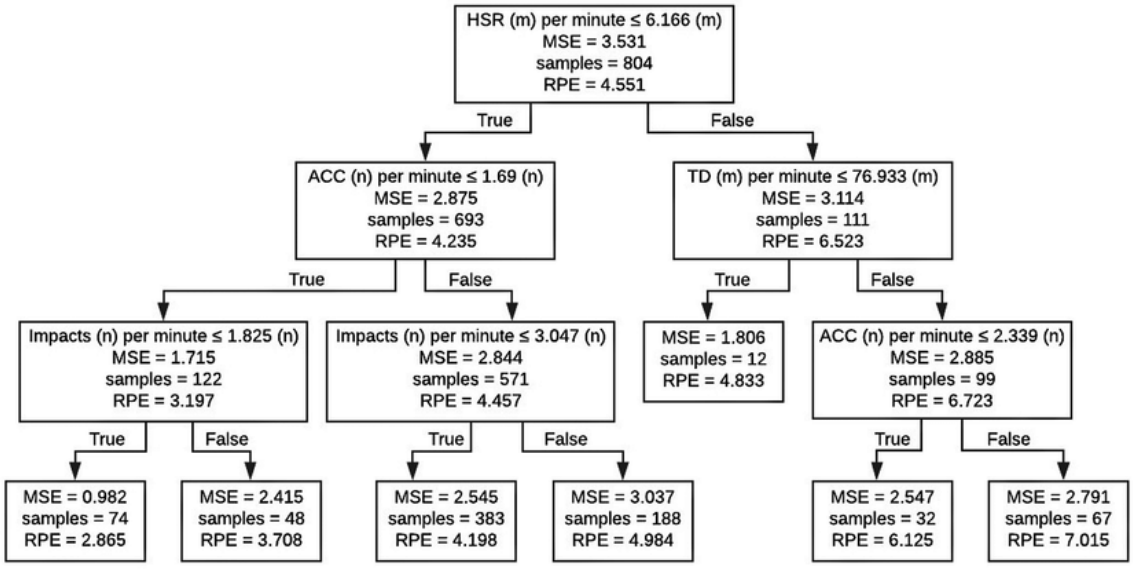
Ένα μειονέκτημα της μεθόδου LASSO, είναι ότι δημιουργεί μεροληπτικές εκτιμήτριες. Αυτό το πρόβλημα επισημάνθηκε πρώτα από τους Fan και Li (2001) και αργότερα από τον Zou (2006). Συγκεκριμένα, ο Zou πρότεινε μια προσαρμοσμένη LASSO, η οποία διαφέρει από την κλασσική ως προς τη συνάρτηση ποινής που χρησιμοποιεί αφού χρησιμοποιεί σταθμισμένη ποινή, η οποία περιγράφεται ενδεικτικά παρακάτω:

$$\sum_{j=1}^{p} w_j|\beta_j| < t, \text{ όπου } \widehat{w_j} = (\frac{1}{|\hat{\beta}_j|})^{\gamma} \text{ με } \gamma \geq 0$$

### 3.4.1.4. Παλινδρόμηση με Δέντρα Απόφασης (Decision Tree regression)

Τα δέντρα αποφάσεων είναι μια μη παραμετρική εποπτευόμενη μέθοδος που χρησιμοποιείται τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Ουσιαστικά, αναπτύσσεται ένα δυαδικό δέντρο απόφασης, όπου σε κάθε κόμβο του (node) εφαρμόζεται ένας έλεγχος που αφορά σε κάθε μεταβλητή ξεχωριστά. Σε κάθε κόμβο, προκύπτει το τετραγωνικό άθροισμα των διαφορών μεταξύ προβλεπόμενης και πραγματικής τιμής. Τα σφάλματα των μεταβλητών συγκρίνονται μεταξύ τους και η μεταβλητή που αποδίδει το χαμηλότερο τετραγωνικό άθροισμα των σφαλμάτων (SSE) επιλέγεται ως ο ριζικός κόμβος διαίρεσης. Αναλόγως το αποτέλεσμα του ελέγχου γίνεται η επιλογή της δεξιά ή αριστερής διακλάδωσης του δέντρου με τη διαδικασία αυτή να είναι επαναλαμβανόμενη και μόλις ολοκληρωθούν όλοι οι έλεγχοι καταλήγουμε στον τελικό κόμβο στον οποίο διενεργείται και η τελική πρόβλεψη.

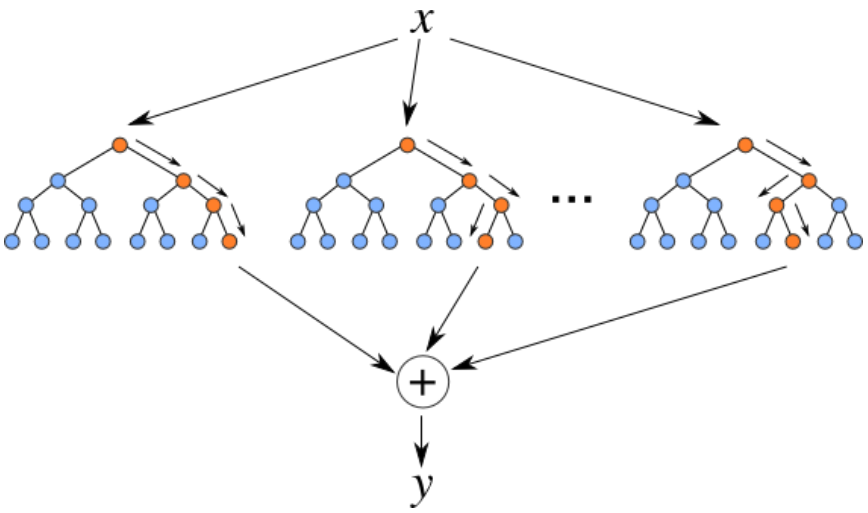
Το παρακάτω διάγραμμα δείχνει τη δομή ενός δέντρου αποφάσεων.




**Σχήμα 3.6**: Διαγραμματική απεικόνιση του δέντρου αποφάσεων για την πρόβλεψη μιας συνεχούς μεταβλητής

### 3.4.1.5. Παλινδρόμηση με Τυχαία Δάση (Random Forest regression)

Ό αλγόριθμος παλινδρόμησης τυχαίων δασών (Random Forest Regression) αποτελεί μια συλλογή δέντρων αποφάσεων αρκετών δέντρων απόφασης και στην συνέχεια υπολογίζει κατά μέσο όρο τα αποτελέσματα. Το τυχαίο δάσος σε αντίθεση με τα δέντρα αποφάσεων δημιουργεί τυχαία υποσύνολα όλων των χαρακτηριστικών και τα οποία χρησιμοποιούνται τυχαία σε κάθε κόμβο μέχρι να κατασκευαστεί η τελική δομή του δέντρου. Στην συνέχεια συνδυάζει τα υποδέντρα αποφάσεων με την τεχνική του bagging (όρος που προτάθηκε από τον Leo Breiman), δηλαδή η μεταβλητή που συμμετέχει πιο πολλές φορές στη λήψη αποφάσεων επιλέγεται ως η μεταβλητή απόφασης μειώνοντας έτσι σημαντικά την διακύμανση και αποφεύγοντας το φαινόμενος της υπερπροσαρμογής (overfitting) του μοντέλου (The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie et. Al. (2008), p. 283-288). Το παρακάτω διάγραμμα δείχνει τη δομή ενός τυχαίου δάσους.

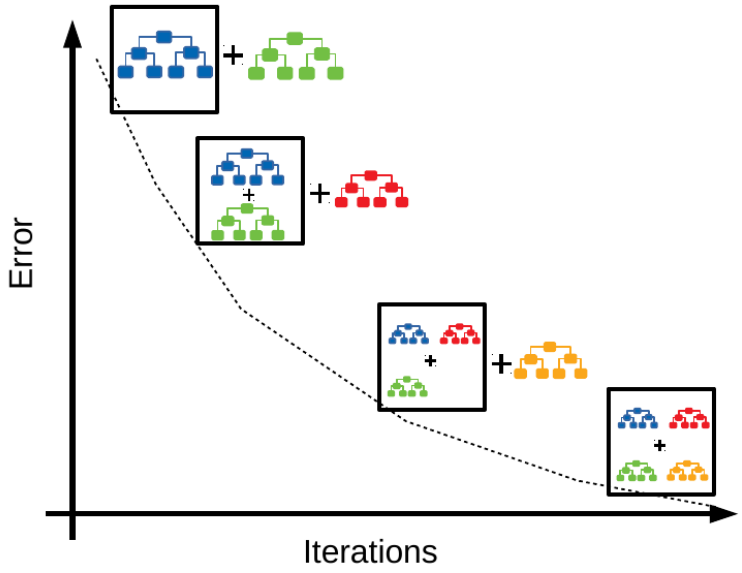



**Σχήμα 3.7**: Διαγραμματική απεικόνιση ενός τυχαίου δάσους σε προβλήματα παλινδρόμησης

### 3.4.1.6. Παλινδρόμηση με Gradient Boosting

Η τεχνική της ενίσχυσης της κλίσης (Gradient Boosting) είναι μια ισχυρή τεχνική συνδυασμού μοντέλων μηχανικής μάθησης τα οποία λειτουργούν ως αδύναμα μοντέλα «μαθητές» και τα οποία χρησιμοποιούν παραμετρικά βάρη κατά την εκπαίδευσή τους με στόχο να συνδυαστούν για την κατασκευή ενός ισχυρότερου μοντέλου. Το όνομά της το πήρε από τον αλγόριθμο Καθόδου της κλίσης (Gradient Descent) ο οποίος αποτελεί μια τεχνική βελτιστοποίησης της συνάρτησης ζημίας και την τεχνική Boosting, μια μέθοδος βασισμένη σε δέντρα αποφάσεων, στην οποία κάθε δέντρο εκπαιδεύεται χρησιμοποιώντας την πληροφορία από τα προηγούμενα δέντρα (The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie et. Al. (2008), p. 283-288). Αρχικά, χρησιμοποιώντας ένα υποσύνολο δεδομένων δημιουργείται ένα μοντέλο που βασίζεται σε δέντρα αποφάσεων και με βάση αυτό το μοντέλο διενεργούνται προβλέψεις για το σύνολο των δεδομένων εκπαίδευσης. Στη συνέχεια δημιουργείται ένα νέο μοντέλο που λαμβάνει υπόψιν τα σφάλματα που δημιουργήθηκαν στο προηγούμενο μοντέλο με σκοπό να τα διορθώσει. Αυτή η διαδικασία λειτουργεί επαναληπτικά και ουσιαστικά κάθε αδύναμο μοντέλο που εκπαιδεύεται αξιοποιεί τα σφάλματα του προηγούμενου με σκοπό τη δημιουργία του τελικού

μοντέλου το οποίο προκύπτει όταν τα υπολειπόμενά σφάλματα δε μπορούν να διορθωθούν ή όταν επιτευχθεί το μέγιστο όριο του αριθμού των μοντέλων.




**Σχήμα 3.8**: Διαγραμματική απεικόνιση της τεχνικής Gradient Boosting

### 3.4.2. Τεχνικές ταξινόμησης (Classification methods)

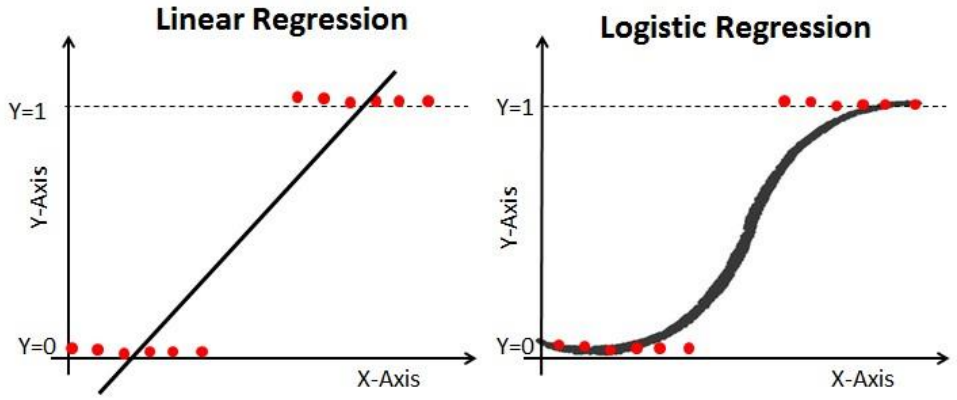
Οι τεχνικές εποπτευόμενης ταξινόμησης είναι από τις πιο δημοφιλείς τεχνικές στην μηχανική μάθηση. Η εποπτευόμενη ταξινόμηση είναι η διαδικασία αναγνώρισης και ομαδοποίησης των δεδομένων σε προκαθορισμένες κατηγορίες – κλάσεις. Απαραίτητη προϋπόθεση για τη διενέργεια της ταξινόμησης είναι να προϋπάρχει στο σύνολο δεδομένων εκπαίδευσης η ετικέτα της κλάσης που ανήκει κάθε σύνολο χαρακτηριστικών, με σκοπό τη δημιουργία προγνωστικών μοντέλων για την εκχώρηση ετικετών κλάσεων σε νέα δεδομένα.

Για παράδειγμα, στην περίπτωση που μια ασφαλιστική εταιρεία θέλει να προβλέψει την ασφαλιστική απάτη για μελλοντικές αξιώσεις, θα χρειαστεί να συλλέξει δεδομένα παρελθοντικών ασφαλιστικών απαιτήσεων που να είναι ήδη κατηγοριοποιημένες ως απάτη ή ειλικρινής δήλωση. Στη συνέχεια θα χρησιμοποιήσει αυτό το σύνολο δεδομένων για τη διερεύνηση τυχόν προτύπων και συσχετίσεων μεταξύ των χαρακτηριστικών με τη χρήση μοντέλων μηχανικής μάθησης, τα οποία στη συνέχεια θα μπορούν να κατηγοριοποιούν τις νέες ασφαλιστικές απαιτήσεις με μία εύλογη ακρίβεια σε μια από τις δύο κλάσεις. Στις παρακάτω υποενότητες ακολουθεί μια σύντομη ανάλυση σε κάποιες από τις πιο ευρέως διαδεδομένες τεχνικές ταξινόμησης στην εποπτευόμενη μηχανική μάθηση.

#### 3.4.2.1. Λογιστική παλινδρόμηση (Logistic regression)

Η λογιστική παλινδρόμηση ερευνά το μη γραμμικό αποτέλεσμα μιας εξαρτημένης κατηγορικής μεταβλητής αναφορικά με τη δράση πολλών ανεξάρτητων μεταβλητών και ουσιαστικά, αποτελεί ένα μοντέλο ταξινόμησης των τιμών της εξαρτημένης μεταβλητής βάσει της θεωρία των πιθανοτήτων. Η βασική διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης είναι η φύση της επιλεγμένης μεταβλητής απόκρισης. Στη γραμμική

παλινδρόμηση η μεταβλητή απόκρισης είναι αποκλειστικά ποσοτική μεταβλητή, ενώ στη λογιστική παλινδρόμηση είναι είτε διχοτομική, είτε κατηγορική (ονομαστική ή διατακτική). Επίσης, στη γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο των ελαχίστων τετραγώνων, ενώ στη λογιστική παλινδρόμηση γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας, δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων προκειμένου να οδηγήσουν στα βέλτιστα παρατηρούμενα αποτελέσματα.




**Σχήμα 3.9**: Διαγραμματική απεικόνιση της διαφοράς μεταξύ Γραμμικής και Λογιστικής Παλινδρόμησης

Το μοντέλο της λογιστικής παλινδρόμησης με δίτιμα δεδομένα αποτελείται από τη μεταβλητή απόκρισης (Y), της οποίας το σύνολο τιμών είναι το {0,1} και από ένα σύνολο ερμηνευτικών μεταβλητών $X_i, i = 1,2, \dots, k$.

Η δίτιμη παλινδρόμηση έχει τη μορφή:

$$f(z) = \frac{e^z}{1 + e^z},$$

Όπου f(z) είναι η πιθανότητα ενός συγκεκριμένου αποτελέσματος το οποίο προκύπτει από την επίδραση της ομάδας των ανεξάρτητων μεταβλητών που εκπροσωπεί η μεταβλητή z.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots + \ \beta_k x_k$$

Λογαριθμίζοντας το μοντέλο παλινδρόμησης ώστε οι συντελεστές παλινδρόμησης να παίρνουν τιμές μέσα στο διάστημα τιμών [0,1], προκύπτει η εξής εξίσωση του μοντέλου με k ερμηνευτικές μεταβλητές:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots + \ \beta_k x_k,$$

όπου το p δηλώνει κάποια πιθανότητα και το πηλίκο $\frac{p}{1-p}$ αντιστοιχεί στην πιθανότητα επιτυχημένης έκβασης (odds).

Οι εκτιμήσεις των συντελεστών της παλινδρόμησης, υπολογίζονται με τη χρήση της μεθόδου της μέγιστης πιθανοφάνειας της οποίας η εξίσωση παρατίθεται στη λογαριθμική της μορφή παρακάτω:

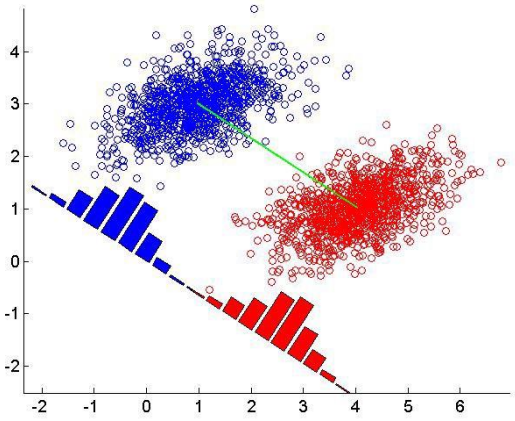
$$L = \prod_{i=1}^{n} log_e f(x_i \theta)$$

### 3.4.2.2. Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis)

Η Γραμμική Διακριτική Ανάλυση (LDA) είναι μια μέθοδος μετασχηματισμού των δεδομένων που ανήκουν σε συγκεκριμένες κλάσεις και χρησιμοποιείται στη μηχανική μάθηση και την τεχνητή νοημοσύνη με σκοπό τον καλύτερο διαχωρισμό των κλάσεων. Για την εφαρμογή της Γραμμικής Διακριτικής Ανάλυσης τα δεδομένα είναι απαραίτητο να είναι αριθμητικά με συνεχείς τιμές και να ανήκουν σε δύο ή περισσότερες γνωστές εκ των προτέρων κλάσεις.

Τα χαρακτηριστικά της LDA είναι ότι:

1. Μετασχηματίζει τα δεδομένα με τέτοιο τρόπο ώστε να μεγιστοποιηθεί η απόσταση μεταξύ των κλάσεων
2. Ελαχιστοποιεί τη διασπορά εντός των κλάσεων με σκοπό τα δεδομένα κάθε κλάσης να είναι συγκεντρωμένα γύρω από τη μέση τιμή τους
3. Μειώνει τις διαστάσεις των δεδομένων με σκοπό τη βελτιστοποίηση της ταξινόμησης




**Σχήμα 3.10**: Γραφική απεικόνιση της Γραμμικής Διακριτικής Ανάλυσης

Αρχικά η διασπορά μεταξύ των κλάσεων υπολογίζεται από την απόσταση μεταξύ των μέσων τιμών των κλάσεων ως εξής:

$$S_b = \sum_{i=1}^{k} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

Στη συνέχεια η διασπορά εντός των κλάσεων υπολογίζεται από το άθροισμα των πινάκων διασπορών κάθε κλάσης (j=1,2…,N) ως εξής:

$$S_w = \sum_{i=1}^{k} \sum_{j=1}^{N} (\bar{x}_{ij} - \bar{x}_i)(\bar{x}_{ij} - \bar{x})^T$$

Και τέλος, με το κριτήριο του Fisher επιτυγχάνεται η μείωση των διαστάσεων των δεδομένων που μεγιστοποιεί τη διακύμανση μεταξύ των κλάσεων και ελαχιστοποιεί τη διακύμανση εντός των κλάσεων. Συγκεκριμένα, πρέπει να βρεθεί ο κατάλληλος πίνακας μετασχηματισμού (P), ώστε να μεγιστοποιηθεί το κριτήριο του Fisher, το οποίο παρατίθεται παρακάτω:

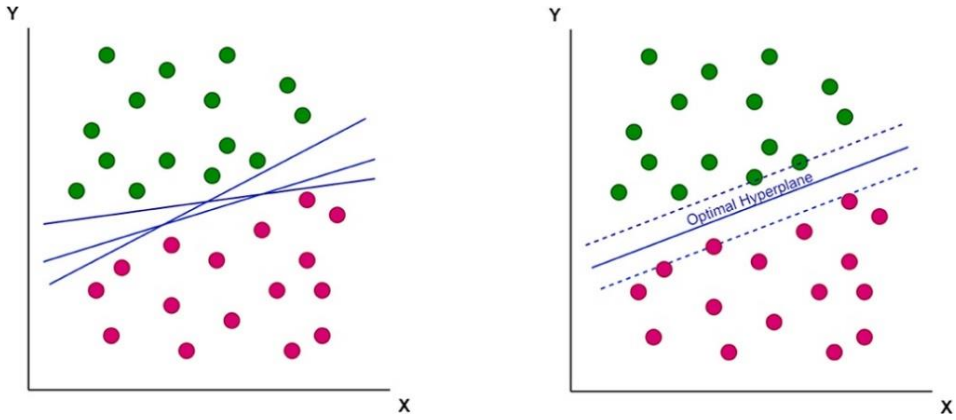
$$J(V) = \frac{V^T S_b V}{V^T S_W V}$$

### 3.4.2.3. Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Οι Μηχανές Διανυσματικής Υποστήριξης (SVM) είναι μία οικογένεια αλγορίθμων εποπτευόμενης μάθησης που αναπτύχθηκαν από τον Vladimir Vapnik και χρησιμοποιείται τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης. Τα SVM είναι μια από τις πιο ισχυρές τεχνικές για γραμμικές ή μη μεθόδους ταξινόμησης, η οποία δε βασίζεται στη θεωρία των πιθανοτήτων, αλλά στη θεωρία της βελτιστοποίησης.

Στόχος του SVM είναι να επιλέξει το βέλτιστο διαχωριστικό υπερεπίπεδο, το οποίο να μεγιστοποιεί το περιθώριο μεταξύ των σημείων δεδομένων και των δυο κατηγοριών. Η μεγιστοποίηση του περιθωρίου παρέχει κάποια ενίσχυση έτσι ώστε τα μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη ακρίβεια σε θετικά (+1) και αρνητικά (-1) παραδείγματα (Εικόνα 12).

Σε περίπτωση προβλήματος 2 διαστάσεων, ο διαχωρισμός των δεδομένων επιτυγχάνεται με τη βέλτιστη διαχωριστική ευθεία, ενώ για παραπάνω διαστάσεις επιτυγχάνεται με το βέλτιστο υπερεπίπεδο (Σχήμα 3.11).

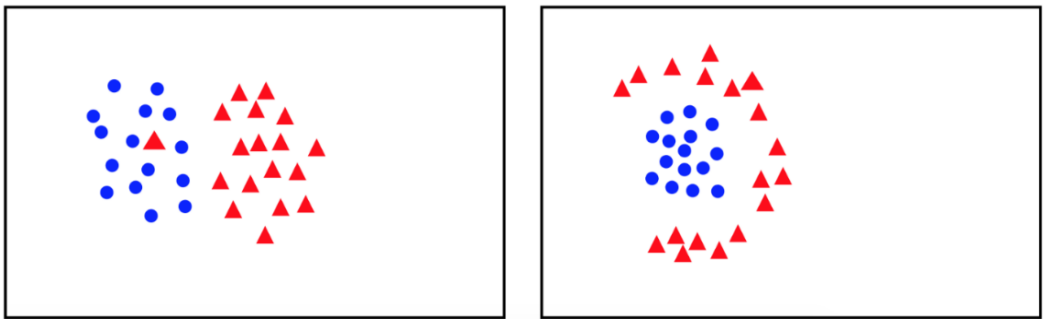



**Σχήμα 3.11**: Γραφική απεικόνιση της επιλογής του μέγιστου περιθωρίου μεταξύ των σημείων δεδομένων και των δυο κατηγοριών

Το πλεονέκτημα του αλγορίθμου SVM είναι ότι διαχειρίζεται πολύ καλά μεγάλο πλήθος χαρακτηριστικών και παρουσιάζει υψηλή απόδοση κατά την κατηγοριοποίηση αντικειμένων μεταξύ δύο κατηγοριών. Επίσης, είναι ικανός να κατασκευάζει μοντέλα αρκετά πολύπλοκα για να επιλύει δύσκολα προβλήματα του πραγματικού κόσμου.

Τα SVM μπορούν να αντιμετωπίσουν 3 ειδών προβλήματα:

1. Γραμμικά πλήρως διαχωρίσιμες κλάσεις όπως φαίνεται στις παραπάνω γραφικές απεικονίσεις (Σχήμα 3.11)
2. Μη απόλυτα γραμμικά διαχωρίσιμες κλάσεις, όπου δε μπορεί να βρεθεί μια ευθεία ή υπερεπίπεδο που να χωρίζει πλήρως τα δεδομένα σε κλάσεις και ορισμένα σημεία βρίσκονται στη λάθος κλάση (Σχήμα 3.12)
3. Γραμμικά μη διαχωρίσιμες κλάσεις, όπου τα σημεία δε μπορούν να διαχωριστούν με μια ευθεία ή υπερεπίπεδο (Σχήμα 3.12)



**Σχήμα 3.12**: Γραφική απεικόνιση προβλημάτων με μη απόλυτα γραμμικά διαχωρίσιμες κλάσεις (Εικόνα στα αριστερά) και με μη γραμμικά διαχωρίσιμες κλάσεις (Εικόνα στα δεξιά)

Σε περίπτωση που οι κλάσεις ενός σετ δεδομένων, δεν είναι γραμμικά διαχωρίσιμες, τότε είναι απαραίτητο να γίνει προβολή των δεδομένων σε τρεις διαστάσεις. Αυτό επιτυγχάνεται με τη χρήση των συναρτήσεων πυρήνα (kernel functions) στο σύνολο εκπαίδευσης, ώστε να μετασχηματίσουν τον αρχικό χώρο υποθέσεων και να βρουν τη βέλτιστη μη γραμμική υπερεπιφάνεια που μεγιστοποιεί την απόσταση μεταξύ των κλάσεων και ελαχιστοποιεί το σφάλμα ταξινόμησης.

Το μοντέλο SVM με δίτιμα δεδομένα αποτελείται από τη μεταβλητή απόκρισης (Y), της οποίας το σύνολο τιμών είναι το {-1,1} και από ένα σύνολο ερμηνευτικών μεταβλητών $X_1, X_2, \ldots, X_k$. Η συνάρτηση απώλειας που πρέπει να ελαχιστοποιηθεί είναι η εξής:

$$min_w(\frac{1}{2})\sum_{i=1}^{n} w_i^2 + C\sum_{j=1}^{m} max(0, 1 - k(x_i, x_j)y_j)$$

Όπου ο πρώτος όρος είναι του αριθμού τον χαρακτηριστικών (n) και ο δεύτερος όρος αθροίζει των αριθμό των δειγμάτων στα δεδομένα (m). Το $w$ εκφράζει τα βάρη των ερμηνευτικών μεταβλητών, το C τη σταθερά κανονικοποίησης και η $k(x_i, x_j)$ εκφράζει τη συνάρτηση πυρήνα (kernel function), της οποίας η μαθηματική έκφραση διαφοροποιείται ανάλογα με τον τύπο του προβλήματος (Γραμμικά, μη πλήρως γραμμικά και μη γραμμικά διαχωρίσιμες κλάσεις). Οι τέσσερις βασικοί τύποι συναρτήσεων πυρήνα είναι οι εξής:

| Linear kernel | Sigmoid kernel | Polynomial kernel | Radial Basis Function |
|---|---|---|---|
| $k(x_i, x_j) = w^T x_j$ | $k(x_i, x_j) = tanh(\gamma w^T x_j + r)$* | $k(x_i, x_j) = (\gamma w^T x_j + r)^d$, $\gamma$>0* | $k(x_i, x_j) = \exp\left(\gamma \left\|x_i - x_j\right\|^2\right)$, $\gamma > 0$ * |

* γ, r, d είναι παράμετροι των kernels

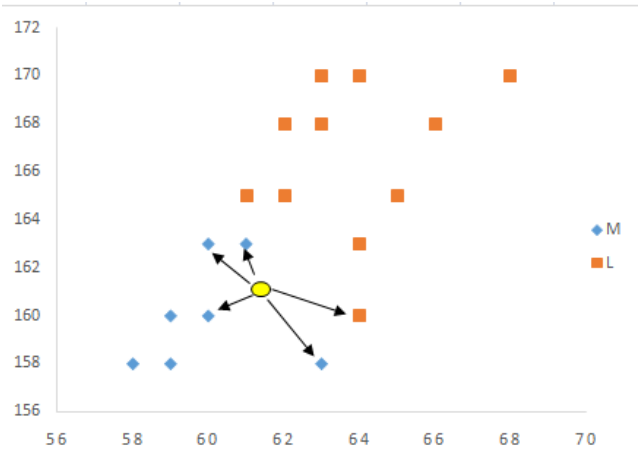
### 3.4.2.4. Κ κοντινότεροι γείτονες (K Nearest Neighbors)

Ο αλγόριθμος Κ κοντινότεροι γείτονες (KNN) είναι μια πολύ διαδεδομένη τεχνική εποπτευόμενης ταξινόμησης που στηρίζεται στη χρήση μέτρων βασισμένων στην απόσταση.

Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων με n ανεξάρτητες μεταβλητές $X_1, X_2 \ldots, X_n$ και την εξαρτημένη μεταβλητή η οποία αποτελείται από 2 κλάσεις. Κάθε σημείο μπορεί να θεωρηθεί ως ένα σημείο στο χώρο των n διαστάσεων, επομένως η απόσταση μεταξύ 2 σημείων X και Y στο χώρο ισούται με d(X,Y). Η απόσταση d(X,Y) υπολογίζεται σύμφωνα με την Ευκλείδεια απόσταση από την παρακάτω εξίσωση:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

όπου $x_i, y_i$ είναι οι τιμές των X, Y για i=1,2,….,n διαστάσεις. Σύμφωνα με τον αλγόριθμο, ο χρήστης προκαθορίζει την τιμή της σταθεράς k και αναζητά στο n-διάστατο χώρο κ παρατηρήσεις που βρίσκονται πλησιέστερα στη νέα παρατήρηση. Στη συνέχεια ο ταξινομητής εκχωρεί τη νέα παρατήρηση στην κλάση που βρίσκονται πλησιέστερα μεταξύ των κ πλησιέστερων γειτόνων.




**Σχήμα 3.13**: KNN Classification

Ένα μειονέκτημα της παραπάνω εξίσωσης της Ευκλείδειας απόστασης είναι ότι για τον υπολογισμό της ομοιότητας μεταξύ των παρατηρήσεων, προϋποθέτει ότι υπάρχει ισότιμη συμμετοχή όλων των διαστάσεων, κάτι το οποίο δεν ισχύει στην πλειοψηφία των περιπτώσεων, καθώς οι μεταβλητές με μεγάλο εύρος τιμών επηρεάζουν περισσότερο το αποτέλεσμα σε σχέση με τις μεταβλητές με μικρότερο εύρος τιμών. Γι αυτό το λόγο, δημιουργήθηκε μια παραλλαγή της Ευκλείδειας απόστασης, η Σταθμισμένη Ευκλείδεια απόσταση στην οποία υπεισέρχονται και πολλαπλασιαστικοί όροι. Αυτό σημαίνει ότι σε κάθε διάσταση αντιστοιχίζεται και ένα βάρος αναλόγως με το εύρος τιμών του.

Η απόσταση d(X,Y) που υπολογίζεται από την Σταθμισμένη Ευκλείδεια απόσταση περιγράφεται από την παρακάτω εξίσωση:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} w_i(x_i - y_i)^2}$$

όπου w είναι το βάρος που αντιστοιχεί στην i-οστή διάσταση.

Υπάρχουν και άλλες συναρτήσεις απόστασης, μερικές από τις οποίες παρατίθενται παρακάτω

- **Απόσταση Mahalanobis**
  $d(x,y)\sqrt{(x-y)^T s^{-1}(x-y)}$, όπου s είναι ένας τετραγωνικός πίνακας N×N
- **Απόσταση Manhattan**
  $$d(x,y) = \sum_{i=1}^{m}|x_i - y_i|$$

- **Απόσταση Minkowski**

$$d(x,y) = (\sum_{i=1}^{m} |x_i - y_i|^r)^{1/r}$$

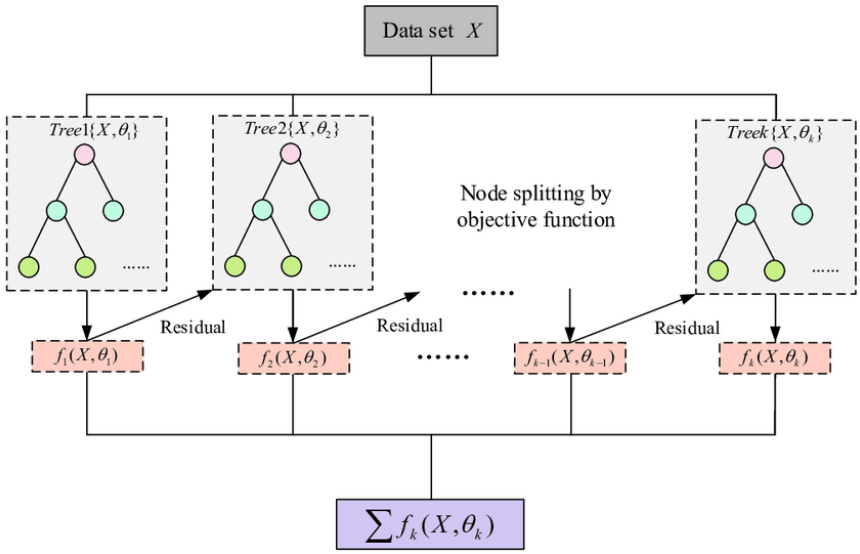
- **Απόσταση Chebychev**

$$d(x,y) = \max_{i=1,2,..,m} |x_i - y_i|$$

### 3.4.2.5. Extreme Gradient Boosting (XGBoost)

Ο αλγόριθμος XGBoost είναι ένας αλγόριθμος εποπτευόμενης μηχανικής μάθησης, που χρησιμοποιεί τα δέντρα αποφάσεων και την τεχνική της ενίσχυσης κλίσης (Gradient Boosting).

Όπως εξηγήθηκε και σε προηγούμενη ενότητα, η τεχνική της ενίσχυσης κλίσης (Gradient Boosting) είναι μια μέθοδος βασισμένη σε δέντρα αποφάσεων, στην οποία κάθε δέντρο εκπαιδεύεται χρησιμοποιώντας την πληροφορία από τα προηγούμενα δέντρα. Συγκεκριμένα, όπως φαίνεται και στην εικόνα 15, κάθε δέντρο απόφασης λειτουργεί σαν «αδύναμος μαθητής», εκπαιδεύεται δηλαδή λαμβάνοντας υπόψιν τα σφάλματα του προηγούμενου δενδροειδές μοντέλου και στη συνέχεια προσπαθεί να τα διορθώσει. Αυτή η διαδικασία λειτουργεί επαναληπτικά έως ότου τα υπολειπόμενά σφάλματα δε μπορούν να διορθωθούν ή όταν επιτευχθεί το μέγιστο όριο του αριθμού των μοντέλων. Απώτερος σκοπός φυσικά, είναι η βελτιστοποίηση της συνάρτησης απώλειας (Ain Shams Engineering Journal, 12, 2021, p. 1545-1556).




**Σχήμα 3.14**: XGBoost Classification

Ο λόγος που ο XGBoost προτιμάται σε σχέση με άλλες υλοποιήσεις ενίσχυσης δέντρων είναι λόγω της υπολογιστικής του ταχύτητας και της απόδοσής του. Συγκεκριμένα:

- ❖ Αποτελείται από ένα σύνολο υπερ-παραμέτρων που μπορούν να συντονιστούν και να βελτιστοποιήσουν την απόδοση του μοντέλου
- ❖ Έχει ενσωματωμένη δυνατότητα χειρισμού ελλειπουσών τιμών
- ❖ Εφαρμόζει Distributed Computing για την εκπαίδευση πολύ μεγάλων μοντέλων
- ❖ Εφαρμόζει βελτιστοποίηση κρυφής μνήμης (Cache Optimization)

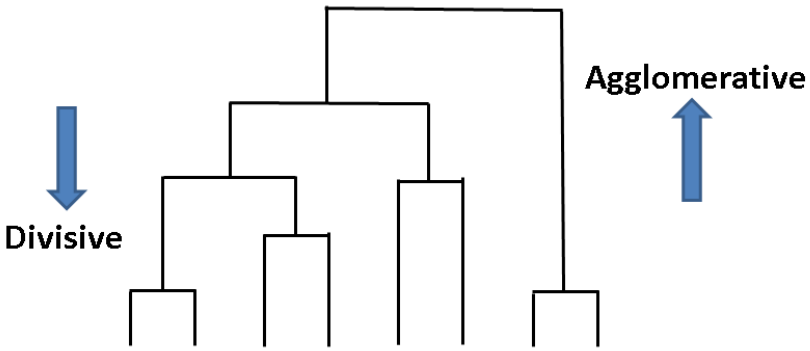
- ❖ Παρέχει Out-of-Core Computing, όπου μεγάλα σετ δεδομένων που είναι αδύνατον να χωρέσουν στη κύρια μνήμη, αποθηκεύονται σε άλλους χώρους

### 3.4.3. Τεχνικές μη εποπτευόμενης μάθησης

Οι μη εποπτευόμενοι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται όταν δεν υπάρχει πρότερη γνώση για την ταξινόμηση των αντικειμένων στο σύνολο δεδομένων εκπαίδευσης και με διάφορές τεχνικές προσπαθούν να ανακαλύψουν κρυφές δομές δεδομένων. Οι πιο διαδεδομένες τεχνικές είναι:

**1) Η ομαδοποίηση κατά συστάδες (Clustering)**, στην οποία ο στόχος είναι να βρεθούν ομοιογενείς υποομάδες στα δεδομένα. Η ομαδοποίηση βασίζεται στον βαθμό ομοιότητας των χαρακτηριστικών και σκοπός είναι οι ομάδες να είναι διακριτές μεταξύ τους και η διακύμανση εντός των ομάδων να είναι μικρή ώστε να προκύψει και καλύτερη ερμηνεία. Υπάρχουν διάφοροι αλγόριθμοι συσταδοποίησης οι οποίοι χωρίζονται στις εξής βασικές κατηγορίες:

a) **Οι Ιεραρχικοί αλγόριθμοι (Hierarchical algorithms)** στους οποίους οι συστάδες δημιουργούνται σε επίπεδα και κάθε επίπεδο αντιπροσωπεύει ένα σύνολο από συστάδες. Οι ιεραρχικοί αλγόριθμοι χωρίζονται με τη σειρά τους σε δύο κατηγορίες οι οποίες έχουν διαφορετικές προσεγγίσεις στα προβλήματα κατηγοριοποίησης:
   - Οι **Συσσωρευτικοί αλγόριθμοι (Agglomerative algorithms)**, στους οποίους αρχικά κάθε στοιχείο είναι μία συστάδα και στη συνέχεια οι συστάδες συγχωνεύονται επαναληπτικά έως ότου δημιουργηθούν οι τελικές συστάδες
   - Οι **Διαιρετικοί αλγόριθμοι (Divisive algorithms)**, στους οποίους αρχικά όλα τα στοιχεία είναι σε μία συστάδα και στη συνέχεια οι συστάδες διαιρούνται έως ότου δημιουργηθούν οι τελικές συστάδες

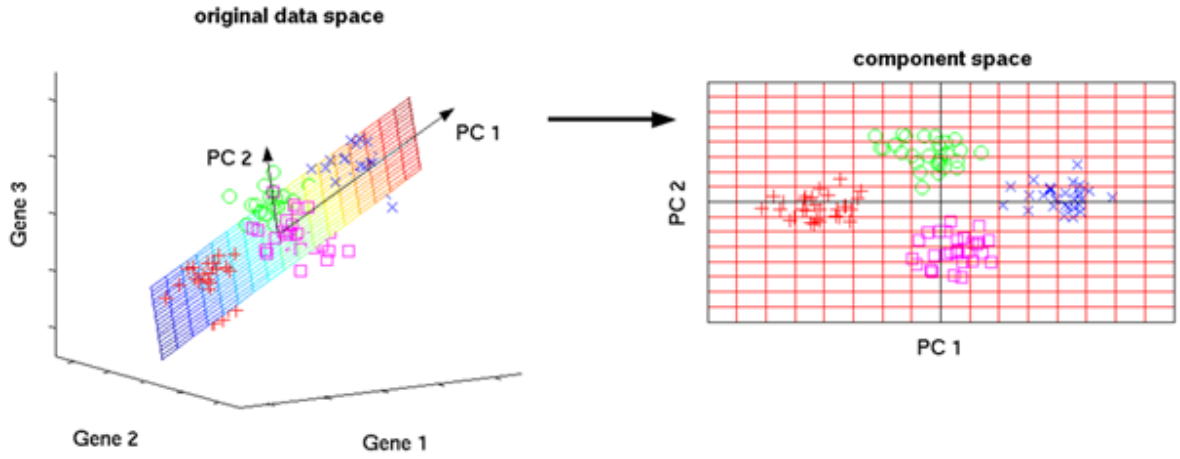



**Σχήμα 3.15**. Ιεραρχική Συσταδοποίηση

b) **Οι Διαμεριστικοί αλγόριθμοι (Partitioning algorithms),** στους οποίους η συσταδοποίηση δημιουργείται σε ένα μόνο βήμα, με τον χρήστη να εισάγει τον αριθμό συστάδων που επιθυμεί
c) **Γενετικοί αλγόριθμοι (Genetic algorithms),** των οποίων η λειτουργία είναι εμπνευσμένη από τη βιολογία και είναι χρήσιμοι σε προβλήματα πολλών διαστάσεων

Το μειονέκτημα της τεχνικής της ομαδοποίησης κατά συστάδες, είναι ότι πάντα θα προκύπτουν ομάδες στα δεδομένα ακόμα και αν ουσιαστικά δεν υπάρχει καμία, επομένως είναι πολύ σημαντικό τα αποτελέσματα των αλγορίθμων να αξιολογούνται κριτικά και να επικυρώνονται χρησιμοποιώντας τυχόν υπάρχουσα γνώση που υπάρχει στον εκάστοτε κλάδο που εφάπτεται η ανάλυση.

**2) Η μείωση διαστάσεων (Dimensionality Reduction)**, η οποία χρησιμοποιείται σε περιπτώσεις όπου ο αριθμός των μεταβλητών που καλείται να διαχειριστεί ο αναλυτής, είναι μεγάλος και είναι απαραίτητη η σύνοψή τους σε λιγότερες διαστάσεις (παράγοντες), οι οποίες πρέπει να έχουν υψηλή εσωτερική συνέπεια για να είναι αξιόπιστες. Λειτουργεί δηλαδή σαν μέθοδος προεπεξεργασίας των δεδομένων πριν την εφαρμογή συνήθως μοντέλων εποπτευόμενης μάθησης. Επίσης, χρησιμοποιείται σε περιπτώσεις που η μείωση των διαστάσεων είναι απαραίτητη για λόγους διευκόλυνσης της οπτικοποίησης των δεδομένων.




**Σχήμα 3.16**. PCA για μείωση διαστάσεων

Στην παρούσα εργασία, δε χρησιμοποιήθηκαν μέθοδοι μη εποπτευόμενης μηχανικής μάθησης, ωστόσο παρακάτω θα ακολουθήσει μια σύντομη παρουσίαση σε έναν αλγόριθμο για κάθε μια από τις παραπάνω τεχνικές. Συγκεκριμένα, θα γίνει η επεξήγηση του αλγορίθμου συσταδοποίησης «K-Means» και της μεθόδους Ανάλυσης Κύριων Συνιστωσών (Principal Components Analysis), οι οποίες είναι από τις ευρέως διαδεδομένες τεχνικές της μη εποπτευόμενης μάθησης.

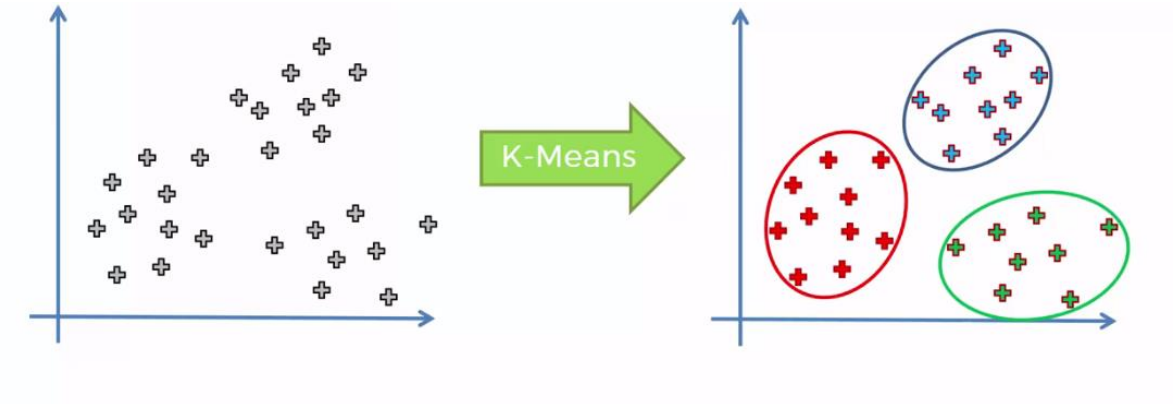
### 3.4.3.1. Αλγόριθμος K-means

Ο αλγόριθμος k-means είναι γενικά η πιο γνωστή και ευρέως χρησιμοποιημένη μέθοδος ομαδοποίησης. Ανήκει στην κατηγορία των διαμεριστικών αλγορίθμων, δηλαδή για την δημιουργία συστάδων ο χρήστης πρέπει να δηλώσει εξαρχής τον επιθυμητό αριθμό συστάδων.

Συγκεκριμένα, ο αλγόριθμος k-means δέχεται ως είσοδο ένα σύνολο προτύπων $x_1, x_2, \dots, x_n$ και τον επιθυμητό αριθμό συστάδων (k) και ξεκινάει μια επαναληπτική διαδικασία, όπου γίνεται η ανάθεση των k τυχαίων σημείων $(K_1, K_2, \dots, K_n)$, τα οποία ονομάζονται κεντροειδή και δηλώνουν το κέντρο βάρους της συστάδας. Για κάθε πρότυπο $x_i$, υπολογίζεται το κοντινότερο κέντρο $K_j$ $(\text{argmin}_j D(x_i, K_j)$ και έπειτα με χρήση κάποιου μέτρου απόστασης (π.χ. Ευκλείδεια απόσταση, Απόσταση Manhattan κ.τ.λ.) αντιστοιχίζεται στη συστάδα $K_j$. Στη συνέχεια, επανυπολογίζονται τα γεωμετρικά κέντρα για κάθε συστάδα $K_1, K_2, \dots, K_n$ βάσει του μέσου όρου της κάθε συστάδας από όλα τα σημεία $x_i$ που προέκυψαν από το προηγούμενο βήμα.

$$K_j = \frac{1}{n_j} \sum_{x_i \to K_J} x_i$$

Η διαδικασία σταματάει όταν τα κεντροειδή των συστάδων μετατοπίζονται ελάχιστα, με αποτέλεσμα να μην προκύπτει πλέον καμία μεταβολή στις συστάδες, δηλαδή όταν όλα τα στοιχεία τους παραμένουν ίδια.




**Σχήμα 3.17**. Αλγόριθμος K-means

### 3.4.3.2. Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis)

Η τεχνική της ανάλυσης κύριων συνιστωσών (PCA), αποτελεί την απλούστερη και πλέον διαδεδομένη πολυμεταβλητή ανάλυση και στοχεύει στην ανεύρεση από ένα πλήθος p μεταβλητών ορισμένων νέων και συνήθως λιγότερων σε πλήθος μεταβλητών, με τέτοιο τρόπο ώστε να είναι αντιπροσωπευτικές και ισχυρά επεξηγηματικές των αρχικών, οι οποίες έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα να μη συσχετίζονται μεταξύ τους. Το πλεονέκτημα της PCA είναι ότι λόγω της συγκεκριμένης τεχνικής οι νέες μειωμένες μεταβλητές, εξηγούν ένα πολύ μεγάλο ποσοστό της ολικής μεταβλητότητας που αναπτύσσεται μεταξύ των αρχικών p μεταβλητών, επομένως το μεγαλύτερο μέρος της πληροφορίας ερμηνεύεται πλέον από τις νέες μειωμένες διαστάσεις. Η διαδικασία της ανάλυσης είναι η εξής:

Αρχικά γίνεται τυποποίηση των αρχικών μεταβλητών $X_1, X_2, \dots, X_p$, ώστε να έχουν μέση τιμή 0 και διακύμανση 1

$$(X_\iota - \bar{X})/s$$

από τις οποίες στη συνέχεια δημιουργούνται p συνδυασμοί $Z_1, Z_2, \dots, Z_p$ με τέτοιο τρόπο ώστε να υπάρχει απουσία συσχετισμού μεταξύ τους και κάθε διάσταση να μετράει διαφορετικές διαστάσεις των στοιχείων. Κάθε συνιστώσα $Z_i$ προκύπτει από τον γραμμικό συνδυασμό p μεταβλητών,

$$Z_i = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} X_i$$

όπου $a_{ij}$ ο ειδικός συντελεστής στάθμισης της j μεταβλητής στην i συνιστώσα, με τον περιορισμό ότι για κάθε συνιστώσα i το άθροισμα των τετραγώνων των ειδικών συντελεστών στάθμισης είναι ίσο με το 0. $\sum_{j=1}^{p} {a_{ij}}^2 = 1$, για κάθε i, όπου i=1,2,…,p

Με την παραπάνω διαδικασία, δημιουργούνται p συνιστώσες i, δηλαδή όσες είναι και οι αρχικές μεταβλητές.

Οι ειδικοί συντελεστές στάθμισης $a_{ij}$ υπολογίζονται με τη βοήθεια του πίνακα C των συνδιακυμάνσεων των αρχικών μεταβλητών,

$$C = \begin{pmatrix} c_{11} & c_{12} \dots & c_{1p} \\ c_{21} & c_{22} \dots & c_{2p} \\ c_{p1} & c_{p2} \dots & c_{pp} \end{pmatrix}$$

Όπου τα διαγώνια στοιχεία $c_{ii}$ είναι οι διακυμάνσεις της $X_i$ και $c_{ij}$ οι συνδιακυμάνσεις των μεταβλητών $X_i$ και $X_j$.

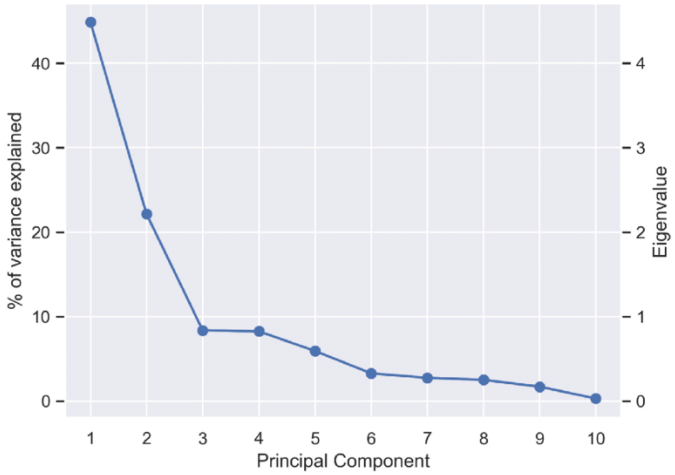
Μετά την τυποποίηση των αρχικών μεταβλητών, ο πίνακας των συνδιακυμάνσεων μετατρέπεται σε πίνακα συσχετίσεων ως εξής,

$$C = \begin{pmatrix} 1 & c_{12} & \dots & c_{1p} \\ c_{21} & 1 & \dots & c_{2p} \\ c_{p1} & c_{p2} & \dots & 1 \end{pmatrix}$$

Επομένως, $c_{ii} = 1$ και $c_{ij}$ είναι οι συσχετίσεις μεταξύ των μεταβλητών $X_i$ και $X_j$.

Οι διακυμάνσεις των κύριων συνιστωσών είναι ίσες με τις ιδιοτιμές $\lambda_i$ (eigenvalues). Στη συνέχεια επιλέγονται οι συνιστώσες που ερμηνεύουν το μεγαλύτερο ποσοστό της ολικής μεταβλητότητας και απορρίπτονται εκείνες που εξηγούν μικρό ποσοστό.

Για παράδειγμα, όπως φαίνεται και στο παρακάτω Scree plot (Σχήμα 3.18), από το σύνολο των 10 διαστάσεων, οι πρώτες τρεις (3) είναι αυτές που εξηγούν περίπου το 80% της μεταβλητότητας του μοντέλου ενώ οι υπόλοιπες επτά (7) διαστάσεις εξηγούν μόνο το υπόλοιπο 20%. Επομένως, οι τελικές διαστάσεις που θα επιλεχθούν ως αντιπροσωπευτικές για το σύνολο των δεδομένων, θα είναι οι πρώτες τρεις (3).




**Σχήμα 3.18**: Scree plot για τη μεταβλητότητα που εξηγείται από τις συνιστώσες

## 3.5. Τεχνικές αξιολόγησης των μοντέλων (Model evaluation technics)

Προκειμένου να αξιολογηθεί η επίδοση των μοντέλων μηχανικής μάθησης και να επιλεχθεί το βέλτιστο μοντέλο ως λύση για το εκάστοτε πρόβλημα, χρησιμοποιούνται συγκεκριμένες τεχνικές.

### 3.5.1. Τεχνικές αξιολόγησης των μοντέλων παλινδρόμησης

Στην **Παλινδρόμηση** χρησιμοποιούνται οι εξής μετρικές για την αξιολόγηση των μοντέλων:

**1) Ο συντελεστής προσδιορισμού ($R^2$)**: Ο συντελεστής προσδιορισμού παίρνει τιμές στο διάστημα τιμών [0,1] και μετράει το ποσοστό της συνολικής μεταβλητότητας που εξηγείται από τις επεξηγηματικές μεταβλητές ($X_1$, $X_2, \ldots, X_n$) του μοντέλου με εξαρτημένη τη μεταβλητή (Y). Για παράδειγμα, εάν ο συντελεστής προσδιορισμού ισούται με 0.85, τότε οι επεξηγηματικές μεταβλητές του μοντέλου ερμηνεύουν το 85% της συνολικής μεταβλητότητας του μοντέλου με εξαρτημένη τη μεταβλητή (Y) και ένα 15 % παραμένει ανερμήνευτο. Ο μαθηματικός τύπος του συντελεστή προσδιορισμού είναι ο εξής:

$$R^2 = 1 - \frac{SSR}{SST}$$

όπου $SSR = \sum_{i=1}^{n}(y_i - \hat{y})^2$ το άθροισμα τετραγώνων των σφαλμάτων και $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ το συνολικό άθροισμα τετραγώνων.

Βέβαια, ο Mordecai Ezekiel πρότεινε μια διόρθωση στον συντελεστή προσδιορισμού:

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{df_{SSR}}}{\frac{SST}{df_{SST}}} = \frac{\frac{SSR}{n-p}}{\frac{SST}{n-1}}$$

Όπου $df_{SST} = n - 1$ οι βαθμοί ελευθερίας του συνολικού αθροίσματος τετραγώνων και $df_{SSR} = n - p$, οι βαθμοί ελευθερίας του αθροίσματος τετραγώνων των σφαλμάτων. Ο προσαρμοσμένος συντελεστής προσδιορισμού ($R^2\,adjusted$) είναι κατάλληλος σε περιπτώσεις που υπάρχουν πολλές επεξηγηματικές μεταβλητές και σε περιπτώσεις σύγκρισης μοντέλων παλινδρόμησης μεταξύ τους. Αυξάνεται όταν προστίθεται στο μοντέλο μια στατιστικά σημαντική μεταβλητή που έχει καλό ποσοστό ερμηνείας και σε αντίθεση με τον $R^2$ μειώνεται όταν η διακύμανση που ερμηνεύει η νέα μεταβλητή δεν είναι μεγαλύτερη από αυτή που θα μπορούσε κανείς να παρατηρήσει τυχαία.

**2) Το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE):** Το μέσο απόλυτο σφάλμα υπολογίζει το μέσο όρο του μεγέθους του σφάλματος, που ορίζεται ως η απόλυτη τιμή της διαφοράς μεταξύ πραγματικής και προβλεπόμενης τιμής της εξαρτημένης μεταβλητής. Δίνεται από τον τύπο:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

**3) Το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE):** Το μέσο τετραγωνικό σφάλμα υπολογίζει το σφάλμα ως το μέσο όρο του αθροίσματος των τετραγώνων των σφαλμάτων και δίνεται από τον τύπο:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

Σε αντίθεση με το μέσο απόλυτο σφάλμα (MAE) που δίνει την ίδια βαρύτητα σε όλα τα κατάλοιπα, το μέσο τετραγωνικό σφάλμα (MSE) χρησιμοποιείται όταν επιθυμείται να τιμωρηθούν οι μεγάλες τιμές καταλοίπων και οι ακραίες τιμές.

**4) Η Τετραγωνική ρίζα του Μέσου Τετραγωνικό Σφάλμα (Root Mean Squared Error - RMSE):** Η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος υπολογίζει το σφάλμα ως την τετραγωνική ρίζα του μέσου όρου του αθροίσματος των τετραγώνων των σφαλμάτων και δίνεται από τον τύπο:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Έχει καλύτερη ερμηνευτική ικανότητα από το μέσο τετραγωνικό σφάλμα, καθώς το σφάλμα εκφράζεται σε μονάδες της ανεξάρτητης μεταβλητής.

Όλες οι παραπάνω μετρικές των σφαλμάτων, (εκτός του συντελεστή προσδιορισμού ($R^2$) και του προσαρμοσμένου συντελεστή προσδιορισμού ($R^2\ adjusted$)), παίρνουν τιμές στο διάστημα τιμών [0,∞) και φυσικά όσο πιο μικρές είναι οι τιμές των σφαλμάτων τόσο καλύτερη είναι η προγνωστική ικανότητα του μοντέλου.

### 3.5.2. Τεχνικές αξιολόγησης των μοντέλων ταξινόμησης

Στην **Ταξινόμηση** χρησιμοποιούνται άλλες μετρικές για την αξιολόγηση των μοντέλων.

**Πίνακας 3.1:** Τύποι απόφασης ενός ταξινομητή

| | Πρόβλεψη | |
|---|---|---|
| **Πραγματικότητα** | **Θετικό** | **Αρνητικό** |
| **Θετικό** | Ορθά Θετικό (True Positive) | Εσφαλμένα Αρνητικό (False Negative) |
| **Αρνητικό** | Εσφαλμένα Θετικό (False Positive) | Ορθά Αρνητικό (True Negative) |

Τα βασικά μεγέθη που αξιοποιούνται από μετρικές για την αξιολόγηση των μοντέλων παρουσιάζονται στον πίνακα 1, στον οποίο η ετικέτα «Θετικό» σημαίνει ότι μια παρατήρηση ταξινομείται στην ετικέτα στόχο και η ετικέτα «Αρνητικό» όταν δεν ταξινομείται στην ετικέτα στόχο. Παραδείγματος χάριν, σε ένα μοντέλο ταξινόμησης για την πρόβλεψη των ασφαλιστικών αξιώσεων που είναι ύποπτες για ασφαλιστική απάτη, η ετικέτα «Θετικό» δηλώνει ύπαρξη ασφαλιστικής απάτης, ενώ η ετικέτα «Αρνητικό» την ύπαρξη ειλικρινούς δήλωσης  Συγκεκριμένα παρατηρούμε τα εξής μεγέθη:

**Ορθή Αποδοχή (True Positive – TP)**
Είναι το πλήθος των παρατηρήσεων που ο ταξινομητής ορθά ταξινόμησε ως θετικά.

**Εσφαλμένη Αποδοχή (False Positive – FP)**
Είναι το πλήθος των παρατηρήσεων που ο ταξινομητής εσφαλμένα ταξινόμησε ως θετικά.

**Ορθή Απόρριψη (True Negative – TN)**
Είναι το πλήθος των παρατηρήσεων που ο ταξινομητής ορθά ταξινόμησε ως αρνητικά.

**Εσφαλμένη Απόρριψη (False Negative – FN)**
Είναι το πλήθος των παρατηρήσεων που ο ταξινομητής εσφαλμένα ταξινόμησε ως αρνητικά.

Φυσικά, ο σκοπός σε κάθε μοντέλο ταξινόμησης είναι η ελαχιστοποίηση των εσφαλμένων αποδοχών και των εσφαλμένων απορρίψεων.

Ακολουθούν οι μετρικές που αξιοποιούν τα παραπάνω μεγέθη για να μετρήσουν την απόδοση του ταξινομητή.

**1) Ορθότητα (Accuracy)**: Είναι το ποσοστό των ορθά ταξινομημένων περιπτώσεων και είναι παραπλανητικό σε περιπτώσεις που τα μεγέθη των κλάσεων δεν είναι ισοπληθή. Υπολογίζεται ως εξής:

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

**2)Ακρίβεια (Precision)**: Είναι το ποσοστό των περιπτώσεων που ταξινομήθηκαν ως θετικά και στην πραγματικότητα είναι θετικά.

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

**3) Ανάκληση (Recall)**: Προσδιορίζει το ποσοστό των δηλώσεων που είναι θετικά και ταξινομήθηκαν ορθά ως θετικά.

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

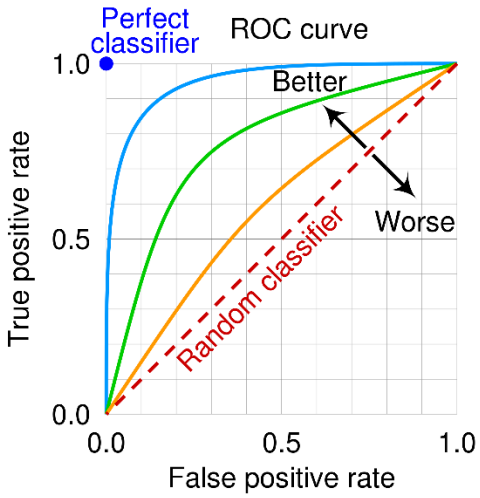
**4) Ειδικότητα (Specificity):** Προσδιορίζει το ποσοστό των δηλώσεων που είναι αρνητικά και ταξινομήθηκαν ως αρνητικά

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

**5) F-Score**: Είναι η μετρική του αρμονικού μέσου όρου της ανάκλησης (Recall) και της ακρίβειας (Precision).

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

**5) Καμπύλη ROC (ROC Curve)**: Η καμπύλη ROC δείχνει την αντιστάθμιση μεταξύ ευαισθησίας (TPR) και ειδικότητας (1-FPR). Οι ταξινομητές που δίνουν καμπύλες πιο κοντά στην πάνω αριστερή γωνία του γραφήματος εμφανίζουν καλύτερη απόδοση. Ένας τυχαίος ταξινομητής αναμένεται να δώσει σημεία που βρίσκονται κατά μήκος της διαγώνιου η οποία ορίζεται και ως η βασική γραμμή (Εικόνα 20). Επομένως όσο πιο μεγάλο είναι το εμβαδό κάτω από την καμπύλη (Area under the curve -AUC) τόσο καλύτερη απόδοση εμφανίζει το μοντέλο.




**Σχήμα 3.19**: ROC Curve

### 3.5.3. Διασταυρούμενη επικύρωση (Cross Validaton)

Στην παρούσα εργασία χρησιμοποιήθηκε η Διασταυρούμενη Επικύρωση 10 τμημάτων (K-fold Cross Validation) και η Στρωματοποιημένη Διασταυρούμενη Επικύρωση 10 τμημάτων (Stratified K-fold Cross Validation) οι οποίες είναι διαδικασίες επαναδειγματοληψίας.

Στη Διασταυρούμενη Επικύρωση 10 τμημάτων υπάρχει μια ενιαία παράμετρος που ονομάζεται k και αναφέρεται στον αριθμό των υποσυνόλων στα οποία πρόκειται να χωριστεί ένα σύνολο δεδομένων, γι αυτό ονομάζεται και k-fold Cross Validation. Η επιλογή της τιμής του k εξαρτάται από το μέγεθος του δείγματος και είναι πολύ σημαντική, καθώς μια λάθος επιλογή της τιμής του k μπορεί να οδηγήσει σε μια εσφαλμένη αντιπροσωπευτική ιδέα για την προγνωστική ικανότητα του μοντέλου. Για την αξιολόγηση των μοντέλων της παρούσας εργασίας διενεργήθηκε διασταυρούμενη επικύρωση για k=10, κάτι που σημαίνει ότι η διαδικασία της διασταυρούμενης επικύρωσης διενεργήθηκε δέκα (10) φορές. Είναι σημαντικό να τονιστεί ότι σε αυτή τη διαδικασία κάθε παρατήρηση του συνόλου δεδομένων εκχωρείται σε μόνο ένα υποσύνολο δεδομένων και από τα k υποσύνολα ένα (1) χρησιμοποιείται ως σετ επικύρωσης και τα υπόλοιπα k-1 ως σετ εκπαίδευσης. Η Διασταυρούμενη Επικύρωση 10 τμημάτων (K-fold Cross Validation) είναι πολύ διαδεδομένη καθώς είναι απλή στην κατανόηση και οδηγεί σε περισσότερο αμερόληπτες εκτιμήσεις σε σχέση με άλλες μεθόδους, όπως η απλή διασταυρούμενη επικύρωση στην οποία ο διαχωρισμός του συνόλου δεδομένων πραγματοποιείται μόνο μια φορά.

Η Στρωματοποιημένη Διασταυρούμενη Επικύρωση 10 τμημάτων, είναι μια άλλη εκδοχή της μεθόδου, στην οποία κάθε υποσύνολο περιέχει περίπου ίσο αριθμό παρατηρήσεων για την κάθε κλάση.

Και οι δύο είναι πολύ χρήσιμες τεχνικές στο στάδιο της επιλογής του κατάλληλου μοντέλου για ανάλυση. Υπάρχουν διάφορες παραλλαγές της διασταυρούμενης επικύρωσης όπως είναι η Διασταυρούμενη επικύρωση Leave-one-out (Leave-one-out Cross validation) και η μέθοδος bootstrap.

## 3.6. Προεπεξεργασία δεδομένων (Data Preprocessing)

Η προεπεξεργασία των δεδομένων είναι μια διαδικασία προετοιμασίας των ακατέργαστων δεδομένων και αποτελεί κρίσιμο βήμα πριν την εφαρμογή ενός μοντέλου μηχανικής μάθησης. Η εφαρμογή της είναι πολύ σημαντική και τις περισσότερες φορές

αποτελεί το πιο χρονοβόρο μέρος μιας ανάλυσης, καθώς η πλειοψηφία των δεδομένων που λαμβάνονται από τον πραγματικό κόσμο είναι χρειάζονται μορφοποίηση, καθαρισμό και μετασχηματισμό για να είναι επεξεργάσιμα και έτοιμα προς εκμετάλλευση. Επιπλέον, η κατάλληλη προεπεξεργασία των δεδομένων μπορεί να αυξήσει την αποτελεσματικότητα και την απόδοση των μοντέλων μηχανικής μάθησης.

Τα πιο συνήθη προβλήματα που πρέπει να αντιμετωπιστούν κατά την προεπεξεργασία των δεδομένων είναι η κωδικοποίηση των κατηγορικών χαρακτηριστικών, ο καθαρισμός των δεδομένων που περιέχουν θορύβους οι οποίοι μπορεί να είναι είτε λάθη είτε ακραίες τιμές, η διαχείριση των ελλειπουσών τιμών και ο μετασχηματισμός των δεδομένων για να βρίσκονται όλα στη ίδια κλίμακα. Ωστόσο, η διαχείριση και η προεπεξεργασία των δεδομένων μπορεί να διαφέρουν ανά αλγόριθμό. Ακολούθως, αναφέρονται συνοπτικά κάποιες από τις μεθόδους:

### 3.6.1. Κωδικοποίηση των κατηγορικών μεταβλητών (Label encoding)

Δεδομένου ότι τα μοντέλα μηχανικής μάθησης στην πλειοψηφία τους υποδέχονται μόνο αριθμητικές τιμές, οι ετικέτες των κατηγορικών μεταβλητών είναι απαραίτητο να κωδικοποιηθούν πριν χρησιμοποιηθούν στη διαδικασία της μοντελοποίησης. Επίσης, σε περίπτωση χρήση κάποιων αλγορίθμων και κυρίως σε προβλήματα παλινδρόμησης, είναι απαραίτητο να παραχθούν ψευδομεταβλητές από τις κατηγορικές μεταβλητές που έχουν πάνω από 2 κατηγορίες.

### 3.6.2. Διαχείριση των ελλειπουσών τιμών (Missing value management)

Η διαχείριση των ελλειπουσών τιμών είναι πολύ σημαντική διαδικασία πριν την εφαρμογή ενός μοντέλου μηχανικής μάθησης. Δεν υπάρχει κάποια συγκεκριμένη διαδικασία που πρέπει να ακολουθηθεί, καθώς η διαχείρισή τους διαφέρει ανάλογα με το είδος των μεταβλητών, τον αλγόριθμο μηχανικής μάθησης που θα χρησιμοποιηθεί και τη φύση του προβλήματος. Ωστόσο, υπάρχουν κάποιες συνήθεις πρακτικές για τη διαχείρισή τους οι οποίες παρουσιάζονται συνοπτικά παρακάτω.

- Σε περίπτωση που το ποσοστό των missing values είναι εξαιρετικά μικρό σε σχέση με το μέγεθος του δείγματος τότε οι παρατηρήσεις αυτές μπορούν να διαγραφούν. Ωστόσο αυτό δε συνίσταται, ειδικά σε περιπτώσεις που το μέγεθος του δείγματος είναι μικρό, καθώς μπορεί να χαθεί πολύ σημαντική πληροφορία από το σύνολο δεδομένων
- Σε περίπτωση που το ποσοστό των missing values σε μία μεταβλητή είναι άνω του 40%, είναι δύσκολο να βγει κάποια αξιόπιστη πληροφορία από αυτή και η αντικατάσταση των τιμών είναι πολύ δύσκολη υπόθεση. Από τη βιβλιογραφία προτείνεται η αφαίρεση της μεταβλητής, ωστόσο σε περίπτωση που η μεταβλητή κρίνεται από τη φύση της πολύ σημαντική για την στατιστική ανάλυση, τότε μπορούν να εφαρμοστούν αλγόριθμοι μηχανική μάθησης για την συμπλήρωση των κενών.
- Η αντικατάσταση των τιμών μπορεί να γίνει με τη μέση τιμή ή την διάμεση τιμή της κατανομής τους στην περίπτωση των ποσοτικών μεταβλητών και η αντικατάσταση με τη διάμεσο ή την επικρατούσα τιμή στην περίπτωση των κατηγορικών μεταβλητών.

### 3.6.3. Εντοπισμός των ακραίων τιμών (Outlier detection)

Οι ακραίες τιμές γενικά ορίζονται ως οι παρατηρήσεις που απέχουν εξαιρετικά από την κύρια ροή των δεδομένων και μπορεί να προκληθούν είτε τυχαία είτε από σφάλμα μέτρησης. Δεν υπάρχου τυποποιημένες μέθοδοι για την ανίχνευση των ακραίων τιμών, καθώς η ανίχνευσή τους εξαρτάται σε μεγάλο βαθμό από τη φύση των δεδομένων. Ωστόσο, υπάρχουν κάποιες παραμετρικές και μη παραμετρικές τεχνικές που μπορούν να χρησιμοποιηθούν για την ανίχνευσή τους, όπως το ενδοτεταρτημοριακό εύρος, το Z-score, η εφαρμογή των αλγορίθμων DBSCAN και Isolation Forest και πολλές άλλες.

### 3.6.4. Διαχωρισμός του συνόλου δεδομένων σε σετ εκπαίδευσης και σετ δοκιμής (Data set split into training & test set)

Απαραίτητη προϋπόθεση για να κατασκευαστεί ένα μοντέλο μηχανικής μάθησης, είναι να προηγηθεί κατά την προεπεξεργασία των δεδομένων ο διαχωρισμός του συνόλου δεδομένων σε σετ εκπαίδευσης και σετ δοκιμής. Σκοπός είναι να δημιουργηθεί ένα μοντέλο το οποίο να έχει υψηλή προβλεπτική ακρίβεια στο σύνολο δεδομένων εκπαίδευσης και στη συνέχεια να ελεγχθεί η ακρίβειά του και στο σύνολο δοκιμής, ώστε να διερευνηθεί η αποτελεσματικότητά του σε ένα διαφορετικό σύνολο. Η συνήθεις αναλογίες διαχωρισμού του συνόλου δεδομένων είναι (70% training set- 30% test set) και (80% training set- 20% test set).

### 3.6.5. Μέθοδοι επαναδειγματοληψίας (Resampling methods)

Πολλές φορές τα σύνολα δεδομένων που λαμβάνονται από τον πραγματικό κόσμο είναι ανεπαρκή και χαρακτηρίζονται από ανομοιογένεια και ασύμμετρες τάξεις κατανομών. Επομένως, αυτό έχει αρνητικό αντίκτυπο στην εξαγωγή επαρκής πληροφορίας από την ανάλυσή τους.

Παραδείγματος χάριν, μια ασφαλιστική εταιρεία χρησιμοποιώντας ιστορικά δεδομένα των πελατών της, θέλει να προβλέψει πότε ένας πελάτης πρόκειται να αποχωρήσει από την εταιρεία ή το ασφαλιστήριο συμβόλαιο. Ωστόσο, το 90% της πληροφορίας αφορά πελάτες που έχουν ανανεώσει το ασφαλιστήριο συμβόλαιο και μόλις το 10% της πληροφορίας αφορά δεδομένα πελατών που αποφάσισαν να το διακόψουν. Κατασκευάζοντας λοιπόν προγνωστικά μοντέλα (customer churn) που βασίζονται στη μηχανική μάθηση, παρατηρείται ότι τα μοντέλα εμφανίζουν αδυναμία στην πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν και αυτό πολλές φορές οφείλεται στη μη επαρκή πληροφορία που υπάρχει γι αυτούς.

Γι αυτό το λόγο έχουν εφευρεθεί κάποιες τεχνικές επαναδειγματοληψίας (resampling techniques) οι οποίες χρησιμοποιούνται από τα μοντέλα μηχανικής μάθησης και επιλέγονται αναλόγως το είδος του προβλήματος . Ενδεικτικά παρατίθενται οι εξής τεχνικές:

- **Τυχαία υπο-δειγματοληψία (Random undersampling):** Οι περιπτώσεις της πλειοψηφικής τάξης της μεταβλητής στόχος διαγράφονται τυχαία από το σύνολο δεδομένων εκπαίδευσης εως ότου επιτευχθεί μια πιο ισορροπημένη κατανομή των τάξεων. Μειονέκτημα αυτής της τεχνικής είναι ότι η τυχαία διαγραφή των τάξεων της πλειοψηφικής τάξης μπορεί να επηρεάσει την επίδοση του ταξινομητή.
- **Τυχαία υπερ-δειγματοληψία (Random oversampling):** Πραγματοποιείται τυχαία επιλογή παραδειγμάτων από τη μειοψηφική τάξη της μεταβλητής στόχος με αντικατάσταση και στη συνέχεια προστίθενται στο σύνολο δεδομένων εκπαίδευσης εως ότου επιτευχθεί μια πιο ισορροπημένη κατανομή των τάξεων. Το πλεονέκτημα της σε σχέση με την τεχνική της τυχαίας υπο-δειγματοληψίας είναι ότι δεν χάνεται κάποια πληροφορία από τα δεδομένα μας, ωστόσο από τη βιβλιογραφία παρατηρείται ότι

συχνά μπορεί να οδηγήσει σε υπερπροσαρμογή των μοντέλων μηχανικής μάθησης, καθώς δημιουργεί ακριβή αντίγραφα των παραδειγμάτων τα οποία δεν αντιστοιχούν σε νέα πραγματικά δεδομένα.

- **Τυχαία υπερ-δειγματοληψίας με συνθετική μειονότητα (Synthetic Minority Oversampling Technique - SMOTE):** Η τεχνική SMOTE έρχεται για να αντιμετωπίσει το πρόβλημα της υπερπροσαρμογής που δημιουργείται στην τεχνική της τυχαίας υπερδειγματοληψίας δημιουργώντας νέα συνθετικά παραδείγματα για την τάξη μειοψηφίας. Συγκεκριμένα, λειτουργεί επιλέγοντας ένα τυχαίο παράδειγμα από την κλάση μειοψηφίας και βρίσκει τους κ πλησιέστερους γείτονες της κλάσης μειοψηφίας. Στη συνέχεια επιλέγεται τυχαία ένας από τους κ πλησιέστερους γείτονες και το συνθετικό στιγμιότυπο δημιουργείται από ένα τυχαίο επιλεγμένο σημείο μεταξύ των δυο παραδειγμάτων στο χώρο των χαρακτηριστικών. Αυτή η διαδικασία συνεχίζεται και για άλλα τυχαία επιλεγμένα παραδείγματα έως ότου επιτευχθεί μια πιο ισορροπημένη κατανομή των τάξεων.

Στην παρούσα εργασία θα χρησιμοποιηθεί η τεχνική της τυχαίας υπερ-δειγματοληψίας με συνθετική μειονότητα (SMOTE).

### 3.6.6. Μετασχηματισμός δεδομένων (Data transformation)

Ο μετασχηματισμός των δεδομένων περιλαμβάνει τα εξής:

- Κανονικοποίηση
  - Κανονικοποίηση των χαρακτηριστικών χρησιμοποιώντας την μέση τιμή και την τυπική απόκλιση της κατανομής τους. Ουσιαστικά γίνεται η τυποποίηση των χαρακτηριστικών, ώστε η κατανομή τους να ακολουθεί την κανονική με μέση τιμή 0 και τυπική απόκλιση 1.

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - Κανονικοποίηση των χαρακτηριστικών στο διάστημα [0,1]. Αυτό σημαίνει ότι η ελάχιστη και η μέγιστη τιμή της κατανομής του χαρακτηριστικού θα είναι 0 και 1 αντίστοιχα. Δε συνίσταται σε περιπτώσεις που η κατανομή των τιμών έχει πολλές ακραίες τιμές. Ο μαθηματικός τύπος είναι:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Δημιουργία χαρακτηριστικών από τον συνδυασμό ήδη υπαρχόντων στο σύνολο δεδομένων

# ΚΕΦΑΛΑΙΟ 4

# 4. Εφαρμογές

Στις παρακάτω υποενότητες γίνεται η παρουσίαση τριών (3) εφαρμογών που χρησιμοποιούν τεχνικές τη στατιστικής μηχανικής μάθησης και της αναλυτικής των δεδομένων για να δώσουν λύσεις σε δύο από τα πιο σημαντικά προβλήματα που κάθε ασφαλιστική εταιρεία καλείται να αντιμετωπίσει.

Η πρώτη εφαρμογή αφορά στη δημιουργία ενός μοντέλου για την πρόβλεψη του ύψους των ιατρικών δαπανών των ασφαλισμένων που καλείται μια ασφαλιστική εταιρία να αποζημιώσει και η δεύτερη εφαρμογή αφορά στη δημιουργία ενός μοντέλου για την πρόβλεψη των πελατών που δύναται να αποχωρήσουν από το ασφαλιστικό συμβόλαιο.

## 4.1. 1η Εφαρμογή - Πρόβλεψη του ύψους των ασφαλιστικών απαιτήσεων

### Πρόβλημα

Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο η ασφαλιστικές απαιτήσεις αποτελούν το σημαντικότερο μέρος του κύκλου ζωής της ασφάλισης. Επομένως, η ακριβής πρόβλεψη των ασφαλιστικών απαιτήσεων που καλείται μια ασφαλιστική εταιρεία να αποζημιώσει αποτελεί θέμα μείζονος σημασίας για τη βιωσιμότητα της, καθώς βάσει αυτών προετοιμάζονται οι ετήσιοι οικονομικοί προϋπολογισμοί και τιμολογούνται τα ασφάλιστρα.

### Σκοπός

Σκοπός της ανάλυσης είναι η δημιουργία ενός προγνωστικού μοντέλου για την πρόβλεψη των ασφαλιστικών απαιτήσεων μιας εταιρείας βάσει ιστορικών συναλλαγών, καθώς και κάποιων δημογραφικών, προσωπικών και ιατρικών τους δεδομένων. Η πρόβλεψη των ασφαλιστικών απαιτήσεων μπορεί να βοηθήσει την εταιρεία να προνοήσει διατηρώντας κάποιο αποθεματικό χρηματικό κεφάλαιο για μελλοντικές αξιώσεις, καθώς και να ανακαλύψει πρότυπα  τα οποία θα οδηγήσουν στην αναπροσαρμογή της τιμολογιακής της πολιτικής στα ασφάλιστρα.

### Περιγραφή του συνόλου δεδομένων

Η ανάλυση αφορά αρχείο αιτήσεων αποζημιώσεων για συμβόλαια ασφάλισης υγείας από ασφαλιστικές εταιρίες των ΗΠΑ, το οποίο αντλήθηκε συνδυαστικά από 2 σύνολα δεδομένων στον ιστότοπο της Kaggle, η οποία είναι μία διαδικτυακή κοινότητα με θεματολογία την επιστήμη των δεδομένων και τη μηχανική μάθηση. Το αρχείο εμπεριέχει δεδομένα από 1338 αιτήσεις αποζημίωσης και τα χαρακτηριστικά που περιλαμβάνονται σε αυτό παρουσιάζονται στον παρακάτω πίνακα 4.1:

**Πίνακας 4.1:** Σύντομη παρουσίαση των χαρακτηριστικών του συνόλου δεδομένων

| ΑΑ | Μεταβλητή | Περιγραφή | Ετικέτες δεδομένων | Είδος μεταβλητής |
|---|---|---|---|---|
| 1 | Patient ID | Μοναδικός κωδικός ασθενή | | Διακριτή |
| 2 | Age | Η ηλικία του ασφαλισμένου | | Συνεχής |
| 3 | Sex | Το φύλο του ασφαλισμένου | • 0=Άντρας<br>• 1=Γυναίκα | Διχοτομική |
| 4 | BMI | Ο δείκτης μάζας σώματος του ασφαλισμένου | | Διατακτική |
| 5 | Steps | Ο μέσος όρος βημάτων ανά ημέρα που κάνει ο ασφαλισμένος | | Συνεχής |
| 6 | Children | Ο αριθμός παιδιών του ασφαλισμένου | | Διακριτή |
| 7 | Bloodpressure | Αρτηριακή πίεση ασφαλισμένου | | Συνεχής |
| 8 | Smoke | Καπνιστική συνήθεια | • 0=Μη καπνιστής<br>• 1=Καπνιστής | Διχοτομική |
| 9 | Region | Η περιοχή κατοικίας του ασφαλισμένου στις Η.Π.Α. | • 0=Βορειοανατολική Αμερική<br>• 1=Βορειοδυτική Αμερική<br>• 2=Νοτιοανατολική Αμερική<br>• 3=Νοτιοδυτική Αμερική | Ονομαστική |
| 10 | Charges | Ποσό ασφαλιστικών απαιτήσεων | | Συνεχής |

## Διερευνητική ανάλυση

Η διερευνητική ανάλυση αποτελεί αναπόσπαστο κομμάτι κάθε ανάλυσης, καθώς είναι ο πρώτος τρόπος κατανόησης των χαρακτηριστικών του συνόλου δεδομένων. Είναι πολύ σημαντική η πλήρης κατανόηση των μεταβλητών, ώστε να γίνει η σωστή διαχείρισή τους, ειδικά όταν στο σύνολο δεδομένων εμπεριέχονται δεδομένα υγείας. Στους παρακάτω πίνακες 4.2 & 4.3 ακολουθεί η περιγραφική ανάλυση των χαρακτηριστικών του συνόλου δεδομένων. Συγκεκριμένα, παρατηρείται ότι από το σύνολο των 1338 ασφαλισμένων το 50,5% (N=662) είναι άντρες και το υπόλοιπο 49,5% (N=676) είναι γυναίκες με τη μέση ηλικία των ασφαλισμένων να είναι ίση 39,2 έτη (Μ.Τ.=39,2, Τ.Α.=14). Όσον αφορά την περιοχή κατοικίας τους, το 24,2% (N=324) διαμένει στη Βορειοανατολική Αμερική, το 24,3% (N=325) στη Βορειοδυτική Αμερική, το 27,2% (N=364) στη Νοτιοανατολική Αμερική και το υπόλοιπο 24,3% (N=324) διαμένει στη Νοτιοδυτική Αμερική. Επίσης, το 42,9% (N=574) των ασφαλισμένων δεν έχει κανένα παιδί στην οικογένειά του, το 53,8% (N=721) έχει από 1 μέχρι 3 παιδιά και το υπόλοιπο 3,3% (N=43) έχει από 4 έως 5 παιδιά. Σχετικά με τα δεδομένα που σχετίζονται με την κατάσταση της υγείας των ασφαλισμένων, παρατηρείται ότι η πλειοψηφία των ασφαλισμένων δεν είναι καπνιστές με ποσοστό 79,5% (N=1064), ενώ μόλις το 20,5% (274) είναι καπνιστές. Επιπλέον, η πλειοψηφία των ασφαλισμένων φαίνεται να παρουσιάζει χαμηλά επίπεδα δραστηριότητας με το 50% εξ αυτών να κάνουν κατά μέσο όρο έως 4007 βήματα καθημερινά και τον μέσο δείκτη μάζας σώματος τους να είναι περίπου ίσος με 30,4 μονάδες, κάτι που δηλώνει ότι η αναλογία βάρους και ύψους είναι πολύ πάνω από τα φυσιολογικά επίπεδα, όπως αυτά ορίζονται από την επιστημονική κοινότητα.

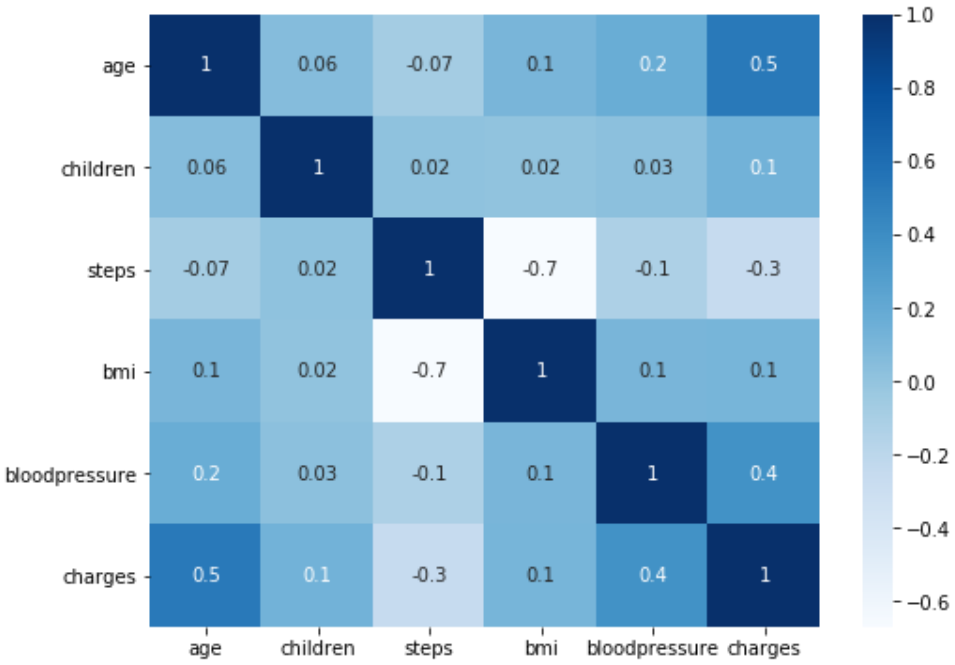
**Πίνακας 4.2:** Περιγραφική παρουσίαση των κατηγορικών χαρακτηριστικών του συνόλου δεδομένων

| | | N | % |
|---|---|---|---|
| **Φύλο** | Άντρας | 662 | 50,5 |
| | Γυναίκα | 676 | 49,5 |
| **Περιοχή κατοικίας** | Βορειοανατολική Αμερική | 324 | 24,2 |
| | Βορειοδυτική Αμερική | 325 | 24,3 |
| | Νοτιοανατολική Αμερική | 364 | 27,2 |
| | Νοτιοδυτική Αμερική | 325 | 24,3 |
| **Αριθμός παιδιών** | 0 | 574 | 42,9 |
| | 1 | 324 | 24,2 |
| | 2 | 240 | 17,9 |
| | 3 | 157 | 11,7 |
| | 4 | 25 | 1,9 |
| | 5 | 18 | 1,4 |
| **Καπνιστής** | Όχι | 1064 | 79,5 |
| | Ναι | 274 | 20,5 |

**Πίνακας 4.3:** Περιγραφική παρουσίαση των ποσοτικών χαρακτηριστικών του συνόλου δεδομένων

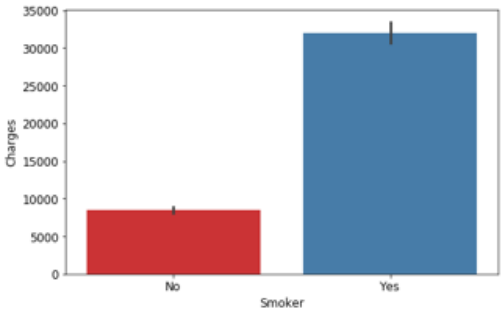
| | **Μ.Τ.** | **Τ.Α.** | **Διάμεσος** | **Ε.Τ.** | **Μ.Τ.** |
|---|---|---|---|---|---|
| **Ηλικία** | 39,2 | 14,0 | 39,0 | 18,0 | 64,0 |
| **Βήματα** | 5.329 | 2.454 | 4.007 | 3.000 | 10.010 |
| **ΔΜΣ** | 30,7 | 6,1 | 30,4 | 16,0 | 53,1 |
| **Αρτηριακή πίεση** | 94,2 | 11,4 | 92,0 | 80,0 | 140,0 |
| **Ποσό ασφαλιστικών απαιτήσεων** | 13.270,4 | 12.110,0 | 9.382,0 | 1.121,9 | 63.770,4 |

Μετά από τη σύντομη περιγραφική ανάλυση των χαρακτηριστικών του συνόλου δεδομένων, είναι σημαντικό να γίνει η διερεύνηση τυχόν συσχετίσεων που παρατηρούνται μεταξύ των χαρακτηριστικών. Σκοπός είναι να βρεθούν τυχόν συσχετίσεις, οι οποίες θα φανούν χρήσιμες κατά τη διαδικασία της μοντελοποίησης. Στο πίνακα 4.4 παρουσιάζονται οι συσχετίσεις μεταξύ των ποσοτικών χαρακτηριστικών του συνόλου δεδομένων με τη χρήση του μη παραμετρικού κριτήριου του Spearman, καθώς τα δεδομένα δεν ακολουθούν την κανονική κατανομή. Συγκεκριμένα, παρατηρείται μια θετική συσχέτιση μέτριας έντασης ($r=0.4$, $p<0.05$) μεταξύ της αρτηριακής πίεσης και του ύψους των ασφαλιστικών απαιτήσεων. Αυτό σημαίνει ότι όσο αυξάνονται οι τιμές τις αρτηριακής πίεσης, το ύψος των ασφαλιστικών απαιτήσεων έχει μια μέτρια τάση προς αύξηση. Αξιόλογο εύρημα είναι και η σχέση μεταξύ της ηλικίας και των ασφαλιστικών απαιτήσεων, στην οποία παρατηρείται ακόμα μια μέτρια θετική συσχέτιση ($r=0.5$, $p<0.05$), με το ύψος των ασφαλιστικών απαιτήσεων να έχει μια μέτρια τάση προς αύξηση, όταν αυξάνεται η ηλικία του ασφαλισμένου. Επίσης, παρατηρήθηκε μια αρνητική συσχέτιση χαμηλής ισχύος, μεταξύ των βημάτων και των ασφαλιστικών απαιτήσεων ($r=-0.3$, $p<0.05$). Τέλος, άλλη μια πολύ σημαντική συσχέτιση η οποία έχει άμεση σχέση με την υγεία του ασφαλισμένου, φαίνεται να είναι αυτή ανάμεσα στα επίπεδα δραστηριότητας και τον ΔΜΣ ($r=-0.7$, $p<0.05$). Συγκεκριμένα, παρατηρείται ότι όσο αυξάνονται τα επίπεδα δραστηριότητας του ασφαλισμένου, ο ΔΜΣ έχει μια πολύ ισχυρή τάση προς μείωση.

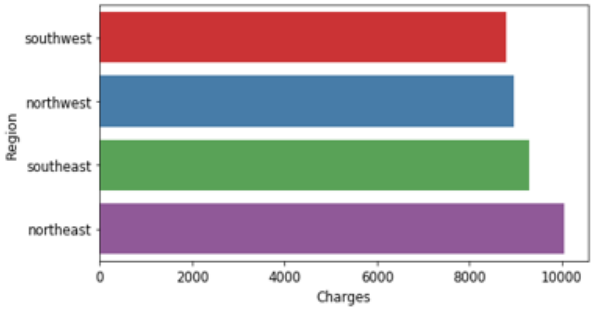



**Πίνακας 4.4**: Correlation matrix

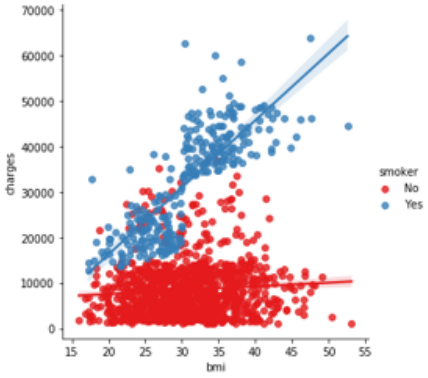
Το ύψος ιατρικών δαπανών δεν ακολουθεί την κανονική κατανομή, επομένως στη διερευνητική ανάλυση χρησιμοποιήθηκε η διάμεσος ως μέτρο κεντρική τάσης. Από τη διερευνητική ανάλυση αξιόλογο εύρημα προέκυψε στη σχέση της καπνιστικής συνήθειας με τις ιατρικές δαπάνες. Συγκεκριμένα, το ενδιάμεσος ύψος ιατρικών δαπανών για τους καπνιστές είναι περίπου ίσο με 34.456$, σχεδόν 3,6 φορές μεγαλύτερο από το αντίστοιχο των μη καπνιστών (δ=7.345$). Επίσης, το ενδιάμεσο ύψος ιατρικών δαπανών παρατηρήθηκε ελάχιστα υψηλότερο σε ασφαλισμένους που διαμένουν στις βορειοανατολικές περιοχές των Η.Π.Α. (δ=10.057$) ενώ στις άλλες περιοχές κυμαίνεται μεταξύ 8.800$ και 9.300$. Επιπλέον, παρατηρήθηκε στατιστικά σημαντική θετική συσχέτιση του δείκτη μάζας σώματος ($r=0.834$, $p<0.05$) καθώς και της αρτηριακής πίεσης ($r=0.411$, $p<0.05$) με το ύψος των ασφαλιστικών απαιτήσεων στον πληθυσμό των καπνιστών, ενώ στον πληθυσμό των μη καπνιστών Οι συσχετίσεις που παρατηρήθηκαν είναι εξαιρετικά αδύναμες. Αυτό είναι ένα επίσης ένα αξιόλογο εύρημα που επιβεβαιώνει φυσικά και την βιβλιογραφία στην οποία το κάπνισμα αποδεδειγμένα επηρεάζει την κατάσταση υγείας του χρήστη, η οποία συνδέεται άμεσα με το κόστος των ιατρικών δαπανών, όπως φαίνεται και από τα παρακάτω διαγράμματα (4.1,4.3,4.4).



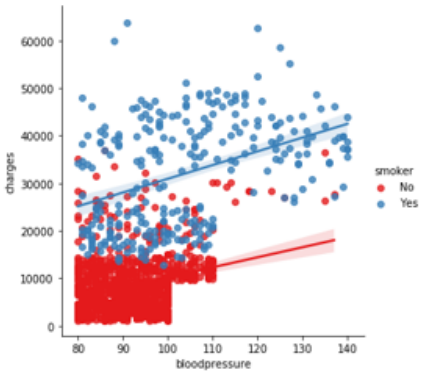

**Σχήμα 4.1:** Ενδιάμεσο κόστος ιατρικών δαπανών ανά καπνιστική συνήθεια




**Σχήμα 4.2:** Ενδιάμεσο κόστος ιατρικών δαπανών ανά περιοχή κατοικίας




**Σχήμα 4.3:** Διαγραμμα διασποράς για το ύψος ιατρικών δαπανών ανά ΔΜΣ και καπνιστική συνήθεια




**Σχήμα 4.4:** Διαγραμμα διασποράς για το ύψος ιατρικών δαπανών ανά αρτηριακή πίεση και καπνιστική συνήθεια

Δεν παρατηρήθηκαν άλλα αξιόλογα ευρήματα προς παρουσίαση, επομένως στη συνέχεια θα διενεργηθεί η διαδικασία της προεπεξεργασίας των δεδομένων ώστε να είναι έτοιμα προς εκμετάλευση κατά τη διαδικασία της δημιουργίας του προγνωστικού μοντέλου.

## Προεπεξεργασία των δεδομένων

Η προεπεξεργασία των δεδομένων στο συγκεκριμένο σετ δεδομένων δεν είχε δυσκολίες καθώς δεν παρατηρήθηκαν καθόλου missing values και τα δεδομένα δεν είχαν καθόλου θόρυβο. Ωστόσο παρατηρήθηκαν ακραίες τιμές στην κατανομή των ασφαλιστικών αποζημιώσεων, των οποίων η εξομάλυνσή ή η διαγραφή δεν κρίθηκε κατάλληλη, καθώς το σύνολο δεδομένων είναι αρκετά μικρό και η αναδιαμόρφωση ή η έλλειψη κάποιας πληροφορίας μπορεί να στοιχίσει στην ακρίβεια του μοντέλου μηχανικής μάθησης. Όσον αφορά την διαδικασία της κωδικοποίησης, όλες οι κατηγορικές μεταβλητές κωδικοποιήθηκαν, ώστε να μπορούν να αναγνωριστούν από τα μοντέλα μηχανικής μάθησης και φυσικά παράχθηκαν ψευδομεταβλητές από τις κατηγορικές μεταβλητές που είχαν πάνω από δύο κατηγορίες, όπως αυτό κρίνεται αναγκαίο για τη διενέργεια τεχνικών ανάλυσης παλινδρόμησης.

Επιπλέον, διενεργήθηκε η τυποποίηση των ποσοτικών χαρακτηριστικών ώστε η κατανομή τους να ακολουθεί την κανονική με μέση τιμή 0 και τυπική απόκλιση 1.

## Διενέργεια των μοντέλων μηχανικής μάθησης

Οι αλγόριθμοι παλινδρόμησης που χρησιμοποιήθηκαν και συγκρίθηκαν με σκοπό τη δημιουργία του πιο αποδοτικού μοντέλου πρόβλεψης του ύψους των ασφαλιστικών απαιτήσεων είναι 1) η Πολλαπλή Γραμμική παλινδρόμηση, 2) η Παλινδρόμηση LASSO, 3) η

Παλινδρόμηση με Δέντρα Απόφασης, 4) η Παλινδρόμηση με Τυχαία Δάση και 5) η Παλινδρόμηση Gradient Boosting.

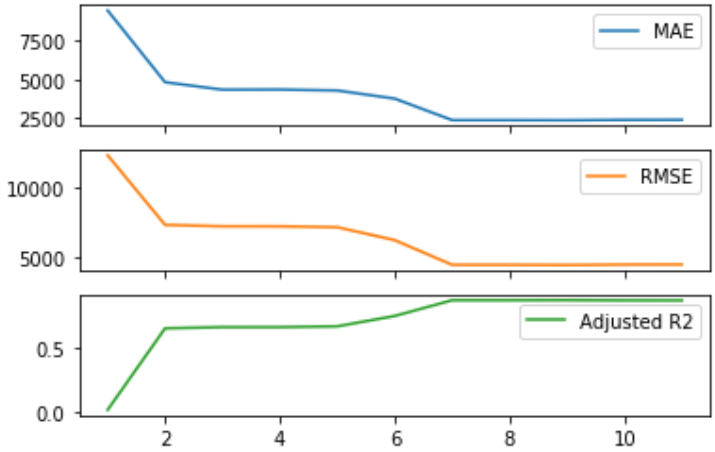
Στα πλαίσια της επιλογής του κατάλληλου μοντέλου, χρησιμοποιήθηκε η τεχνική της διασταυρούμενης επικύρωσης 10-fold (Cross Validation 10-fold), με δύο ειδών αναλογίες διαχωρισμού του συνόλου δεδομένων (70%-30% και 80%-20%) λαμβάνοντας υπόψιν τρεις (3) μετρικές για την αξιολόγηση της απόδοσης των μοντέλων παλινδρόμησης. Στον Πίνακα 4.5 παρουσιάζεται για κάθε μία από τις δύο αναλογίες, ο μέσος όρος των 10 επαναλήψεων της διασταυρούμενης επικύρωσης 10-fold για το μέσο απόλυτο σφάλμα της εκτίμησης (MAE), το ριζικό μέσο τετραγωνικό σφάλμα της εκτίμησης (RMSE) και τον προσαρμοσμένο συντελεστή προσδιορισμού (R squared), σε κάθε μοντέλο παλινδρόμησης.

**Πίνακας 4.5:** Αξιολόγηση της απόδοσης των μοντέλων παλινδρόμησης

| Split ratio | Model | MAE | RMSE | R squared |
|---|---|---|---|---|
| (70% - 30%) | Linear Regression | 3923,793 | 5713,398 | 77,12% |
| | Lasso Regression | 3923,797 | 5713,401 | 77,12% |
| | Decision Tree Regression | 2614,413 | 5685,647 | 77,45% |
| | Random Forest Regression | 2143,71 | 4050,89 | 88,44% |
| | Gradient Boosting Regression | 2154 | 4029,41 | 88,6% |
| (80% - 20%) | Linear Regression | 3923,792 | 5713,398 | 77,12% |
| | Lasso Regression | 3923,80 | 5713,402 | 77,12% |
| | Decision Tree Regression | 2606,57 | 5576,647 | 76,86% |
| | Random Forest Regression | 2143,72 | 4050,89 | 88,44% |
| | Gradient Boosting Regression | 2146,11 | 4029,9 | 88,52% |

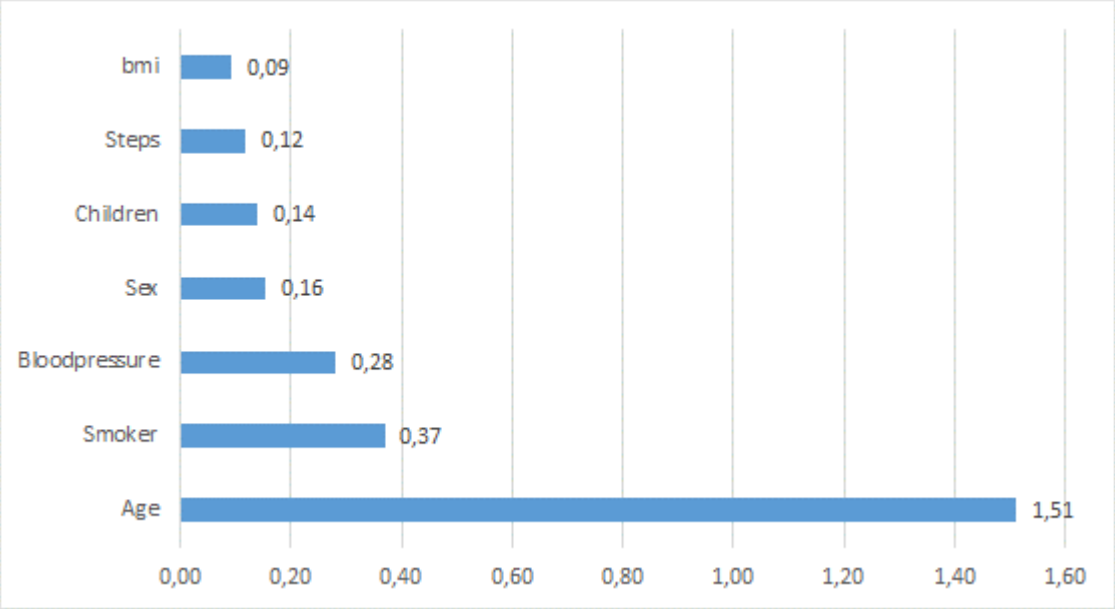
Από τα αποτελέσματα του παραπάνω πίνακα προκύπτει ότι τα μοντέλα που φαίνεται να λειτουργούν καλύτερα για την πρόβλεψη του ύψους των ασφαλιστικών απαιτήσεων είναι αυτά που χρησιμοποιούν τα δέντρα αποφάσεων, ενώ η πολλαπλή γραμμική παλινδρόμηση και η παλινδρόμηση Lasso, φαίνεται να κάνουν πιο αδύναμες προβλέψεις και με περισσότερα σφάλματα.

Συγκεκριμένα, καλύτερη απόδοση παρουσίασε η παλινδρόμηση που χρησιμοποιεί την τεχνική Gradient Boosting με αναλογία διαχωρισμού 80%-20%, επομένως το προγνωστικόμοντέλο που θα κατασκευαστεί θα βασίζεται σε αυτή την τεχνική. Για την κατασκευή του μοντέλου δε χρησιμοποιήθηκαν όλες οι ανεξάρτητες μεταβλητές, παρά μόνο οι επτά (7) πιο σημαντικές για το μοντέλο, καθώς οι υπόλοιπες δε φάνηκε να προσθέτουν κάποια επιπλέον ερμηνευτιική αξία στο μοντέλο (Σχήμα 4.5). Συγεκριμένα οι μεταβλητές που κρίθηκαν πιο σημαντικές για το τελικό μοντέλο,




**Σχήμα 4.5**:Number of features selected

σύμφωνα με τη συνεισφορά τους στη διακύμανση του μοντέλου ήταν η ηλικία, η καπνιστική συνήθεια, η αρτηριακη πίεση, το φύλο, ο μέσος αριθμός βημάτων και ο ΔΜΣ, όπως φαίνεται και στο σχήμα 4.6 που παρουσιάζει τη σημαντικότητα των μεταβλητών στο μοντέλο.




**Σχήμα 4.6**:Feature importance chart of Gradient Boosting Regressor

Η παλινδρόμηση που χρησιμοποιεί την τεχνική Gradient Boosting έχει ένα σύνολο παραμέτρων οι οποίοι μπορούν να βελτιστοποιηθούν και να δώσουν καλύτερες εκτιμήσεις στο τελικό μοντέλο. Οι βασικοί παράμετροι ενλισχυσης κλίσης που χρησιμοποιήθηκαν για τον ορισμό ενός δέντρου απόφασης, σύμφωνα με το πακέτο scikit-learn της python, είναι οι εξής:

1. **Number of estimators (Default =100)**
   Ο αριθμός των δέντρων που θα χρησιμοποιηθούν για την ενίσχυση κλίσης. Συνήθως, ένας μεγαλύτερος αριθμός δέντρων οδηγούν σε καλύτερη απόδοση του μοντέλου.
2. **Max depth (Default =3)**
   Είναι το μέγιστο βάθος ενός δυαδικού δέντρου, δηλαδή ο αριθμός των κόμβων στο ύψος του δυαδικού δέντρου
3. **Minimum samples to split (Default =2)**
   Καθορίζει τον ελάχιστο αριθμό δειγμάτων ή παρατηρήσεων που απαιτούνται σε ένα κόμβο για να ληφθούν υπόψη για διαχωρισμό. Χρησιμοποιείται για τον έλεγχο της υπερπροσαρμογής
4. **Learning rate (Default =0,1)**
   Είναι ο ρυθμός εκμάθησης, η οποία είναι μια υπερ-παράμετρος στο αλγόριθμος παλινδρόμησης ενίσχυσης κλίσης, που καθορίζει την ποσότητα ανανέωσης των βαρών κατά τη διάρκεια της εκπαίδευσης του μοντέλου. Δέχεται τιμές στο διάστημα τιμών [0,1]

Για τη διαδικασία βελτιστοποίησης των παραμέτρων διενεργήθηκε διασταυρούμενη επικύρωση 5-fold (Cross validation 5-fold) και το βέλτιστο μοντέλο προέκυψε για: number of estimators=300, max_depth=4, minimum samples to split = 4 και learnng rate = 0.01.

Στον πίνακα 4.6 παρουσιάζεται η απόδοση του τελικού μοντέλου παλινδρόμησης με την τεχνική Gradient Boosting.

**Πίνακας 4.6:** Αξιολόγηση της απόδοσης του μοντέλου παλινδρόμησης Gradient Boosting στο Training set και στο test set

| Set | Model | MAE | RMSE | Adjusted R squared |
|---|---|---|---|---|
| Training | Gradient Boosting Regression | 1658,58 | 3.195,89 | 91,75% |
| Test | Gradient Boosting Regression | 2361,14 | 4479,46 | 87,22% |

Όπως φαίνεται και στον πίνακα 4.6, η ερμηνευτική δυνατότητα του μοντέλου στο σύνολο δεδομένων εκπαίδευσης είναι ίση με 91,75% και στο σύνολο δεδομένων δοκιμής είναι ίση με 87,22%, βάσει του προσαρμοσμένου συντελεστή προσδιορισμού. Αυτό σημαίνει ότι ανεξάρτητες μεταβλητές του μοντέλου ερμηνεύουν το 87,22% της μεταβλητότητας του μοντέλου με εξαρτημένη τη μεταβλητή «Ύψος ασφαλιστικών απαιτήσεων».

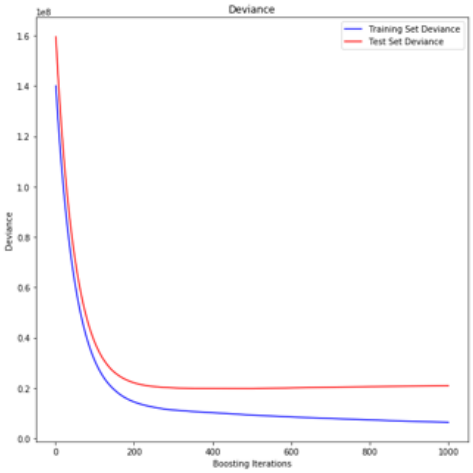
Το προγνωστικό μοντέλο θα είναι άρρηκτα συνδεδεμένο με τη διαχείριση του κινδύνου της ασφαλιστικής εταιρείας και με τη βοήθεια του η εταιρεία μπορεί να μειώσει το διοικητικό της κόστος, να προνοήσει, διατηρώντας κάποιο αποθεματικό χρηματικό κεφάλαιο για μελλοντικές αξιώσεις καθώς και να ανακαλύψει πρότυπα τα οποία μπορεί να οδηγήσουν στην αναπροσαρμογή της τιμολογιακής της πολιτικής στα ασφάλιστρα. Συγκεκριμένα, η ασφαλιστική εταιρεία γνωρίζοντας το μίγμα των ασφαλισμένων και τα χαρακτηριστικά τους τα οποία τα αντλεί από το αίτημα συμβολαίου, μπορεί να υπολογίσει με μεγάλη ακρίβεια το ύψος των μελλοντικών ασφαλιστικών απαιτήσεων καθώς και να ενδυναμώσει την ακρίβεια της διαδικασίας αξιολόγησης της επικινδυνότητας των ασφαλισμένων. Επίσης, με περαιτέρω ανάλυση, το προγνωστικό μοντέλο μπορεί να χρησιμοποιηθεί για δυναμική τιμολόγηση των ασφαλίστρων.

**Πίνακας 4.7:** Ενδεικτική παρουσίαση των 10 πρώτων ποσών των ασφαλιστικών απαιτήσεων που προβλέφθηκαν από το μοντέλο σε σχέση με το πραγματικό ύψος των απαιτήσεων.

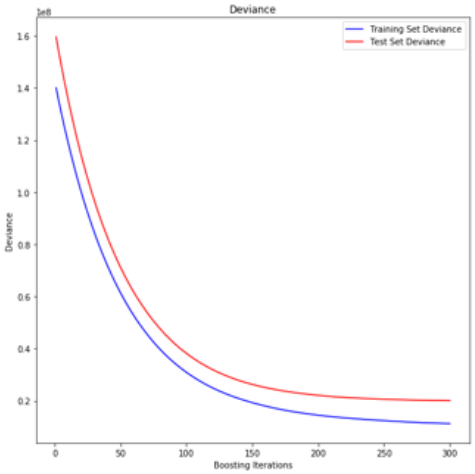
| | True | Prediction |
|---|---|---|
| 0 | 8.116,26 | 8.662,89 |
| 1 | 8.534,67 | 9.795,29 |
| 2 | 7.986,47 | 7.845,76 |
| 3 | 18.648,42 | 18.710,33 |
| 4 | 1.705,62 | 2.881,03 |
| 5 | 15.019,76 | 14.839,93 |
| 6 | 1.628,47 | 3.519,89 |
| 7 | 39.836,52 | 38.196,49 |
| 8 | 36.397,57 | 34.542,53 |
| 9 | 47.896,79 | 44.197,96 |

Όπως φαίνεται στο σχήμα 4.7, το σφάλμα εκπαίδευσης, φαίνεται ότι μειώνεται γρήγορα στα πρώτα 200 δέντρα και στη συνέχεια φαίνεται να συνεχίζεται να μειώνεται όσο προστίθεντε και άλλα δέντρα ενίσχυσηςε κλίσης. Όσον αφορά το σφάλμα δοκιμής, φαίνεται να μειώνεται επίσης γρήγορα στην αρχή, αλλά στη συνέχεια επιβραδύνεται και φτάνει στο ελάχιστό του

αρκετά γρήγορά (~300 δέντρα), ενώ στη συνέχεια αρχίζει ακόμη και να αυξάνεται. Αυτό είναι σημάδι υπερπροσαρμογής, καθώς φαίνεται ότι μετά το σημείο προσθήκης των 300 δέντρων ενίσχυσης κλίσης, το μοντέλο έχει τόση χωρητικότητα που αρχίζει να προσαρμόζει τις ιδιοσυγκρασίες του συνόλου δεδομένων εκπαίδευσης. Γι αυτό το λόγο, για να αποτραπεί η υπερπροσαρμογή, επιλέχθηκε το μοντέλο με αριθμό δέντρων ίσο με 300, όπως φαίνεται και στο σχήμα 4.8.




**Σχήμα 4.7**:Training & Test set deviance




**Σχήμα 4.8**:Training & Test set deviance

## 4.2. 2η Εφαρμογή - Πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο με τη λήξη του

### Πρόβλημα

Οι ασφαλιστικές εταιρείες δραστηριοποιούνται σε ένα άκρως ανταγωνιστικό και δυναμικό περιβάλλον με χιλιάδες ή εκατομμύρια πελάτες και τεράστιες δεξαμενές δεδομένων που σχετίζονται με αυτούς. Για τις ασφαλιστικές εταιρίες, η απόκτηση και η διατήρηση των πελατών είναι εξίσου σημαντικές, ωστόσο η πρώτη είναι αρκετά πιο δαπανηρή διαδικασία. Επομένως, οι εταιρίες βασίζονται στα δεδομένα που έχουν στην κατοχή τους, με σκοπό να κατανοήσουν τη συμπεριφορά των πελατών και να αποτρέψουν την αποχώρησή τους από την εταιρία ή το ασφαλιστήριο συμβόλαιο.

### Σκοπός

Σκοπός της ανάλυσης είναι η δημιουργία ενός προγνωστικού μοντέλου για την πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο. Αυτό το προγνωστικό μοντέλο, θα κατασκευαστεί χρησιμοποιώντας ιστορικά δεδομένα πελατών και μπορεί να βοηθήσει την ασφαλιστική εταιρεία να αναπροσαρμόσει τις στρατηγικές της για τη διατήρηση των πελατών της σε ένα πιο συγκεκριμένο δείγμα πελατών των οποίων τα συμπεριφορικά δεδομένα είναι ύποπτα για αποχώρηση από το ασφαλιστήριο συμβόλαιο.

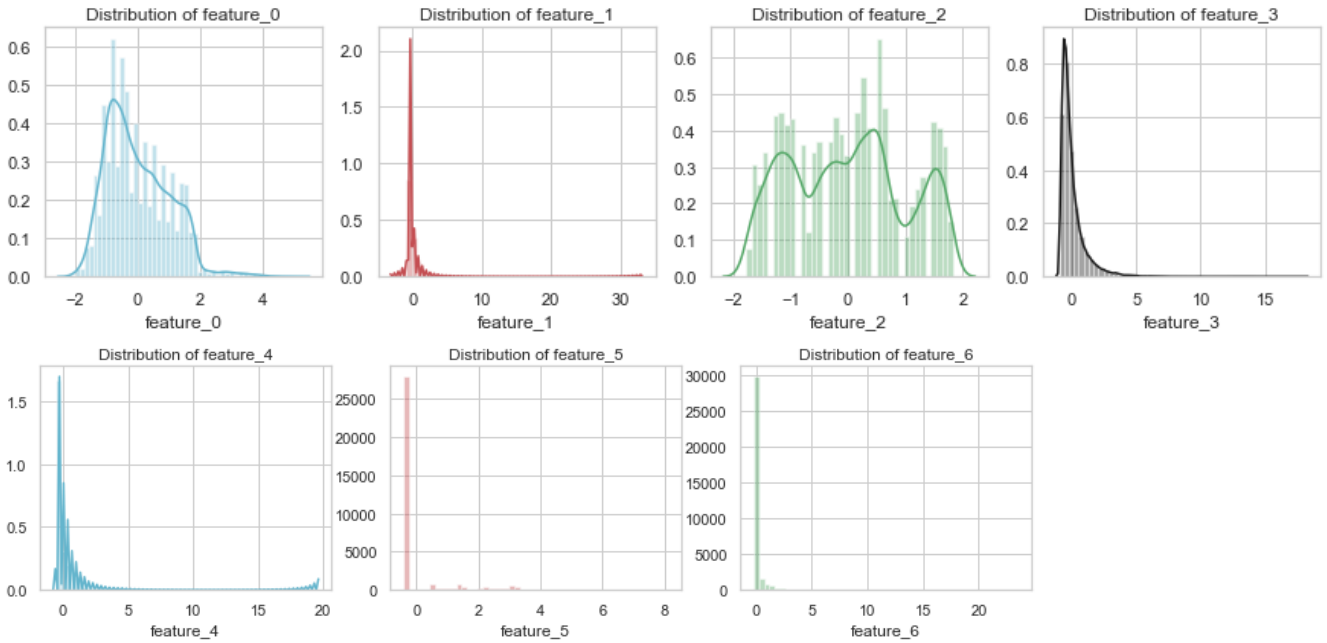
### Περιγραφή του συνόλου δεδομένων

Η ανάλυση αφορά ένα αρχείο που περιέχει δεδομένα σε ανωνυμοποιημένη μορφή για 33,908 πελάτες μιας ασφαλιστικής εταιρίας και αντλήθηκε από τον ιστότοπο της Kaggle.

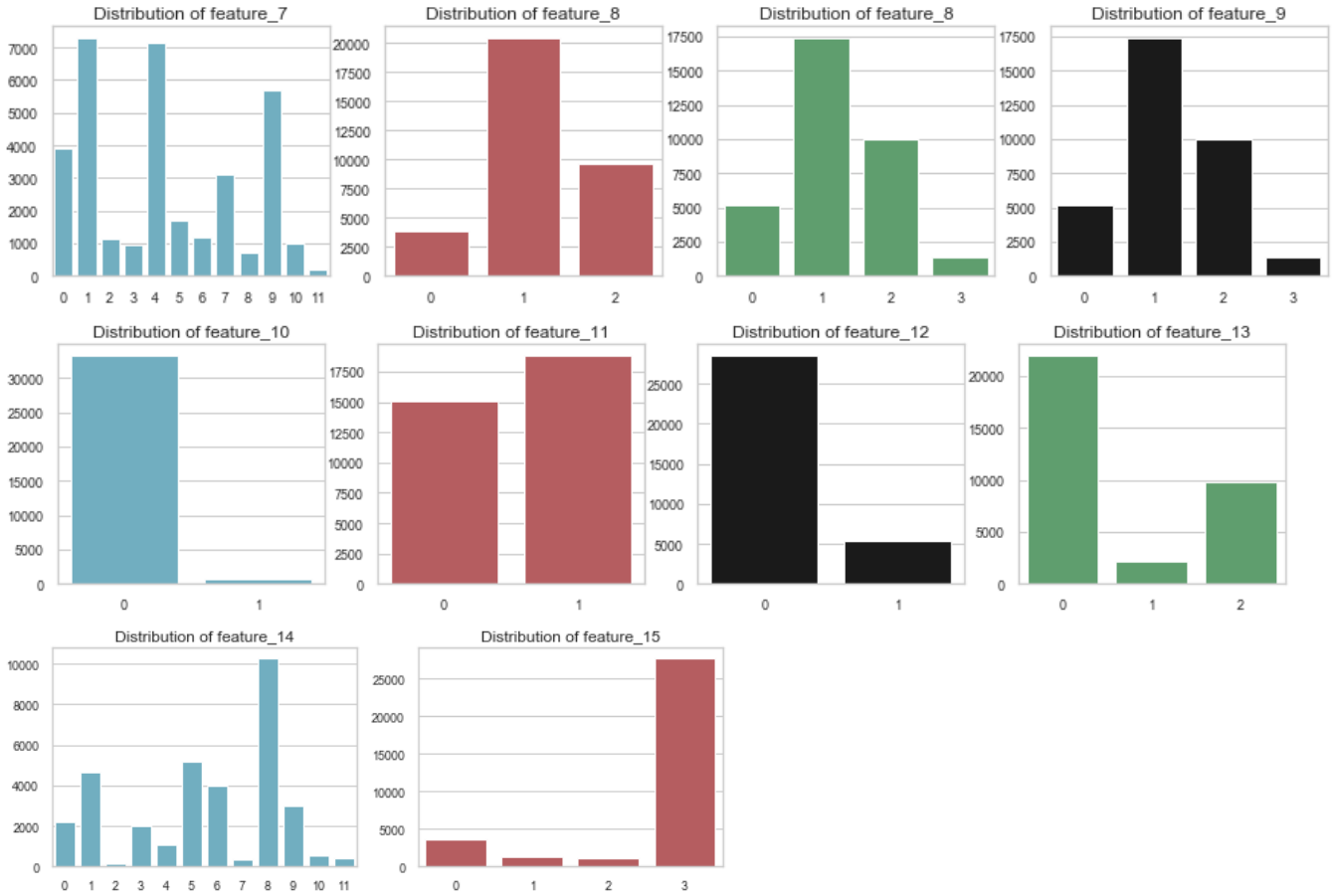
Συγκεκριμένα, το αρχείο εμπεριέχει δεδομένα από 33908 πελάτες με 16 ανωνυμοποιημένους παράγοντες που αφορούν δημογραφικά, προσωπικά και συμπεριφορικά δεδομένα των πελατών, καθώς και την πληροφορία αν ο πελάτης έχει αποχωρήσει από το ασφαλιστήριο συμβόλαιο ή όχι από ιστορικά δεδομένα. Η πληροφορία για το είδος των χαρακτηριστικών δεν υπάρχει, επομένως σε πρώτη φάση προηγήθηκε μια διερευνητική ανάλυση με τη χρήση κάποιων λογικών κανόνων, για την κατηγοριοποίηση των μεταβλητών σε ποσοτικές και κατηγορικές μεταβλητές, ώστε να επιτευχθεί η σωστή εκμετάλλευσή τους κατά τη διαδικασία της μοντελοποίησης. Στον πίνακα 4.8, παρουσιάζονται οι μεταβλητές και ο τύπος των μεταβλητών, ενώ στα σχήματα 4.9, 4.10 και 4.11 δίνεται η διαγραμματική τους απεικόνιση.

**Πίνακας 4.8:** Χαρακτηριστικά του συνόλου δεδομένων

| **Ποσοτικές μεταβλητές** | **Κατηγορικές μεταβλητές** |
|---|---|
| Feature_1 | Feature_7 |
| Feature_2 | Feature_8 |
| Feature_3 | Feature_9 |
| Feature_4 | Feature_10 |
| Feature_5 | Feature_11 |
| Feature_6 | Feature_12 |
| | Feature_13 |
| | Feature_14 |
| | Feature_15 |
| | Churn |

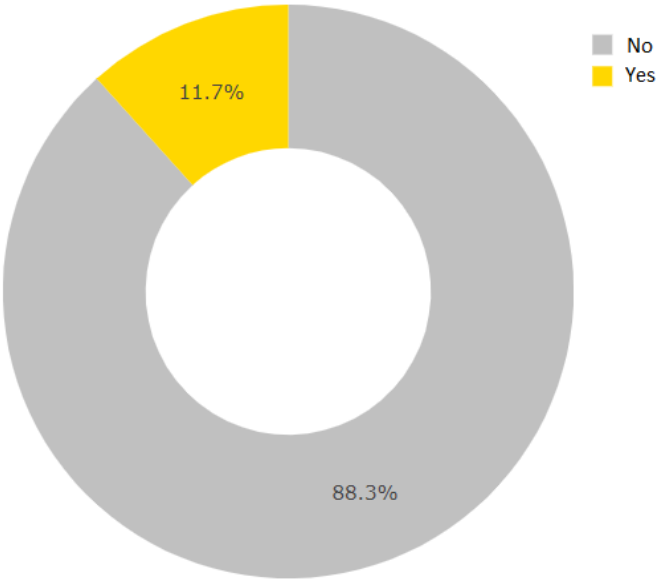



**Σχήμα 4.9**:Ιστογράμματα των ποσοτικών μεταβλητών




**Σχήμα 4.10**:Ραβδογράμματα των κατηγορικών μεταβλητών

Στο σύνολο δεδομένων, παρατηρείται ότι το 88,3% (N=29.941) των πελατών ανανέωσε το ασφαλιστήριο συμβόλαιο και μόλις το 11,7% (N=3967) αποφάσισε να αποχωρήσει στη λήξη του (Σχήμα 4.11).




**Σχήμα 4.11**:Κατανομή των πελατών που αποχώρησαν ή όχι από το ασφαλιστήριο συμβόλαιο

**Προεπεξεργασία των δεδομένων**

Διενεργήθηκε η τυποποίηση των ποσοτικών χαρακτηριστικών ώστε η κατανομή τους να ακολουθεί την κανονική με μέση τιμή 0 και τυπική απόκλιση 1. Δεν παρατηρήθηκαν ελλείπουσες τιμές στο σύνολο δεδομένων.

**Διενέργεια των μοντέλων μηχανικής μάθησης**

Οι αλγόριθμοι εποπτευόμενης κατηγοριοποίησης που χρησιμοποιήθηκαν και συγκρίθηκαν με σκοπό τη δημιουργία του πιο αποδοτικού μοντέλου πρόβλεψης των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο είναι 1) η Λογιστική παλινδρόμηση, 2) η Γραμμική Διακριτική Ανάλυση, 3) οι Μηχανές διανυσματικής Υποστήριξης, 4) οι Κ κοντινότεροι γείτονες και 5) ο Extreme Gradient Boosting.

Στα πλαίσια της επιλογής του κατάλληλου μοντέλου, χρησιμοποιήθηκε η τεχνική της διασταυρούμενης επικύρωσης 10-fold (Cross Validation 10-fold), με την αναλογία διαχωρισμού του συνόλου δεδομένων να είναι (70%-30%) λαμβάνοντας υπόψιν τέσσερις (4) μετρικές για την αξιολόγηση της απόδοσης των μοντέλων παλινδρόμησης. Στον Πίνακα 4.8 παρουσιάζεται ο μέσος όρος των 10 επαναλήψεων της διασταυρούμενης επικύρωσης 10-fold με το Accuracy, το Precision, το Recall, το F-score και την τιμή AUC για κάθε μοντέλο ταξινόμησης. Ο μέσος όρος για τα Precision, Recall και F-score υπολογίστηκε βάσει της κλάσης ενδιαφέροντος, που είναι η αποχώρηση από το ασφαλιστήριο συμβόλαιο.

**Πίνακας 4.9:** Αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης

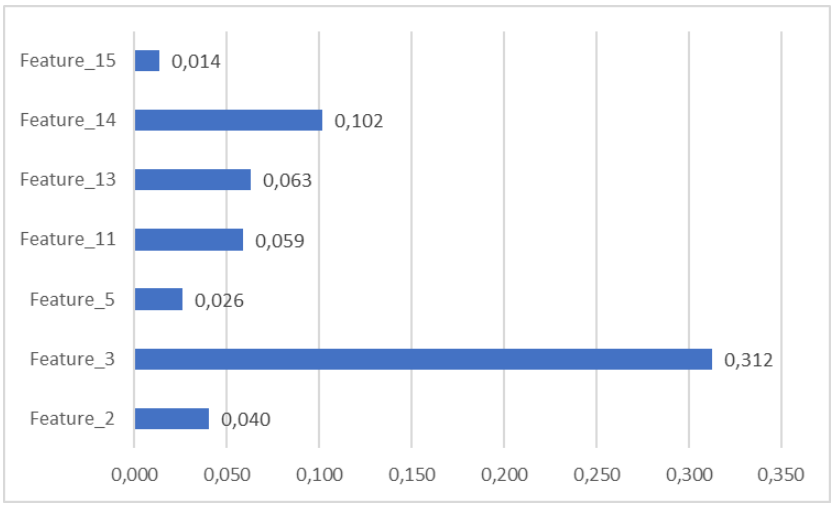
| Model | Accuracy (%) | Precision (%) | Recall (%) | F-score (%) | AUC (%) |
|---|---|---|---|---|---|
| Logistic Regression | 89,03 | 59,00 | 20,70 | 30,65 | 86,13 |
| Linear Discriminant | 88,97 | 55,72 | 28,22 | 37,47 | 86,47 |
| SVM | 89,52 | 68,17 | 19,56 | 30,40 | 80,60 |
| Kneighbors | 89,57 | 57,25 | 42,83 | 49,00 | 84,95 |
| XGBoost | 90,38 | 62,62 | 44,18 | 51,81 | 91,80 |

Από τα αποτελέσματα του πίνακα 4.9 προκύπτει ότι όλα τα μοντέλα ταξινόμησης, έχουν πολύ μεγάλη ακρίβεια πρόβλεψης στο προφίλ των πελατών που δε πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο, ενώ εμφανίζουν αδυναμία στην πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν. Ωστόσο, ο κύριος λόγος κατασκευής του προγνωστικού μοντέλου είναι η πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο και μάλλον η αδυναμία των μοντέλων οφείλεται στο γεγονός ότι οι υπάρχει ανισομερής κατανομή των ετικετών της μεταβλητής στόχου καθώς στο σύνολο δεδομένων όπως προαναφέρθηκε, το 88,3% (N=29.941) των πελατών ανανέωσε το ασφαλιστήριο συμβόλαιο και μόλις το 11,7% (N=3967) αποφάσισε να αποχωρήσει στη λήξη του. Από τη βιβλιογραφία ένας από τους τρόπους που αναφέρεται για τη διαχείριση των ανομοιογενών δεδομένων είναι η υπερδειγματοληψία. Επομένως, θα επαναληφθεί η διαδικασία αξιολόγησης των μοντέλων, αφού πρώτα προηγηθεί η συνθετική τεχνική υπερδειγματοληψίας μειονότητας (Synthetic minority oversampling technique - SMOTE), για να επιτευχθεί μια πιο ισορροπημένη κατανομή των τάξεων και να βελτιωθεί η απόδοση του μοντέλου.

**Πίνακας 4.10:** Αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης με την εφαρμογή της συνθετικής τεχνικής υπερδειγματοληψίας μειονότητας (SMOTE)

| Model | Accuracy (%) | Precision (%) | Recall (%) | F-score (%) | AUC (%) |
|---|---|---|---|---|---|
| Logistic Regression | 89,04 | 58,46 | 20,34 | 30,18 | 86,78 |
| Linear Discriminant | 89,04 | 56,18 | 28,89 | 38,16 | 87,26 |
| SVM | 89,16 | 65,97 | 15,42 | 25,00 | 86,29 |
| Kneighbors | 89,89 | 59,83 | 41,41 | 48,94 | 85,87 |
| XGBoost | 90,58 | 62,30 | 49,28 | 55,03 | 92,68 |

Από τους παραπάνω αλγόριθμους οι K πλησιέστεροι γείτονες και ο Extreme Gradient Boosting φαίνεται να αποδίδουν καλύτερα με τον δεύτερο να παρουσιάζει την καλύτερη απόδοση από όλους τους αλγόριθμους μηχανικής μάθησης και πριν και μετά την υπερδειγματοληψία με την τεχνική SMOTE. Επομένως το μοντέλο που θα χρησιμοποιηθεί για προγνωστικά θα είναι το XGBoost το οποίο μάλιστα φαίνεται να βελτιώθηκε αρκετά μετά την εφαρμογή της τεχνικής υπερδειγματοληψίας SMOTE. Για την κατασκευή του τελικού μοντέλου δε χρησιμοποιήθηκαν όλες οι μεταβλητές, παρά μόνο εκείνες που προσέφεραν κάποια μεταβλητότητα στο τελικό μοντέλο. Στο παρακάτω σχήμα 4.12 παρουσιάζεται η προσφορά των επιλεχθέντων χαρακτηριστικών στη συνολική μεταβλητότητα του μοντέλου. Η κατασκευή του μοντέλου καθώς και η επιλογή των βέλτιστων παραμέτρων έγινε με έμφαση στη μεγιστοποίηση του f-score, δηλαδή του αρμονικού μέσου μεταξύ του precision και recall.



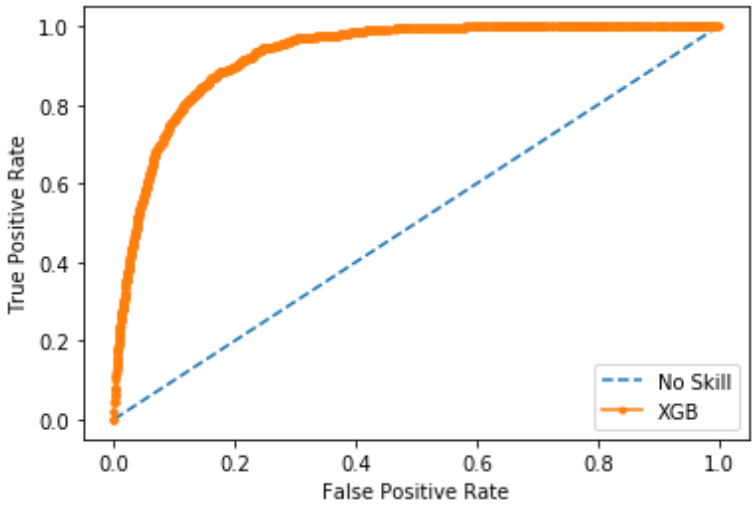

**Σχήμα 4.12**:Διακύμανση των χαρακτηριστικών στο μοντέλο

**Πίνακας 4.11:** Αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης XGBoost στο Training set και στο test set

| Set | Class | Accuracy (%) | Precision (%) | Recall (%) | F-score (%) | AUC (%) |
|---|---|---|---|---|---|---|
| Training | Not Churn | 95,83 | 97,00 | 95,00 | 96,00 | 99,4 |
| | Churn | | 95,14 | 96,60 | 95,86 | |
| Test | Not Churn | 90,00 | 96,00 | 93,00 | 94,00 | 92,7 |
| | Churn | | 56,09 | 68,75 | 62,32 | |

**Πίνακας 4.12:** Πίνακας ταξινόμησης του μοντέλου XGBoost

| | Predicted | | Total |
|---|---|---|---|
| **Actual** | **Churn** | **Not Churn** | |
| **Churn** | 548 | 249 | 797 |
| **Not Churn** | 429 | 5.556 | 5.985 |
| **Total** | 977 | 5.805 | 6782 |




**Σχήμα 4.13**:ROC Curve

Τα αποτελέσματα της ανάλυσης για τα ιστορικά στοιχεία τα οποία φαίνονται στον Πίνακα 4.6, έδειξαν ότι ταξινομήθηκε σωστά το 90% των περιπτώσεων (Ορθότητα), ωστόσο αυτό δεν είναι κατάλληλο μέτρο αξιολόγησης του μοντέλου, καθώς τα δεδομένα στο σετ δοκιμής είναι ανομοιογενή και το συνολικό ποσοστό ταξινόμησης επηρεάζεται από το γεγονός ότι το μοντέλο ταξινόμησης έχει πολύ ισχυρή προβλεπτική ακρίβεια στην αρνητική τάξη (Ασφαλισμένοι που δε θα αποχωρήσουν). Συγκεκριμένα, φαίνεται ότι από το σύνολο των ασφαλισμένων που ταξινομήθηκαν ως μη επικίνδυνοι για αποχώρηση, το 96% από αυτούς στην πραγματικότητα δεν αποχώρησαν (Ακρίβεια). Ωστόσο, στο σύνολο των ασφαλισμένων που ταξινομήθηκαν ως επικίνδυνοι για αποχώρηση, μόλις το 56% από αυτούς στην πραγματικότητα αποχώρησαν (Ακρίβεια). Επίσης, το μοντέλο φαίνεται ότι αναγνώρισε το 69% από τους ασφαλισμένους που στην πραγματικότητα αποχώρησαν το οποίο είναι ένα ποσοστό το οποίο είναι αρκετά ικανοποιητικό και σίγουρα αναδυκνύει τη σημαντικότητα του μοντέλου.

Στο σχήμα 4.13 παρατίθεται η διαγραμματική απεικόνιση της καμπύλης ROC. Σύμφωνα με την καμπύλη ROC, η προβλεπτική αξία του μοντέλου είναι ίση με 92,7%, ένα ποσοστό που είναι εξαιρετικά υψηλό και επιβεβαιώνει την απόδοσή του.

Η αναγνώριση του προφίλ των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο ή την εταιρεία είναι διαδικασία ζωτικής σημασίας για τις ασφαλιστικές εταιρείες οι οποίες θέτουν σαν προτεραιότητα τη διατήρηση των πελατών τους και όχι την εύρεση νέων, καθώς το δεύτερο καθίσταται πολύ κοστοβώρο.

Το τελικό μοντέλο που θα προκύψει θα δίνει ένα score για το προφίλ του κάθε ασφαλισμένου. Με βάση το score αυτό, το προφίλ ενός ασφαλισμένου θα ταξινομείται ως πιθανό ή όχι για αποχώρηση από την εταιρεία. Παρακάτω ακολουθούν 2 περιπτώσεις που έλαβε υπόψιν του αλγόριθμος από το σύνολο δεδομένων δοκιμής, στα οποία παρουσιάζεται ποιες μεταβλητές έχουν συνεισφορά σε κάθε κλάση της μεταβλητής στόχος καθώς και η πιθανότητα αποχώρησης.

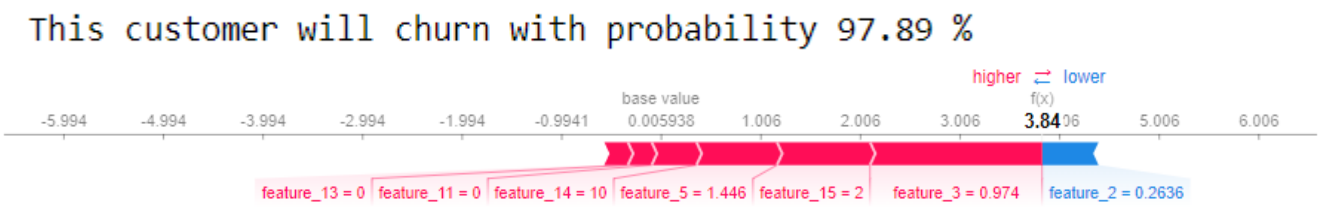
**Πίνακας 4.13:** Στοιχεία του ασφαλισμένου στην θέση 27520 στο σύνολο δεδομένων δοκιμής

| Index | Feature_2 | Feature_3 | Feature_5 | Feature_11 | Feature_13 | Feature_14 | Feature_15 |
|---|---|---|---|---|---|---|---|
| 27520 | 0,02326 | -0,404478 | 0,787017 | 0 | 0 | 10 | 0 |


**Σχήμα 4.14**:Force plot

**Πίνακας 4.14:** Στοιχεία του ασφαλισμένου στην θέση 32354 στο σύνολο δεδομένων δοκιμής

| Index | Feature_2 | Feature_3 | Feature_5 | Feature_11 | Feature_13 | Feature_14 | Feature_15 |
|---|---|---|---|---|---|---|---|
| 32354 | 0,263576 | 0,97403 | 1,446176 | 0 | 0 | 10 | 2 |




**Σχήμα 4.15**:Force plot

Σύμφωνα με το μοντέλο και ορίζοντας ως κατώφλι για λήψη αποφάσεων το 50%, προκύπτει ότι ο ασφαλισμένος του πρώτου παραδείγματος έχει 38,25% πιθανότητα να αποχωρήσει από την ασφαλιστική, κάτι που δε τον κατατάσει στην κατηγορία κινδύνου για αποχώρηση, ενώ ο ασφαλισμένος του δεύτερου παραδείγματος έχει 97,89% πιθανότητα να αποχωρήσει από την ασφαλιστική, κάτι που τον κατατάσει στην κατηγορία υψηλού κινδύνου για αποχώρηση.

Τα διαγράμματα force plot 4.14 και 4.15 παρουσιάζουν διαγραμματικά ποιες μεταβλητές έχουν συνεισφορά σε κάθε κλάση της μεταβλητής στόχος. Με κόκκινο συμβολίζεται η θετική τάξη και με μπλε η αρνητική τάξη. Όσο ποιο μεγάλη είναι η κόκκινη λωρίδα, τόσο πιο μεγάλη είναι η πιθανότητα μια περίπτωση να ανήκει στην τάξη κινδύνου και όσο μεγαλύτερο εμβαδόν πιάνει στη λωρίδα η κάθε μεταβλητή, τόσο μεγαλύτερη είναι και η συνεισφορά της στη λήψη της τελικής απόφασης.

Για να γίνει αντιληπτή η σημαντικότητα του μοντέλου ας γίνει μια υπόθεση ότι το σύνολο των 6782 ασφαλισμένων έχουν συνάψει με την ασφαλιστική εταιρεία ένα ετήσιο συμβόλαιο υγείας το οποίο έχει ασφάλιστρο ύψους 1.200€. Αυτό σημαίνει ότι η συνολική αξία

των πελατών της είναι ίση με 8.138.400€. Η ασφαλιστική εταιρεία με την ανανέωση του συμβολαίου προσφέρει κουπόνια αξίας 200€ για αγορές, στους πελάτες που προβλέπονται από το μοντέλο ως επικίνδυνοι να αποχωρήσουν. Από το σύνολο των 6782 ασφαλισμένων, ο αλγόριθμος ταξινόμησε ως επικίνδυνους για αποχώρηση 977 ασφαλισμένους. Αυτό σημαίνει, ότι η ασφαλιστική εταιρεία συμβουλευόμενη το μοντέλο για την διατήρηση των πελατών της θα έπρεπε να μοιράσει κουπόνια σε 977 πελάτες αξίας (977×200) 195.400€. Από ιστορικά δεδομένα φαίνεται ότι θα αποχωρήσουν 797 πελάτες με το μοντέλο να βρίσκει επιτυχώς το 69% (N=548) από αυτούς. Το κόστος αποχώρησης των 548 πελατών από την ασφαλιστική είναι ίσο με (548*1.200) 657.600€, επομένως η ασφαλιστικη χρησιμοποιώντας το μοντέλο θα έχει μέγιστο αναμενόμενο συνολικό όφελος που θα ανέρχεται περίπου στο ύψος των 462.200€ (657.600€ - 195.400€). Η χρησιμότητα του τελικού μοντέλου δεν έγκειται μόνο στην αναγνώριση των πελατών που πρόκειται να ανανεώσουν ή όχι το ασφαλιστήριο συμβόλαιο, αλλά και στον εντοπισμό των κύριων παραγόντων που σχετίζονται με την απόφαση αυτή για κάθε μεμονωμένη περίπτωση. Τα στελέχη των ασφαλιστικών εταιρειών έχοντας στη διάθεσή τους αυτή την πληροφορία μπορούν να εντοπίσουν τα κρίσιμα σημεία που οδηγούν στην απόφαση ανανέωσης ή αποχώρησης του πελάτη από το συμβόλαιο, να αναδιαμορφώσουν τις προσφορές τους αναλόγως το προφίλ του πελάτη και να δημιουργήσουν πιο προσωποποιημένες προσφορές. Οι αναλογιστές επίσης, μπορούν να αξιοποιήσουν τις προβλέψεις του μοντέλου για τον αριθμό των πελατών που πρόκειται να ανανεώσουν και να βελτιστοποιήσουν τα σχέδια τιμολόγησής τους.

## 4.3. 3η Εφαρμογή - Πρόβλεψη πώλησης ασφάλισης οχήματος σε πελάτες με συμβόλαια ασφάλισης υγείας

### Πρόβλημα

Η διαδικασία της διασταυρούμενης πώλησης (cross-selling), περιλαμβάνει την πώληση συμπληρωματικών προϊόντων σε υπάρχοντες πελάτες. Είναι μια από τις εξαιρετικά αποτελεσματικές τεχνικές στον κλάδο του μάρκετινγκ, όχι μόνο επειδή ενισχύει τα έσοδα από τις πωλήσεις σε μια υπάρχουσα πελατειακή βάση, αλλά επειδή αυξάνει την ικανοποίηση των πελατών, δημιουργεί αφοσίωση και βοηθά στη δημιουργία σταθερών και διαρκών σχέσεων με τους πελάτες. Οι ασφαλιστικές εταιρείες, δαπανούν πολλούς οικονομικούς πόρους για τηλεφωνικές προωθητικές ενέργειες προϊόντων ή υπηρεσιών σε υφιστάμενους πελάτες. Η αναλυτική των δεδομένων, η μηχανική μάθηση και η τεχνητή νοημοσύνη αποτελούν εξαιρετικά εργαλεία στη διαδικασία της στόχευσης δηλαδή, στη διαδικασία εντοπισμού των υφιστάμενων πελατών, οι οποίοι θα αποκρίνονταν θετικά, με υψηλή πιθανότητα, στην πρόταση για αγορά κάποιας επιπλέον υπηρεσίας ή προϊόντος.

### Σκοπός

Σκοπός της ανάλυσης είναι αρχικά η διερεύνηση των δεδομένων και στη συνέχεια η διαμόρφωση κατάλληλου μοντέλου, βάσει των διαθέσιμων δεδομένων, ούτως ώστε μελλοντικά το μοντέλο αυτό να υποδεικνύει αυτόματα τους πελάτες που έχουν μεγαλύτερη πιθανότητα να ενδιαφέρονται για αγορά Ασφάλισης οχήματος. Το μοντέλο, όπως θα αποδειχθεί, είναι εξαιρετικά χρήσιμο για μία ασφαλιστική επιχείρηση, καθώς της επιτρέπει έναν πιο ουσιαστικό σχεδιασμό της στρατηγικής επικοινωνίας και προσέγγισης των πελατών της, βελτιστοποιώντας έτσι το επιχειρηματικό της μοντέλο.

## Περιγραφή του συνόλου δεδομένων

Τα διαθέσιμα δεδομένα αφορούν πελάτες μιας ασφαλιστικής επιχείρησης, οι οποίοι είναι κάτοχοι ασφαλιστηρίων συμβολαίων υγείας. Το αρχείο εμπεριέχει δεδομένα για 95.277 πελάτες, όπως ο κωδικός πελάτη, φύλο, ηλικία, μοναδικός κωδικός για την καταγωγή του πελάτη, η πληροφορία για τον αν ο ασφαλισμένος είναι κάτοχος διπλώματος αυτοκινήτου, η πληροφορία για τον αν ο ασφαλισμένος έχει ήδη ασφάλιση οχήματος, η ηλικία του οχήματος, η πληροφορία για τον αν το όχημα του ασφαλισμένου είχε υποστεί κάποια ζημιά στο παρελθόν, το ετήσιο ασφάλιστρο για το συμβόλαιο υγείας, ο ανωνυμοποιημένος κωδικός για το κανάλι προσέγγισης του πελάτη (π.χ. μέσω τηλεφώνου, ταχυδρομείου κλπ.), ο χρόνος (σε ημέρες) που ο ασφαλισμένος είναι πελάτης στη συγκεκριμένη ασφαλιστική εταιρεία και τέλος εάν ο ασφαλισμένος ενδιαφέρεται ή όχι για ασφάλιση οχήματος (Πίνακας 4.15).

Η ασφαλιστική επιχείρηση, προκειμένου να εφαρμόσει τη διαδικασία της διασταυρούμενης πώλησης (cross-selling), έκανε τηλεφωνική επικοινωνία (μέση διάρκεια 20 λεπτά) με το σύνολο των πελατών της και διαπίστωσε ότι το ποσοστό των πελατών που προχώρησε στην αγορά ασφάλισης οχήματος ήταν 12,2% (Σχήμα 4.16). Η διαδικασία της προωθητικής επικοινωνίας απαίτησε την εργασία 18 υπαλλήλων για 12 μήνες και το κόστος της, συμπεριλαμβανομένων των τηλεπικοινωνιακών τελών, ήταν $450.000. Τα μικτά κέρδη που προέκυψαν από τη διαδικασία του cross-selling ήταν περίπου $600.000, ενώ τα αντίστοιχα καθαρά ανήλθαν σε $150.000.

**Πίνακας 4.15:** Σύντομη παρουσίαση των χαρακτηριστικών του συνόλου δεδομένων

| ΑΑ | Μεταβλητή | Περιγραφή | Ετικέτες δεδομένων | Είδος μεταβλητής |
|---|---|---|---|---|
| 1 | Customer ID | Μοναδικός κωδικός πελάτη | | Διακριτή |
| 2 | Gender | Φύλο | • 1=Άντρας<br>• 2=Γυναίκα | Διχοτομική |
| 3 | Age | Ηλικία | | Συνεχής |
| 4 | Driving License | Κατοχή διπλώματος οδήγησης | • 0=Όχι<br>• 1=Ναι | Διχοτομική |
| 5 | Region Code | Καταγωγή | | Διακριτή |
| 6 | Previously Insured | Ο ασφαλισμένος έχει ήδη ασφάλιση οχήματος | • 0=Όχι<br>• 1=Ναι | Διχοτομική |
| 7 | Vehicle age | Ηλικία οχήματος | • 1=<1 χρόνου<br>• 2=1-2 χρόνια<br>• 3=<2 χρόνια | Διατάξιμη |
| 8 | Vehicle damage | Το όχημα του ασφαλισμένου είχε υποστεί κάποια ζημιά στο παρελθόν | • 0= Όχι<br>• 1= Ναι | Διχοτομική |
| 9 | Annual premium | Ετήσιο ασφάλιστρο υγείας | | Συνεχής |
| 10 | Policy sales channel | Ο ανωνυμοποιημένος κωδικός για το κανάλι προσέγγισης του πελάτη | | Διακριτή |
| 11 | Vintage | Η χρονική διάρκεια που είναι πελάτης της ασφαλιστικής | | Διακριτή |
| 12 | Response | Ενδιαφέρον για ασφάλιση οχήματος | | Διχοτομική |

**Διερευνητική ανάλυση**

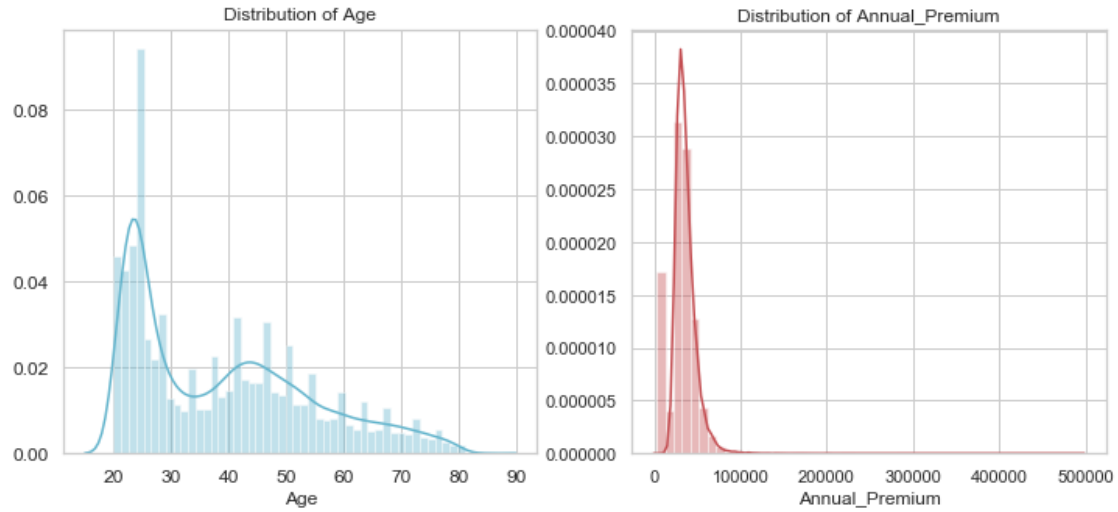
Από το σύνολο των 95.277 πελατών της ασφαλιστικής εταιρείας, το 53,9% (N= 51.360) είναι άνδρες και το υπόλοιπο 46,1% (N= 43.917) είναι γυναίκες, με τη μέση ηλικία των ασφαλισμένων να είναι ίση με τα 39 έτη. Η συντριπτική πλειοψηφία των ασφαλισμένων διαθέτουν άδεια οδήγησης (99,8%), με το 50,6% των ασφαλισμένων να έχει υποστεί ζημιά στο αυτοκίνητό τους κατά το παρελθόν. Σχετικά με το ετήσιο ασφάλιστρο που πληρώνουν για το συμβόλαιο υγείας που διαθέτουν, βρέθηκε πως ανέρχεται στα $30.684 κατά μέσο όρο, ενώ η μέση διάρκεια συνεργασίας των πελατών με την ασφαλιστική εταιρεία ισούται με τις 154 ημέρες (βλ. Πίνακες 4.16 & 4.17).

**Πίνακας 4.16:** Βασικά περιγραφικά μέτρα για τις ποσοτικές μεταβλητές του συνόλου δεδομένων

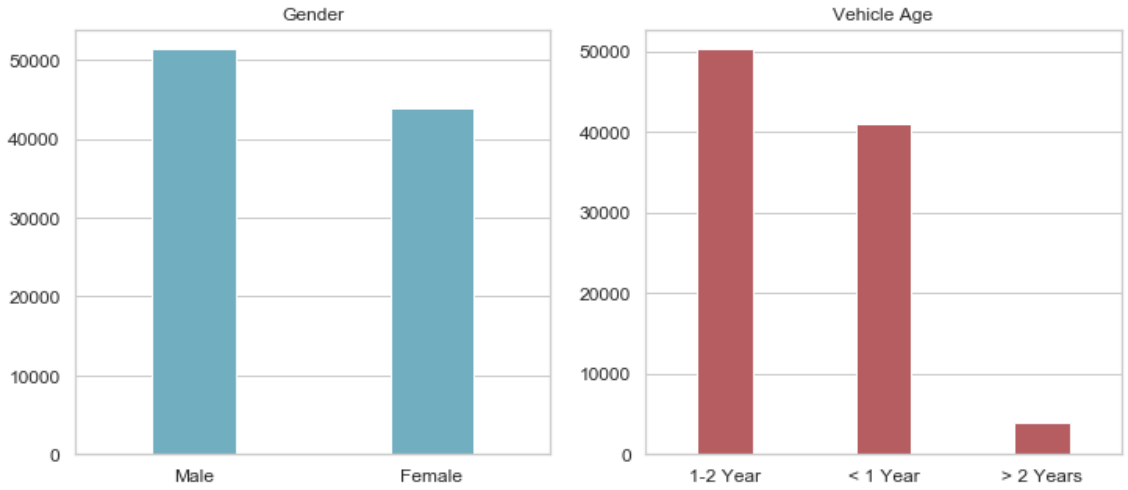
| | ***Age*** | ***Annual Premium*** | ***Vintage*** |
|---|---|---|---|
| ***count*** | 95277,0 | 95277,0 | 95277,0 |
| ***mean*** | 38,9 | 30684,4 | 154,0 |
| ***std*** | 15,6 | 17316,8 | 83,6 |
| ***min*** | 20,0 | 2630,0 | 10,0 |
| ***25%*** | 25,0 | 24470,0 | 82,0 |
| ***50%*** | 36,0 | 31695,0 | 153,0 |
| ***75%*** | 50,0 | 39505,0 | 227,0 |
| ***Max*** | 85,0 | 495106,0 | 299,0 |

**Πίνακας 4.17:** Βασικά περιγραφικά μέτρα για τις ποιοτικές μεταβλητές του συνόλου δεδομένων

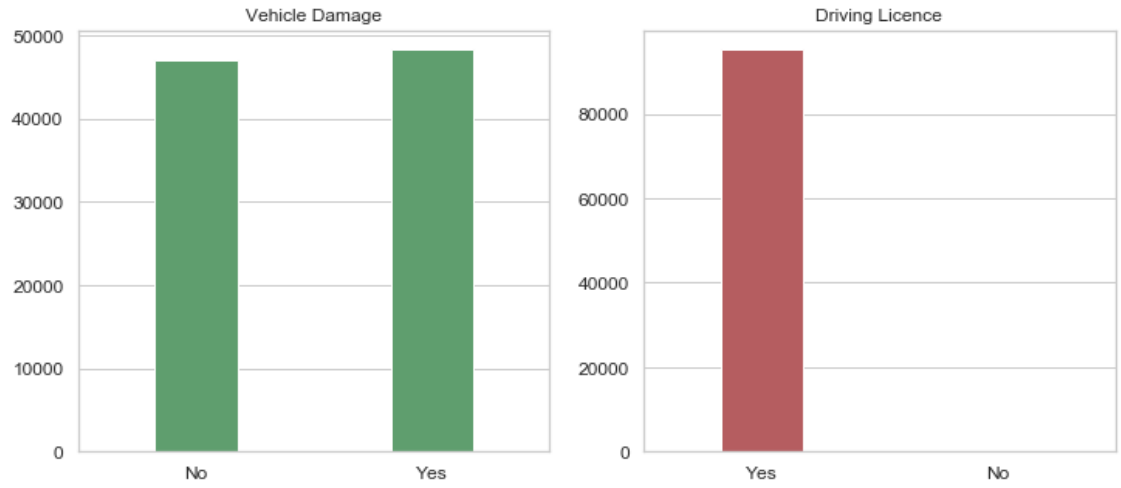
| | ***Gender*** | ***Vehicle Age*** | ***Vehicle Damage*** | ***Driving License*** | ***Region Code*** | ***Previously Insured*** | ***Response*** |
|---|---|---|---|---|---|---|---|
| ***count*** | 95277 | 95277 | 95277 | 95277 | 95277 | 95277 | 95277 |
| ***unique*** | 2 | 3 | 2 | 2 | 53 | 2 | 2 |
| ***mode*** | Male | 1-2 Year | Yes | Yes | 28 | No | No |
| ***N*** | 51360 | 50251 | 48214 | 95083 | 26755 | 51759 | 83639 |
| ***%*** | 53,90 | 52,74 | 50,60 | 99,79 | 28,08 | 54,32 | 87,78 |



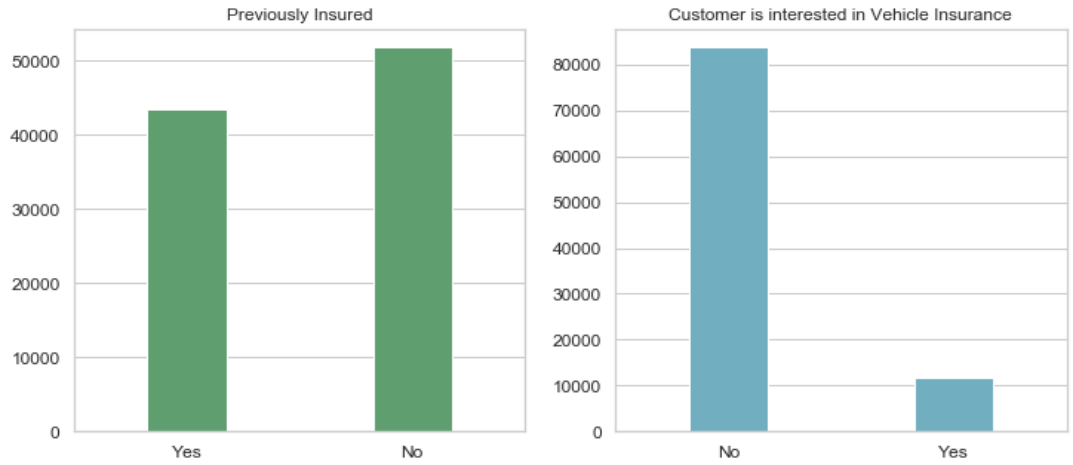

**Σχήμα 4.16**:Κατανομή της ηλικίας του ασφαλισμένου και του ετήσιου ασφάλιστρου για ασφάλιση υγείας που πληρώνει στη ασφαλιστική εταιρία




**Σχήμα 4.17**:Κατανομή του φύλου των ασφαλισμένων και της ηλικίας του αυτοκινήτου που έχουν στην κατοχή τους

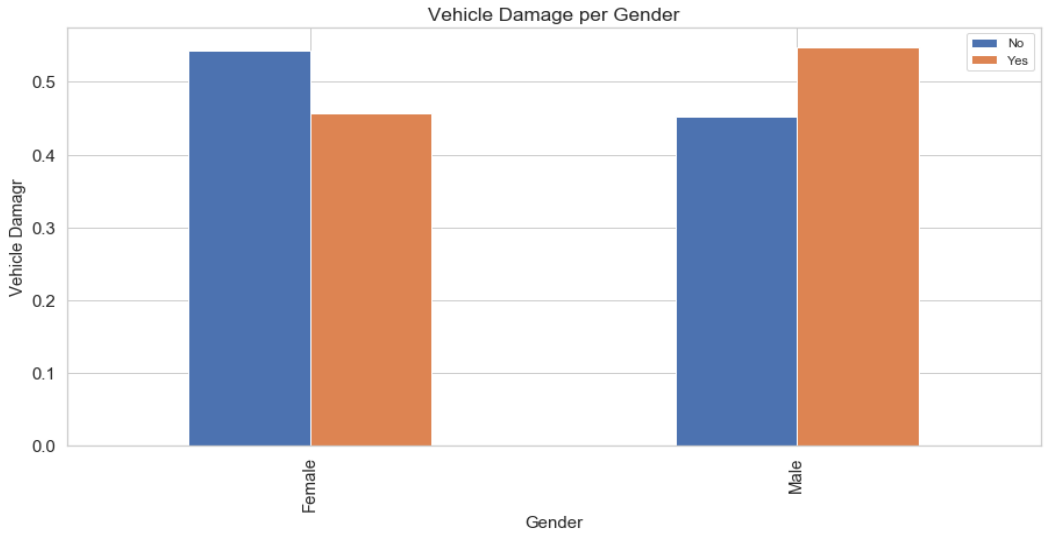


**Σχήμα 4.18**:Κατανομή των ασφαλισμένων που έχουν υποστεί ζημιά στο όχημά τους στο παρελθόν και των ασφαλισμένων που έχουν στην κατοχή τους άδεια οδήγησης



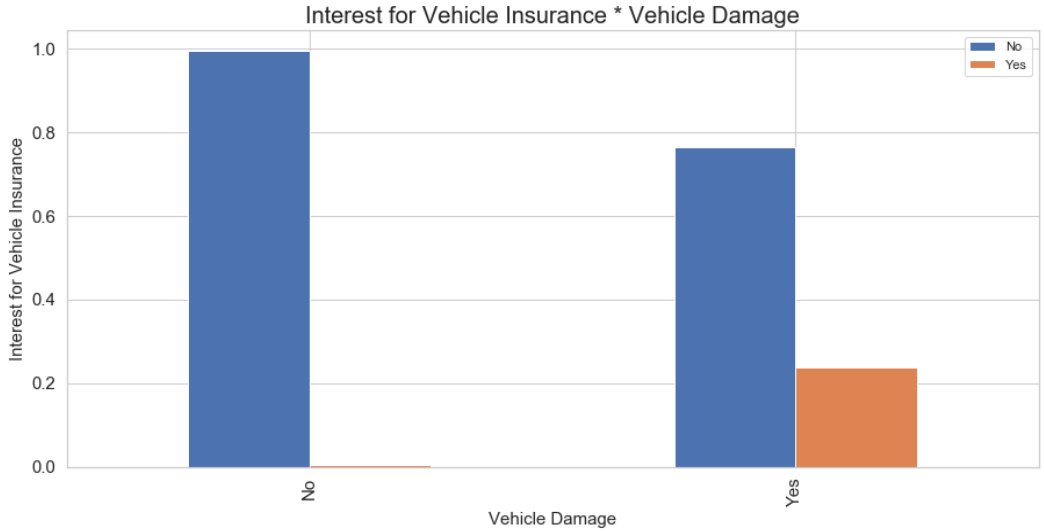

**Σχήμα 4.19**:Κατανομή των ασφαλισμένων που έχουν ασφαλίσει το όχημά τους στο παρελθόν και των ασφαλισμένων που προχώρησαν στην αγορά ασφάλισης οχήματος

Έπειτα από διερεύνηση με τη χρήση κατάλληλων στατιστικών ελέγχων, προέκυψε στατιστικά σημαντική διαφοροποίηση της κατανομής του φύλου των ασφαλισμένων σε σχέση με το αν έχουν υποστεί ζημία στο όχημά τους στο παρελθόν ($X^2$=778.79, p<0.05). Συγκεκριμένα, παρατηρήθηκε ότι η πλειοψηφία των αντρών με ποσοστό 54.78% (N=28137) έχουν υποστεί ζημιά στο όχημά τους στο παρελθόν, σε αντίθεση με τις γυναίκες που έχουν χαμηλότερο ποσοστό ίσο με 45.71% (N=20077) (Σχήμα 4.20).



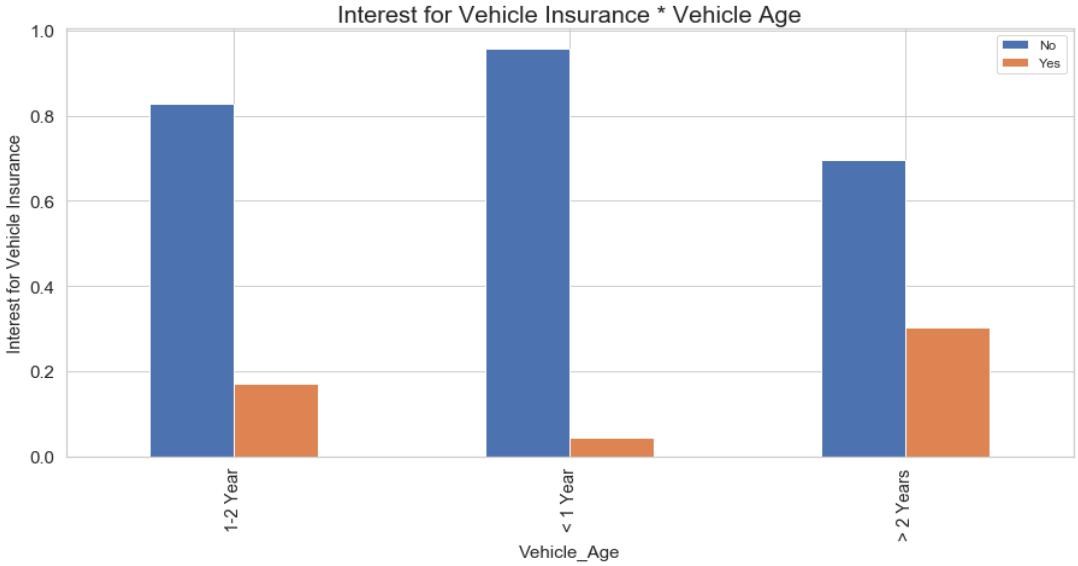

**Σχήμα 4.20**:Κατανομή του φύλου των ασφαλισμένων σε σχέση με το αν έχουν υποστεί ζημία στο όχημά τους στο παρελθόν

Επίσης, παρατηρήθηκε στατιστικά σημαντική διαφοροποίηση στην κατανομή των ασφαλισμένων που έχουν υποστεί ζημία στο όχημά τους στο παρελθόν σε σχέση με τους ασφαλισμένους που προχώρησαν στην αγορά ασφάλισης ($X^2$=11857, p<0.05). Πιο αναλυτικά, μόλις το 0.53% (N=246) των ασφαλισμένων που δεν έχουν υποστεί ζημία στο όχημά τους στο παραλθόν προχώρησαν σε αγορά ασφάλισης οχήματος σε αντίθεση με εκείνους που είχαν υποστεί ζημία, όπου ανταποκρίθηκαν θετικά με ποσοστό 23.63% (N=11392) (Σχήμα 4.21).




**Σχήμα 4.21**:Κατανομή των ασφαλισμένων που έχουν υποστεί ζημία στο όχημά τους στο παρελθόν σε σχέση με τους ασφαλισμένους που προχώρησαν στην αγορά ασφάλισης

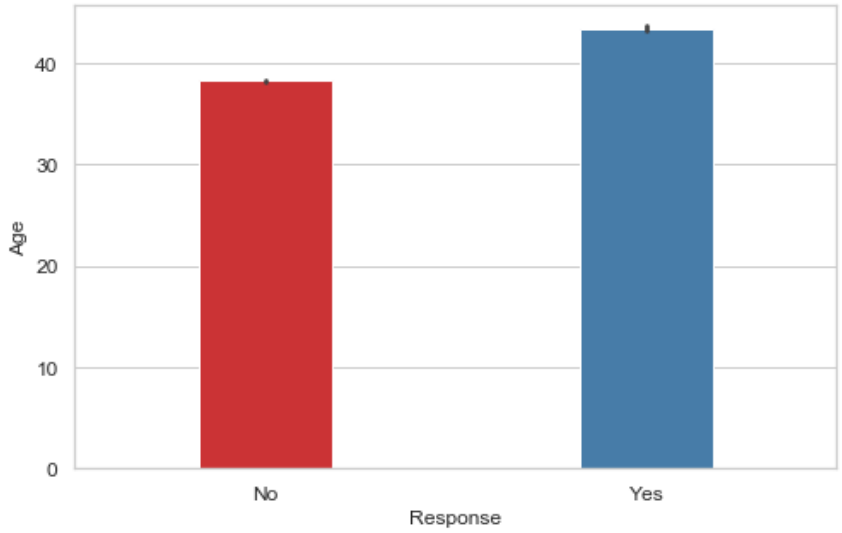
Επιπλέον, παρατηρήθηκε ότι η κατανομή της ηλικίας του οχήματος των ασφαλισμένων διαφοροποιείται σε σχέση με τους ασφαλισμένους που προχώρησαν στην αγορά ασφάλισης οχήματος ($X^2$=4750, p<0.05) με τα ποσοστά της θετικής ανταπόκρισης για αγορά ασφάλισης οχήματος να αυξάνονται όσο αυξάνεται και η ηλικία του οχήματος που έχουν στην κατοχή τους οι ασφαλισμένοι (Σχήμα 4.22).




**Σχήμα 4.22**:Κατανομή της ηλικίας του οχήματος σε σχέση με τους ασφαλισμένους που προχώρησαν στην αγορά ασφάλισης οχήματος

Τέλος, από διερευνητική ανάλυση προέκυψε ότι η κατανομή της ηλικίας των ασφαλισμένων διαφοροποιείται σε σχέση με το αν προχώρησαν στην αγορά ασφάλισης οχήματος (t=-33.86, p<0.05). Συγκεκριμένα, παρατηρήθηκε ότι η μέση ηλικία των ασφαλισμένων που προχώρησαν σε αγορά ασφάλισης οχήματος ήταν τα 43,4 έτη, κατά 5

χρόνια μεγαλύτερη δηλαδή από την αντίστοιχη εκείνων που δε προχώρησαν σε αγορά. (Σχήμα 4.23).




**Σχήμα 4.23**:Κατανομή της ηλικίας των ασφαλισμένων σε σχέση με το αν προχώρησαν στην αγορά ασφάλισης οχήματος

**Προεπεξεργασία των δεδομένων**

Διενεργήθηκε η τυποποίηση των ποσοτικών χαρακτηριστικών ώστε η κατανομή τους να ακολουθεί την κανονική με μέση τιμή 0 και τυπική απόκλιση 1. Δεν παρατηρήθηκαν ελλείπουσες τιμές στο σύνολο δεδομένων.

**Διενέργεια των μοντέλων μηχανικής μάθησης**

Ο αλγόριθμος εποπτευόμενης κατηγοριοποίησης που χρησιμοποιήθηκε με σκοπό τη δημιουργία του μοντέλου πρόβλεψης των πελατών που ενδιαφέρονται για ασφάλιση οχήματος είναι ο Extreme Gradient Boosting (XGBoost). Ωστόσο, το σύνολο δεδομένων δεν είναι ισορροπημένο με μόλις το 12,22% των περιπτώσεων να αφορά περιπτώσεις όπου οι πελάτες ανταποκρίθηκαν θετικά στην πώληση ασφάλισης οχήματος. Αυτό, δημιουργεί προβλήματα στην απόδοση των αλγορίθμων μηχανικής μάθησης, με την απόδοσή τους συνήθως να είναι καλύτερη στην εύρεση περιπτώσεων που αφορά την πλειοψηφική τάξη σε σχέση με την τάξη μειονότητας, καθώς στη δεύτερη δεν υπάρχει ικανοποιητική πληροφορία για την ανάλυση προτύπων. Μια πιθανή λύση σε αυτό δίνεται μέσω των τεχνικών επαναδειγματοληψίας.

Στον Πίνακα 4.18 παρουσιάζεται ο μέσος όρος των 10 επαναλήψεων της διασταυρούμενης επικύρωσης 10 τμημάτων με το Accuracy, το Precision, το Recall, το F-score και την τιμή AUC τόσο στο αρχικό σύνολο δεδομένων όσο και στο σύνολο δεδομένων στο οποίο έχει εφαρμοστεί η συνθετική τεχνική υπερδειγματοληψίας μειονότητας (Synthetic minority oversampling technique - SMOTE), για να επιτευχθεί μια πιο ισορροπημένη κατανομή των τάξεων. Η αξιολόγηση έγινε δοκιμάζοντας δυο (2) αναλογίες διαχωρισμού του συνόλου δεδομένων (70%-30%) και (80%-20%) και ο μέσος όρος για τα Precision, Recall και F-score υπολογίστηκε βάσει της κλάσης ενδιαφέροντος, που είναι η αγορά ασφάλισης οχήματος.

**Πίνακας 4.18:** Αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης «XGBoost»

| Split ratio | Sampling Technique | Accuracy (%) | Precision (%) | Recall (%) | F-score (%) | AUC (%) |
|---|---|---|---|---|---|---|
| (70% - 30%) | Original | 87,59 | 51,75 | 5,81 | 10,45 | 52,74 |
| | SMOTE | 77,53 | 30,81 | 69,89 | 42,77 | 74,23 |
| (80% - 20%) | Original | 87,78 | 49,91 | 6,21 | 11,04 | 52,54 |
| | SMOTE | 77,35 | 30,63 | 70,06 | 42,63 | 74,2 |

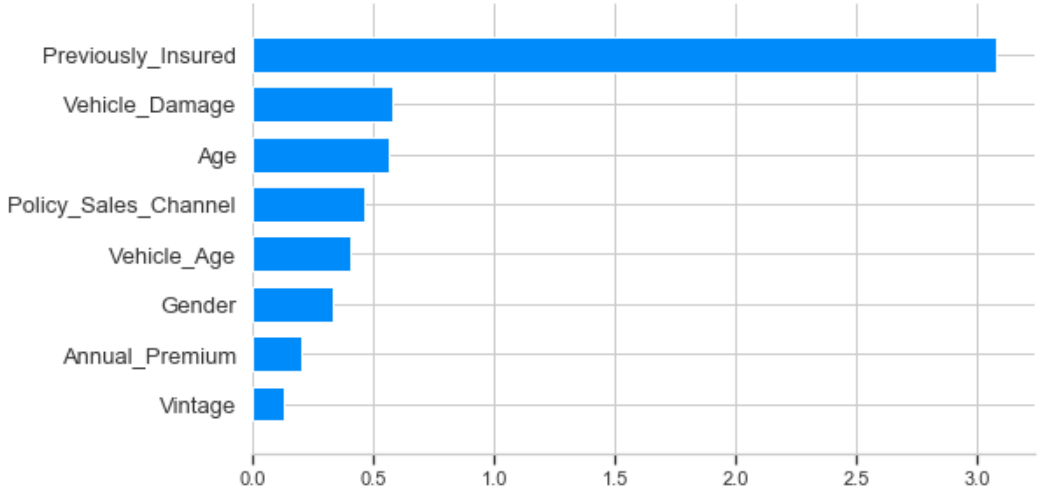
Σύμφωνα με τα αποτελέσματα του πίνακα 4.18, φαίνεται να λειτουργεί καλύτερα για το μοντέλο, ο διαχωρισμός του συνόλου δεδομένων με την αναλογία 70%-30%.

Η απόδοση του μοντέλου «XGBoost» χωρίς να εφαρμόσουμε κάποια τεχνική επαναδειγματοληψίας δε φαίνεται να είναι ικανοποιητική, καθώς σύμφωνα με την τιμή της ανάκλησης (Recall=5,81%) το μοντέλο καταφέρνει να αναγνωρίσει μόνο το 5,81% από το σύνολο των πελατών που ενδιαφέρονται για ασφάλιση οχήματος. Επίσης, από το σύνολο των περιπτώσεων που το μοντέλο ταξινόμησε στην τάξη ενδιαφέροντος, το 51,75% από αυτές ταξινομήθηκε ορθά. Ένα καλό μοντέλο μηχανικής μάθησης πρέπει να πετυχαίνει μια ικανοποιητική αναλογία στο f-score που είναι ο αρμονικός μέσος των precision και recall, ωστόσο στην περίπτωση αυτή το μοντέλο έχει πολύ χαμηλό f-score και ίσο με 10,45%, ένα ποσοστό αρκετά χαμηλό που δημιουργεί αμφιβολίες για τη χρησιμότητα του μοντέλου. Η συνολική ακρίβεια του μοντέλου είναι υψηλή (87,59%), ωστόσο δε μπορεί να ληφθεί υπόψιν σε περιπτώσεις που υπάρχει μη ισορροπημένο σύνολο δεδομένων καθώς δεν είναι αντιπροσωπευτική της απόδοσής του.

Η απόδοση του μοντέλου «XGBoost» εφαρμόζοντας τη συνθετική τεχνική υπερδειγματοληψίας μειονότητας (SMOTE), φαίνεται να παρουσιάζει σημαντική βελτίωση με το f-score να είναι ίσο με 42,77%, δηλαδή περίπου 4 φορές μεγαλύτερο σε σχέση με το μοντέλο που εκπαιδεύτηκε με το αρχικό σύνολο δεδομένων. Συγκεκριμένα, το μοντέλο αναγνωρίζει το 70,06% από το σύνολο των πελατών που ενδιαφέρονται για ασφάλιση οχήματος και ταξινομεί σωστά το 30,81% από το σύνολο των περιπτώσεων που ταξινομήθηκαν στην κλάση ενδιαφέροντος.

Επομένως το τελικό μοντέλο θα κατασκευαστεί διαχωρίζοντας το σύνολο δεδομένων σε 70% σύνολο εκπαίδευσης και 30% σύνολο δοκιμής και εφαρμόζοντας στο σύνολο δεδομένων εκπαίδευσης την τεχνική υπερδειγματοληψίας «SMOTE».

Από τη διαδικασία της επιλογής μεταβλητών, που διενεργήθηκε βάση της μεταβλητότητας που προσφέρουν στο μοντέλο, εξαιρέθηκε από το μοντέλο, η μεταβλητή «Κατοχή άδειας οδήγησης», όπως και ήταν φυσιολογικό καθώς το 99,79% των πελατών είχαν στην κατοχή τους άδεια και η χρήση της στην λήψη αποφάσεων από το μοντέλο θα ήταν ανούσια.




**Σχήμα 4.23**:Διακύμανση των χαρακτηριστικών στο μοντέλο

Επίσης, διενεργήθηκε η διαδικασία της εύρεσης βέλτιστων παραμέτρων για το μοντέλο «XGBoost» και στη συνέχεια κατασκευάστηκε το τελικό μοντέλο (Πίνακας 4.19).

**Πίνακας 4.19:** Αξιολόγηση της απόδοσης του μοντέλου ταξινόμησης XGBoost στο Training set και στο test set

| Set | Class | Accuracy (%) | Precision (%) | Recall (%) | F-score (%) | AUC (%) |
|---|---|---|---|---|---|---|
| Training | Not Interested | 95,83 | 96,00 | 78,00 | 86,00 | 78,78 |
| | Interested | | 34,00 | 79,60 | 48,00 | |
| Test | Not Interested | 77,00 | 95,00 | 78,00 | 84,00 | 74,90 |
| | Interested | | 31,00 | 72,00 | 43,00 | |

**Πίνακας 4.20:** Πίνακας ταξινόμησης των περιπτώσεων από το μοντέλο «XGBoost»

| | **Predicted** | | **Total** |
|---|---|---|---|
| **Actual** | **Interested** | **Not Interested** | |
| **Interested** | 2.468 | 966 | 3.434 |
| **Not Interested** | 5.616 | 19.534 | 25.150 |
| **Total** | 8.084 | 20.500 | 28.584 |

Το τελικό μοντέλο, ταξινομεί σωστά το 77% των ασφαλισμένων (Ορθότητα). Επίσης, αναγνωρίζει το 72% από το σύνολο των πελατών που ενδιαφέρονται για ασφάλιση οχήματος (Ανάκληση) και ταξινομεί σωστά το 31% από το σύνολο των περιπτώσεων που πραγματικά αγόρασαν ασφάλιση οχήματος (Ακρίβεια).
Συγκεκριμένα, όπως προκύπτει από τα παραπάνω, στο μέλλον η ασφαλιστική επιχείρηση, εφαρμόζοντας το μοντέλο που προέκυψε από την εφαρμογή του αλγόριθμου XGBoost, θα είχε τα ακόλουθα αποτελέσματα:

- Θα έκανε τελικά 8.084 τηλεφωνικές πωλήσεις, δηλαδή μόλις το 28,28% του συνόλου των επικοινωνιών που έγιναν κατά την αρχική διαδικασία των τηλεφωνικών επικοινωνιών

- Το ποσοστό των επικοινωνιών που θα κατέληγε σε πώληση (conversion rate) θα ήταν το 31%, δηλαδή η ασφαλιστική εταιρεία συμβουλευόμενη το μοντέλο για τη διενέργεια προωθητικών ενεργειών θα έπρεπε να κάνει τηλεφωνική επικοινωνία με 8084 πελάτες κάτι που θα είχε κόστος 127.260$. Ωστόσο το μοντέλο αναγνωρίζει σωστά το 72% (2.468 πελάτες) από το σύνολο των πελατών που θα προχωρήσουν σε ασφάλιση οχήματος (3.434 πελάτες), μειώνοντας τα μικτά κέρδη από 600.000$ σε 431.217$, αυξάνοντας όμως τα καθαρά κέρδη από 150.000$ σε (431.217$ - 127.260$) 303.957$.

Αυτό σημαίνει ότι αν η ασφαλιστική εταιρία χρησιμοποιούσε το συγκεκριμένο μοντέλο θα είχε αύξηση στα καθαρά της κέρδη κατά σχεδόν 103% (από 150.000$ σε 303.957$).

# ΚΕΦΑΛΑΙΟ 5

## 5. Συμπεράσματα

Στην παρούσα διπλωματική εργασία έγινε αναφορά στη σημαντικότητα της υιοθέτησης της αναλυτικής των δεδομένων και της μηχανικής μάθησης στην ασφαλιστική «βιομηχανία» και συγκεκριμένα στον κλάδο ασφάλισης υγείας. Στη συνέχεια, έγινε αναφορά σε αλγόριθμους της μηχανικής μάθησης, αναλύθηκε ο τρόπος λειτουργίας τους και μερικοί εξ' αυτών εφαρμόστηκαν σε μελέτες περιπτώσεων για να προσδώσουν λύσεις και κατευθύνσεις σε τρία (3) βασικά προβλήματα που αντιμετωπίζει κάθε ασφαλιστική εταιρεία και αφορούν νευραλγικούς επιχειρηματικούς τομείς της. Οι τεχνικές που χρησιμοποιήθηκαν βασίζονται σε αλγορίθμους παλινδρόμησης και κατηγοριοποίησης με τις σύγχρονες επεκτάσεις τους.

Η πρώτη μελέτη περίπτωσης αφορά τη διαχείριση των ασφαλιστικών απαιτήσεων που αποτελούν το σημαντικότερο μέρος του κύκλου ζωής της ασφάλισης και ο σκοπός ήταν η δημιουργία ενός μοντέλου για τη πρόβλεψη του ύψους των ασφαλιστικών απαιτήσεων μιας ασφαλιστικής εταιρείας. Το ύψος των ασφαλιστικών απαιτήσεων είναι μία ποσοτική μεταβλητή, επομένως για την πρόβλεψή της χρησιμοποιήθηκαν τεχνικές παλινδρόμησης. Κατά τη διαδικασία αξιολόγησης και σύγκρισης των αλγορίθμων, πλέων κατάλληλος κρίθηκε ο αλγόριθμος Gradient Boosting ο οποίος βασίζεται σε δέντρα αποφάσεων και αποτελεί μια νέα ισχυρή τεχνική συνδυασμού μοντέλων μηχανικής μάθησης τα οποία συνδυάζονται με στόχο την κατασκευή ενός ισχυρότερου μοντέλου. Το πλεονέκτημα του αλγορίθμου Gradient Bosting είναι ότι διαθέτει ένα σύνολο παραμέτρων των οποίων η παραμετροποίηση μπορεί να βελτιστοποιήσει την απόδοση του. Επομένως, το τελικό μοντέλο δημιουργήθηκε έπειτα από μια διαδικασία βελτιστοποίησης των παραμέτρων κατά την οποία διενεργήθηκε διασταυρούμενη επικύρωση 5 τμημάτων (Cross validation 5-fold) και επιλέχθηκε ο συνδιασμός των παραμέτρων που ελαχιστοποιούσε το μέσο απόλυτο σφάλμα, ενώ παράλληλα απέτρεπε τον κίνδυνο υπερπροσαρμογής. Τέλος, οι ανεξάρτητες μεταβλητές, επιλέχθηκαν βάσει της συνεισφοράς τους στη διακύμανση του μοντέλου ερμηνεύοντας το 87.2% της συνολικής μεταβλητότητας του. Αυτό το ποσοστό είναι ιδιαίτερα υψηλό και υποδηλώνει την εξαιρετική ακρίβεια του μοντέλου, καθώς και τον αντίκτυπο που μπορεί να έχει κατά τη διαδικασία λήψης αποφάσεων.

Η δεύτερη μελέτη περίπτωσης, αφορά τη διαδικασία διατήρησης των πελατών, ο οποίος αποτελεί βασικό πυλώνα της οικονομικής βιωσιμότητας κάθε ασφαλιστικής εταιρείας και ο σκοπός ήταν η δημιουργία ενός προγνωστικού μοντέλου για την πρόβλεψη των πελατών που πρόκειται να αποχωρήσουν από το ασφαλιστήριο συμβόλαιο. Αυτό το προγνωστικό μοντέλο, κατασκευάστηκε χρησιμοποιώντας ιστορικά δεδομένα πελατών στα οποία συμπεριλαμβανόταν και η πληροφορία για το πότε κάποιος πελάτης ανανέωσε ή διέκοψε το συμβόλαιό του, επομένως για την πρόβλεψή της αποχώρησης των πελατών χρησιμοποιήθηκαν τεχνικές ταξινόμησης. Η πρόκληση σε τέτοιου είδους προβλήματα είναι η διαχείριση των ανομοιογενών κατανομών. Στα δεδομένα που χρησιμοποιήθηκαν, υπήρχε εξαιρετικά ανισομερής κατανομή των κλάσεων της μεταβλητής στόχος «αποχώρηση από το ασφαλιστήριο συμβόλαιο» με το ποσοστό των πελατών που έχουν αποχωρήσει από το ασφαλιστήριο

συμβόλαιο να είναι μόλις το 11.7%, δημιουργώντας έτσι προβλήματα κατά την εκπαίδευση των δεδομένων από τους αλγορίθμους μηχανικής μάθησης. Από τη βιβλιογραφία στο πρόβλημα της ανομοιογένειας των κατανομών των τάξεων της μεταβλητής στόχος, έρχονται να δώσουν λύσεις τεχνικές επαναδειγματοληψίας των δεδομένων. Συγκεκριμένα, κατά τη διαδικασία αξιολόγησης των μοντέλων ταξινόμησης εφαρμόστηκε και η συνθετική τεχνική υπερδειγματοληψίας μειονότητας (SMOTE), για να επιτευχθεί μια πιο ισορροπημένη κατανομή των τάξεων και να ελεγχθεί αν αυτό βελτιώνει την απόδοση των αλγορίθμων. Τα αποτελέσματα έδειξαν ότι η τεχνική υπερδειγματοληψίας (SMOTE) δεν βελτίωσε σημαντικά την απόδοση όλων των αλγορίθμων, παρά μόνο των αλγορίθμων «Κ κοντινότεροι γείτονες» και «Extreme Gradient Boosting (XGBoost)» με τον δεύτερο να επιλέχθηκε ως το πιο ισχυρό μοντέλο από τα υπόλοιπα. Για την κατασκευή του τελικού μοντέλου δε χρησιμοποιήθηκαν όλες οι μεταβλητές, παρά μόνο εκείνες που προσέφεραν κάποια μεταβλητότητα στο τελικό μοντέλο για το οποίο η επιλογή των βέλτιστων παραμέτρων έγινε βάσει με έμφαση στη μεγιστοποίηση του αρμονικού μέσου (f-score) μεταξύ των precision και recall, τα οποία είναι τα πιο κρίσιμα μέτρα αξιολόγησης των μοντέλων που σχετίζονται με προβλήματα πρόβλεψης αποχώρησης πελατών. Το τελικό μοντέλο εμφάνισε εξαιρετικές επιδόσεις στην αρνητική κλάση ταξινομώντας ορθώς το 93% από το σύνολο των περιπτώσεων που δεν αποχώρησαν από το ασφαλιστήριο συμβόλαιο, ενώ στην θετική κλάση εμφάνισε αρκετά χαμηλότερες αλλά ικανοποιητικές επιδόσεις αναγνωρίζοντας το 69% από τους ασφαλισμένους που στην πραγματικότητα αποχώρησαν, το οποίο είναι ένα ποσοστό αρκετά ικανοποιητικό και σίγουρα αναδυκνύει τη σημαντικότητα του μοντέλου κατά τη διαδικασία της λήψης αποφάσεων.

Η τρίτη μελέτη περίπτωσης, αφορά τη διαδικασία της διασταυρούμενης πώλησης (cross-selling), δηλαδή την πώληση συμπληρωματικών προϊόντων σε υπάρχοντες πελάτες. Σκοπός ήταν η διαμόρφωση κατάλληλου μοντέλου, βάσει των διαθέσιμων δεδομένων, ούτως ώστε μελλοντικά το μοντέλο αυτό να υποδεικνύει αυτόματα τους υπάρχοντες πελάτες (κατόχους ασφάλισης υγείας) που έχουν μεγαλύτερη πιθανότητα να ενδιαφέρονται για αγορά Ασφάλισης οχήματος. Για την κατασκευή του μοντέλου χρησιμοποιήθηκε ο αλγόριθμος εποπτευόμενης κατηγοριοποίησης «Extreme Gradient Boosting (XGBoost)» λόγω της υψηλής απόδοσής του, ωστόσο το σύνολο δεδομένων δεν ήταν ισορροπημένο με μόλις το 12,22% των περιπτώσεων να αφορά περιπτώσεις όπου οι πελάτες ανταποκρίθηκαν θετικά στην πώληση ασφάλισης οχήματος, κάτι που επηρεάζει αρνητικά την απόδοση του αλγορίθμου, ιδιαίτερα στην πρόβλεψη της μειοψηφικής τάξης. Επομένως, κατά τη διαδικασία αξιολόγησης του μοντέλου, με τη διαδικασία της διασταυρούμενης επικύρωσης 10 τμημάτων, διενεργήθηκε η αξιολόγηση τόσο με το αρχικό σύνολο δεδομένων όσο και με τα δεδομένα έπειτα από την εφαρμογή της συνθετικής τεχνικής υπερδειγματοληψίας μειονότητας (SMOTE) και τα αποτελέσματα έδειξαν ότι η δεύτερη εμφάνισε αισθητά καλύτερα αποτελέσματα. Συγκεκριμένα, η ευαισθησία (recall) του μοντέλου με την τεχνική SMOTE ήταν 11 φορές μεγαλύτερη και ο αρμονικός μέσος όρος της ευαισθησίας και της ακρίβειας (f-score) ήταν πάνω από 3 φορές μεγαλύτερος σε σχέση με τα αντίστοιχα αποτελέσματα του μοντέλου που εφαρμόστηκε με το αρχικό σύνολο δεδομένων. Το τελικό μοντέλο προέκυψε έπειτα από την επιλογή μεταβλητών, βάσει της προσφοράς τους στη συνολική διακύμανση του και το συνολικό ποσοστό ορθής ταξινόμησης που προέκυψε ήταν 77%. Το μοντέλο αναγνώριζε το 72% από το σύνολο των πελατών που ενδιαφέρονταν για ασφάλιση οχήματος, όμως ταξινομούσε σωστά μόλις το 31% από το σύνολο των περιπτώσεων που πραγματικά αγόρασαν ασφάλιση οχήματος (Ακρίβεια). Ωστόσο, αυτό δεν επηρέασε τη σημαντικότητα της εκμετάλλευσής του, καθώς όπως προέκυψε αν η ασφαλιστική εταιρία χρησιμοποιούσε το συγκεκριμένο μοντέλο θα είχε αύξηση στα καθαρά της κέρδη κατά σχεδόν 103%.

Οι παραπάνω εφαρμογές επιβεβαιώνουν τη βιβλιογραφία που αναφέρει ότι οι σύγχρονες τεχνικές μηχανικής μάθησης όπως είναι ο αλγόριθμος Gradient Boosting για την παλινδρόμηση και ο αλγόριθμος Extreme Gradient Boosting για την εποπτευόμενη κατηγοριοποίηση, εμφανίζουν καλύτερες επιδόσεις από τα παραδοσιακά μοντέλα. Να αναφερθεί ότι θα είχε ιδιαίτερο ενδιαφέρον η προσπάθεια επίλυσης τέτοιων προβλημάτων με τη χρήση νευρωνικών δικτύων.

Συνοψίζοντας, επιβεβαιώνεται από τα παραπάνω ότι η χρήση μεθόδων της μηχανικής μάθησης και της αναλυτικής των δεδομένων μπορεί να έχει τεράστια επιρροή σε νευραλγικούς επιχειρηματικούς τομείς των ασφαλιστικών εταιρειών αποτελώντας σημαντικό γνώμονα για τη λήψη αποφάσεων, με τις ασφαλιστικές εταιρείες να λαμβάνουν υπόψιν όλους του παράγοντες που μπορούν να προκαθορίσουν ένα γεγονός. Προβλέποντας το μέλλον οι ασφαλιστικές εταιρείες, μπορούν να αποτρέπουν κινδύνους και να δημιουργούν νέα έσοδα, βελτιώνοντας την κερδοφορία τους και αναβαθμίζοντας την θέση τους σε μία εξαιρετικά ανταγωνιστική αγορά. Να τονιστεί ότι η μείωση της ζημίας των ασφαλιστικών εταιρειών έχει διπλό αντίκτυπο, τόσο στην ίδια, αυξάνοντας τα κέρδη της όσο και στους ασφαλισμένους καθώς η ζημία περνάει μέσα από τα ασφαλιστήρια συμβόλαια άρα επηρεάζει και τα οικονομικά του ίδιου του ασφαλισμένου.

# Παραρτήματα

## Π1 Πηγαίος κώδικας σε Python για την 1η Εφαρμογή

```
#Συνάρτηση για αξιολόγηση των μοντέλων παλινδρόμησης

def evaluate(true, predicted):
    MAE = metrics.mean_absolute_error(true, predicted)
    #MSE = metrics.mean_squared_error(true, predicted)
    RMSE = np.sqrt(metrics.mean_squared_error(true, predicted))
    Adjusted_R_square_ = 1-((1-metrics.r2_score(true, predicted))*((len(X_test)-1)/(len(X_test)-len(X_test[0])-1)))
    return MAE, RMSE, Adjusted_R_square_

#Συνάρτηση για αξιολόγηση των μοντέλων παλινδρόμησης με τη χρήση διασταυρούμενης επικύρωσης 10 τμημάτων

def cross_val(model):
    cv = KFold(n_splits=10, random_state=42, shuffle=True)
    pred1 = abs(cross_val_score(model, X, y, scoring='neg_mean_absolute_error',cv=cv, n_jobs=-1))
    pred2 = abs(cross_val_score(model, X, y, scoring='neg_root_mean_squared_error',cv=cv, n_jobs=-1))
    pred3 = abs(cross_val_score(model, X, y, scoring='r2',cv=cv, n_jobs=-1))
    pred4 = abs(cross_val_score(model, X, y, scoring='explained_variance',cv=cv, n_jobs=-1))
    gbr.score(X_test, y_test)
    return pred1.mean(),pred2.mean(),pred3.mean(),pred4.mean()

#Εισαγωγή του αρχείου excel με τα δεδομένα

df = pd.read_excel(r'df.xlsx')

#Μερική προεπεξεργασία δεδομένων
column_names=["age","sex","region","children","steps","bmi","bloodpressure","smoker",'charges']
df=df.reindex(columns=column_names)
df["sex"]=df["sex"].map({0:"male",1:"female"})
df["smoker"]=df["smoker"].map({0:"No",1:"Yes"})
df["region"]=df["region"].map({0:"northeast",1:"northwest",2:"southeast",3:"southwest"})

#Διερευνητική ανάλυση
df.stb.freq(['smoker'])
table=pd.DataFrame(df,columns=["age","children","steps","bmi","bloodpressure","charges"])
```

#Γραφήματα

```
plt.figure(figsize=(10,5))
corr = table.corr()
mask = np.triu(np.ones_like(corr, dtype=bool))
sns.heatmap(corr, mask = mask, annot=True, cmap='Dark2');

fig, ax = plt.subplots(figsize=(8,6))
sns.heatmap(table.corr(method="spearman"), annot=True, fmt='.1g', cmap="Blues", cbar=True);

f, ax = plt.subplots(1, 1, figsize=(8, 5))
ax = sns.barplot(x='sex', y='charges', data=df, palette='Set1')
ax.set_ylabel('Charges', fontsize = 12.0) # Y label
ax.set_xlabel('Sex', fontsize = 12) # X label
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )
plt.show()

claim = df['charges'].groupby(df.sex).median().sort_values(ascending = True)
f, ax = plt.subplots(1, 1, figsize=(8, 4))
ax = sns.barplot(claim.head(), claim.head().index, palette='Set1')
ax.set_ylabel('Charges', fontsize = 12.0) # Y label
ax.set_xlabel('Sex', fontsize = 12) # X label
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )
plt.show()

f, ax = plt.subplots(1, 1, figsize=(8, 5))
ax = sns.barplot(x='smoker', y='charges', data=df, palette='Set1')
ax.set_ylabel('Charges', fontsize = 12.0)
ax.set_xlabel('Smoker', fontsize = 12)
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )
plt.show()

f, ax = plt.subplots(1, 1, figsize=(8, 5))
ax = sns.barplot(x='region', y='charges', data=df, palette='Set1')
ax.set_ylabel('Charges', fontsize = 12.0) # Y label
ax.set_xlabel('Region', fontsize = 12) # X label
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )
plt.show()

claim = df['charges'].groupby(df.region).median().sort_values(ascending = True)
f, ax = plt.subplots(1, 1, figsize=(8, 4))
ax = sns.barplot(claim.head(), claim.head().index, palette='Set1')
ax.set_ylabel('Region', fontsize = 12.0) # Y label
```

```
ax.set_xlabel('Charges', fontsize = 12) # X label
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )
plt.show()

ax = sns.lmplot(x = 'bmi', y = 'charges', data=df, hue='smoker', palette='Set1')
sns.lmplot(x="bmi", y="charges", col="smoker", data=df)
ax = sns.lmplot(x = 'bloodpressure', y = 'charges', data=df, hue='smoker', palette='Set1')
f, ax = plt.subplots(1, 1, figsize=(8, 5))
ax = sns.barplot(x='region', y='charges', hue='sex', data=df, palette='Set1')

#Ελεγχος κανονικότητας των ποσοτικών μεταβλητών
for col in table:
   print(f"ks test for {col}")
   print(stats.kstest(table[col], 'norm'))
   print("============================")

#Άλλες συσχετίσεις που ελέγχθηκαν με το κριτήριο του Spearman

group_by_smoker = df.groupby(['smoker'])
group_by_smoker_N=group_by_smoker.get_group('No')
group_by_smoker_Y=group_by_smoker.get_group('Yes')

coef, p = spearmanr(group_by_smoker_N.charges, group_by_smoker_N.bloodpressure)
print('Spearmans correlation coefficient: %.3f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
   print('Variables are uncorrelated p=%.3f' % p)
else:
   print('Variables are correlated (p=%.3f)' % p)

coef, p = spearmanr(group_by_smoker_Y.charges, group_by_smoker_Y.bloodpressure)
print('Spearmans correlation coefficient: %.3f' % coef)
# interpret the significance
alpha = 0.05
if p > alpha:
   print('Variables are uncorrelated p=%.3f' % p)
else:
   print('Variables are correlated (p=%.3f)' % p)

#Προεπεξεργασία δεδομένων

#Ελεγχος ελλειπουσών τιμών

df.isnull().any()
plt.figure(figsize = (9, 3))
missings = df.isnull().sum() / len(df)
```

```
missings.plot.bar()
plt.axhline(0.4, color = 'r')
plt.show()
print("The total number of rows in my dataset is:",len(df))

#----------See the percentage of missing----------
print("Percentage of missing values")
percent_missing = df.isnull().sum()
num_missing = df.isnull().sum()* 100 / len(df)
missing_value_df = pd.DataFrame({
                        'Number of missing': num_missing,
                        'Percent of missing': percent_missing})


#Δημιουργία ψευδομεταβλητών

df1=pd.get_dummies(df)
df1=df1.drop(columns=['age', 'children', 'steps', 'bmi', 'bloodpressure', 'charges',
     'sex_female', 'sex_male','smoker_No', 'smoker_Yes'])
df["region_northeast"]=df1["region_northeast"]
df["region_northwest"]=df1["region_northwest"]
df["region_southeast"]=df1["region_southeast"]
df["region_southwest"]=df1["region_southwest"]

df=df.drop(columns=["region"])

#Κωδικοποίηση μεταβλητών

df["sex"]=df["sex"].map({"male":0,"female":1})
df["smoker"]=df["smoker"].map({"No":0,"Yes":1})
#df["region"]=df["region"].map({"southeast":0,"southwest":1,"northeast":2,"northwest":3})

df=df.drop(columns=["region_northeast",'region_northwest','region_southeast','region_south
west'])

#Διαχωρισμός των δεδομένων σε train και test set

X=df.drop(columns=["charges"],axis=1)
y=df["charges"]


#Επαναδιάταξη των μεταβλητών

column_names=["age","smoker","bloodpressure","sex","children","steps","bmi"]
X=X.reindex(columns=column_names)

#Διαχωρισμός των δεδομένων σε train 80% και test set 20%
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```
#Κανονικοποίηση των μεταβλητών με τη χρήση της τυποποιημένης κανονικής κατανομής
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

#Αξιολόγηση των μοντέλων παλινδρόμησης με τη χρήση διασταυρούμενης επικύρωσης 10 τμημάτων

#LinearRegression
lr = LinearRegression()
lr.fit(X_train, y_train)
lr_metrics=cross_val(lr)

#Make the predictions
#y_pred_lr = lr.predict(X_test)
#train_pred_lr=lr.predict(X_train)
#Check the score
#lr.score(X_test,y_test)
#print_evaluate(y_test,y_pred_lr)
#params=np.append(lr.intercept_,lr.coef_)

results_df = pd.DataFrame(data=[["Linear Regression", lr_metrics[0]
,lr_metrics[1],lr_metrics[2]]],
                          columns=['Model','MAE','RMSE','R squared'])
results_df

#Lasso Regression
lasso_reg=linear_model.Lasso(alpha=0.2, max_iter=1000,tol=0.1)
lasso_reg.fit(X_train,y_train)
lasso_metrics=cross_val(lasso_reg)

results_df1 = pd.DataFrame(data=[["Lasso Regression", lasso_metrics[0]
,lasso_metrics[1],lasso_metrics[2]]],
                          columns=['Model','MAE','RMSE','R squared'])
results_df1=results_df.append(results_df1,ignore_index=True)

#Decision Tree Regressor
tree = DecisionTreeRegressor()
tree.fit(X_train,y_train)

tree_metrics=cross_val(tree)
tree_metrics

results_df_2 = pd.DataFrame(data=[["Decision Tree Regression", tree_metrics[0]
,tree_metrics[1],tree_metrics[2]]],
                          columns=['Model','MAE','RMSE','R squared'])
```

```
results_df2=results_df1.append(results_df_2,ignore_index=True)
results_df2

#Random Forest Regressor
rf = RandomForestRegressor(n_estimators = 1000,max_depth=6,random_state=1234)
rf.fit(X_train,y_train)

rf_metrics=cross_val(rf)
rf_metrics

results_df_3 = pd.DataFrame(data=[["Random Forest Regression", rf_metrics[0]
,rf_metrics[1],rf_metrics[2]]],
                    columns=['Model','MAE','RMSE','R squared'])
results_df3=results_df2.append(results_df_3,ignore_index=True)
results_df3

#Gradient Boosting Regressor
gbr_params={'n_estimators':1000,
        'max_depth':3,
        'min_samples_split':5,
        'learning_rate':0.01,
        'loss':'ls'}
gbr=GradientBoostingRegressor()
gbr.fit(X_train, y_train)

gbr_metrics=cross_val(gbr)
gbr_metrics

#Τελικός πίνακας αξιολόγησης των μοντέλων

results_df_4 = pd.DataFrame(data=[["Gradient Boosting Regression", gbr_metrics[0]
,gbr_metrics[1],gbr_metrics[2]]],
                    columns=['Model','MAE','RMSE','R squared'])
results_df4=results_df3.append(results_df_4,ignore_index=True)

#Μοντέλο Gradient Boosting Regressor

gbr_params={'n_estimators':1000,
        'max_depth':4,
        'learning_rate':0.01,
        'min_samples_split':4,
        'loss':'ls',
        "random_state":1234}

gbr=GradientBoostingRegressor(**gbr_params)
gbr.fit(X_train, y_train)
```

```
#Επιλογή μεταβλητών

def select_features(X_train, y_train, X_test):
    # configure to select all features
    fs = SelectKBest(score_func=mutual_info_regression, k='all')
    # learn relationship from training data
    fs.fit(X_train, y_train)
    # transform train input data
    X_train_fs = fs.transform(X_train)
    # transform test input data
    X_test_fs = fs.transform(X_test)
    return X_train_fs, X_test_fs, fs

X_train_fs, X_test_fs, fs = select_features(X_train, y_train, X_test)

#Σημαντικότητα μεταβλητών στο μοντέλο

for i in range(len(fs.scores_)):
    print('Feature %d: %f' % (i, fs.scores_[i]))

# feature selection
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import mutual_info_regression
def select_features(X_train, y_train, X_test):
    # configure to select all features
    fs = SelectKBest(score_func=mutual_info_regression, k=7)
    # learn relationship from training data
    fs.fit(X_train, y_train)
    # transform train input data
    X_train_fs = fs.transform(X_train)
    # transform test input data
    X_test_fs = fs.transform(X_test)
    return X_train_fs, X_test_fs, fs

#Εκπαίδευση του μοντέλου με τις επιλεγμένες μεταβλητές
X_train_fs, X_test_fs, fs = select_features(X_train, y_train, X_test)
gbr=GradientBoostingRegressor(**gbr_params)
gbr.fit(X_train_fs, y_train)

# evaluate the model
yhat = gbr.predict(X_test_fs)
# evaluate predictions
print_evaluate(y_test, yhat)

gbr.feature_importances_

#Σχήμα 4.5
a11=(2336.8997968734334, 4590.246450469801, 0.8637308078558545)
```

```
a10=(2321.3229242649345, 4554.082504384601, 0.8658695248892473)
a9=(2241.0837733022036, 4493.541296985641, 0.8694120365363418)
a8=(2216.9695886269783, 4491.560537299132, 0.8695271378993339)
a7=(2264.3015209417745, 4571.157368485195, 0.8648618342607739)
a6=(3734.5229641908204, 6432.625887480604, 0.732390046407665)
a5=(4151.276363671282, 7153.205702406919, 0.6690768858906319)
a4=(4324.276581269162, 7368.56408306218, 0.6488510257923692)
a3=(4345.742594884085, 7433.788905710592, 0.642606933152698)
a2=(4722.316257831281, 7439.942720980969, 0.6420149762753298)
a1=(9401.958250284051, 12370.646832374166, 0.010284227716489402)

x1=[a1[0],a1[1],a1[2]]
x2=[a2[0],a2[1],a2[2]]
x3=[a3[0],a3[1],a3[2]]
x4=[a4[0],a4[1],a4[2]]
x5=[a5[0],a5[1],a5[2]]
x6=[a6[0],a6[1],a6[2]]
x7=[a7[0],a7[1],a7[2]]
x8=[a8[0],a8[1],a8[2]]
x9=[a9[0],a9[1],a9[2]]
x10=[a10[0],a10[1],a10[2]]
x11=[a11[0],a11[1],a11[2]]

df = pd.DataFrame({
   'MAE': [a1[0], a2[0], a3[0], a4[0], a5[0], a6[0], a7[0], a8[0], a9[0],a10[0],a11[0]],
   'RMSE': [a1[1], a2[1], a3[1], a4[1], a5[1], a6[1], a7[1], a8[1], a9[1],a10[1],a11[1]],
   'Adjusted R2': [a1[2], a2[2], a3[2], a4[2], a5[2], a6[2], a7[2], a8[2], a9[2],a10[2],a11[2]],
   }, index=[1, 2, 3, 4, 5,6,7,8,9,10,11])
lines = df.plot.line()

axes = df.plot.line(subplots=True)
type(axes)


feature_names=["age","smoker","bloodpressure","sex","children","steps","bmi","region_nort
hwest",'region_northeast',
            'region_southeast','region_southwest']

range(len(feature_names))

imp=[]
for i in range(len(feature_names)):
    imp.append(fs.scores_[i])

feature_names=["age","smoker","bloodpressure","sex","children","steps","bmi","region_nort
hwest",'region_northeast',
            'region_southeast','region_southwest']
feature_importance = imp
```

```
feature_importance=gbr.feature_importances_
sorted_idx=np.argsort(feature_importance)
#largest_indices = sorted_idx[::-1][:11]
pos = np.arange(sorted_idx.shape[0]) + 0.5
fig = plt.figure(figsize=(8, 8))
plt.subplot(1, 1, 1)
plt.barh(pos, feature_importance, align="center")
plt.yticks(pos, np.array(feature_names)[sorted_idx])
plt.title("Feature Importance")


y_pred_gbr=gbr.predict(X_test)
train_pred_gbr=gbr.predict(X_train)
print_evaluate(y_train,train_pred_gbr)
print_evaluate(y_test,y_pred_gbr)
y_test=pd.DataFrame(y_test);y_test.head(10)
y_pred_gbr=gbr.predict(X_test)
y_pred=pd.DataFrame(y_pred_gbr);y_pred.head(10)
importance = gbr.feature_importances_

results = pd.DataFrame(data=[["Gradient Boosting Regressor", *importance ]],
                       columns=['Model','age', 'sex', 'children', 'steps', 'bmi', 'bloodpressure', 'smoker',
                            'region_northeast',                                    'region_northwest',
'region_southeast','region_southwest'])

results=results.transpose();results

#Οι παράμετροι που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου

gbr_params={'n_estimators':300,
        'max_depth':4,
        'learning_rate':0.01,
        'min_samples_split':4,
        'loss':'ls',
        "random_state":1234}

gbr=GradientBoostingRegressor(**gbr_params)
gbr.fit(X_train, y_train)

#Επιλογή μεταβλητών σύμφωνα με τη συνεισφορά τους στο μοντέλο

thres=gbr.feature_importances_.sort()
selection = SelectFromModel(gbr, threshold=thres, prefit=True)
select_X_train = selection.transform(X_train)

#Εκπαίδευση του μοντέλου μόνο με τις επιλεγμένες μεταβλητές

selection_model = GradientBoostingRegressor(**gbr_params)
```

```
selection_model.fit(select_X_train, y_train)

features = ['age', 'sex', 'children', 'steps', 'bmi', 'bloodpressure', 'smoker','charges',
'region_northeast',
        'region_northwest', 'region_southeast','region_southwest']

select_X_test = selection.transform(X_test)
y_pred = selection_model.predict(select_X_test)

#Πρόβλεψη του ύψους των ασφαλιστικών απαιτήσεων από το μοντέλο

len(y_test)
predictions = [round(value) for value in y_pred]
print_evaluate(y_test,predictions)
```

## Π2 Πηγαίος κώδικας σε Python για την 2η Εφαρμογή

```
def cross_val(model):
    cv = StratifiedKFold(n_splits=10, random_state=2, shuffle=True)
    pred1 = cross_val_score(model, X, y, scoring='accuracy',cv=cv)
    pred2 = cross_val_score(model, X, y, scoring='precision',cv=cv)
    pred3 = cross_val_score(model, X, y, scoring='recall',cv=cv)
    pred4 = cross_val_score(model, X, y, scoring='roc_auc',cv=cv)
    return pred1.mean(),pred2.mean(),pred3.mean(),pred4.mean()

def print_evaluate(true, predicted):
    Accuracy = metrics.accuracy_score(true, predicted)
    Precision = metrics.precision_score(true, predicted)
    Recall = metrics.recall_score(true, predicted)
    print('Accuracy:', Accuracy)
    print('Precision:', Precision)
    print('Recall:', Recall)

def evaluate(true, predicted):
    Accuracy = metrics.accuracy_score(true, predicted)
    Precision = metrics.precision_score(true, predicted)
    Recall = metrics.recall_score(true, predicted)
    return Accuracy, Precision, Recall

def grab_col_names(dataframe, cat_th=10, car_th=10):

    # cat_cols, cat_but_car
    cat_cols = [col for col in dataframe.columns if dataframe[col].dtypes == "O"]
    num_but_cat = [col for col in dataframe.columns if dataframe[col].nunique() <
cat_th and
                dataframe[col].dtypes != "O"]
    cat_but_car = [col for col in dataframe.columns if dataframe[col].nunique() >
car_th and
```

```
                dataframe[col].dtypes == "O"]
    cat_cols = cat_cols + num_but_cat
    cat_cols = [col for col in cat_cols if col not in cat_but_car]

    # num_cols
    num_cols = [col for col in dataframe.columns if dataframe[col].dtypes != "O"]
    num_cols = [col for col in num_cols if col not in num_but_cat]

    dict= {"Observations": dataframe.shape[0],
           "Variables": dataframe.shape[1],
           "Categorical": len(cat_cols),
           "Descrete": len(num_but_cat),
           "cardinal": len(cat_but_car),
           "Numeric": len(num_cols)}

variable_info=pd.DataFrame.from_dict(dict,orient="index",columns=["Variable_Info
"])
    display(variable_info)
    return cat_cols, num_cols, cat_but_car


#Read the data set
df = pd.read_csv('Train.csv')

df.columns=['feature_0', 'feature_1', 'feature_2', 'feature_3', 'feature_4',
       'feature_5', 'feature_6', 'feature_7', 'feature_8', 'feature_9',
       'feature_10', 'feature_11', 'feature_12', 'feature_13', 'feature_14',
       'feature_15', 'Churn']

df.isnull().sum()

cat_cols, num_cols, cat_but_car = grab_col_names(df,cat_th=15, car_th=12)

my_dict={"categorical":cat_cols,"descrete":cat_but_car,"numeric":num_cols}
table = pd.DataFrame.from_dict(my_dict, orient='index')
table=table.transpose()

table.drop(columns="descrete")

#Visualizations

sns.set(style='whitegrid')

f=plt.figure(figsize=(15, 3))
ax = f.add_subplot(141)
ax = sns.distplot(df['feature_0'], kde = True, color = 'c',ax=ax)
ax.set_title('Distribution of feature_0')
```

```
ax = f.add_subplot(142)
ax = sns.distplot(df['feature_1'], kde = True, color = 'r',ax=ax)
ax.set_title('Distribution of feature_1')

ax = f.add_subplot(143)
ax = sns.distplot(df['feature_2'], kde = True, color = 'g',ax=ax)
ax.set_title('Distribution of feature_2')

ax = f.add_subplot(144)
ax = sns.distplot(df['feature_3'], kde = True, color = 'k',ax=ax)
ax.set_title('Distribution of feature_3')

sns.set(style='whitegrid')

f=plt.figure(figsize=(13, 3))
ax = f.add_subplot(131)
ax = sns.distplot(df['feature_4'], kde = True, color = 'c',ax=ax)
ax.set_title('Distribution of feature_4')

ax = f.add_subplot(132)
ax = sns.distplot(df['feature_5'], kde = False, color = 'r',ax=ax)
ax.set_title('Distribution of feature_5')

ax = f.add_subplot(133)
ax = sns.distplot(df['feature_6'], kde = False, color = 'g',ax=ax)
ax.set_title('Distribution of feature_6')

sns.set(style='whitegrid')

f=plt.figure(figsize=(15, 3))
ax = f.add_subplot(141)
ax = sns.countplot(df['feature_7'],color = 'c',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_7')

ax = f.add_subplot(142)
ax = sns.countplot(df['feature_8'], color = 'r',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_8')

ax = f.add_subplot(143)
ax = sns.countplot(df['feature_9'], color = 'g',ax=ax)
```

```
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_8')

ax = f.add_subplot(144)
ax = sns.countplot(df['feature_9'],color = 'k',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_9')

sns.set(style='whitegrid')

f=plt.figure(figsize=(15, 3))
#f.subplots_adjust(hspace = 3)

ax = f.add_subplot(141)
ax = sns.countplot(df['feature_10'],color = 'c',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_10')

ax = f.add_subplot(142)
ax = sns.countplot(df['feature_11'],color = 'r',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_11')

ax = f.add_subplot(143)
ax = sns.countplot(df['feature_12'], color = 'k',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_12')

ax = f.add_subplot(144)
ax = sns.countplot(df['feature_13'], color = 'g',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
```

```
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_13')

sns.set(style='whitegrid')

f=plt.figure(figsize=(9, 3))
#f.subplots_adjust(hspace = 3)

ax = f.add_subplot(121)
ax = sns.countplot(df['feature_14'],color = 'c',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_14')

ax = f.add_subplot(122)
ax = sns.countplot(df['feature_15'],color = 'r',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=9)
ax.yaxis.set_tick_params(labelsize=9)
ax.set_title('Distribution of feature_15')

churn = df['Churn'].value_counts()
label_churn = churn.index
size_churn = churn.values
colors = ['silver', 'gold']

trace = go.Pie(
        labels = label_churn, values = size_churn, marker = dict(colors = colors), name
= 'Customer Churn', hole = 0.5)
table = [trace]
layout = go.Layout(
          title = 'Distribution of Churn')
fig = go.Figure(data = table, layout = layout)
py.iplot(fig)
table=pd.DataFrame(df,columns=["feature_0","feature_1","feature_2","feature_3","f
eature_4","feature_5","feature_6"])

#Normality test
for col in table:
    print(f"ks test for {col}")
    print(stats.kstest(table[col], 'norm'))
    print("==============================")

#Having a look at the correlation matrix
fig, ax = plt.subplots(figsize=(8,6))
```

```
sns.heatmap(table.corr(method="spearman"), annot=True, fmt='.1g', cmap="Blues",
cbar=False);

#perform the Mann-Whitney U test
stats.mannwhitneyu(df.feature_0, df.Churn, alternative='two-sided')

#Select the features and labels
no=df.loc[df["Churn"]==0]
yes=df.loc[df["Churn"]==1]

for col in table:
    print(f"mannw hitney test for {col}")
    print(stats.mannwhitneyu(no[col],yes[col],alternative='two-sided'))
    print("============================")

df=df.drop(columns=['feature_8', 'feature_9','feature_10'])

X=df.drop(columns=["Churn"],axis=1)
y=df["Churn"]

# set aside 20% of train and test data for evaluation
X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=0.2, shuffle = True, random_state = 42)

print("X_train shape: {}".format(X_train.shape))
print("X_test shape: {}".format(X_test.shape))
print("y_train shape: {}".format(y_train.shape))
print("y_test shape: {}".format(y_test.shape))

model = DecisionTreeClassifier()
# define pipeline
over = SMOTE(sampling_strategy=0.1)
under = RandomUnderSampler(sampling_strategy=0.5)
steps = [('o', over), ('u', under), ('m', model)]
pipeline = Pipeline(steps=steps)

#sm = SMOTE(random_state=42)
sm = SMOTENC(random_state=42, categorical_features=[7,8,9,10,11,12,13,14,15])
X_train, y_train = sm.fit_resample(X_train, y_train)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

mutual_info=mutual_info_classif(X_train,y_train)
mutual_info=pd.DataFrame(mutual_info,columns=["mutual_info"])
mutual_info["columns"]=['feature_0', 'feature_1', 'feature_2', 'feature_3', 'feature_4',
```

```
      'feature_5', 'feature_6', 'feature_7', 'feature_11', 'feature_12', 'feature_13',
'feature_14',
      'feature_15']
mutual_info.sort_values(ascending=False,by='mutual_info')

# import the class
lr.fit(X_train, y_train)
cross_val(lr)

y_pred_lr=lr.predict(X_test)

print("Training Accuracy: ", lr.score(X_train, y_train))
print('Testing Accuarcy: ', lr.score(X_test, y_test))

# making a classification report
cr = classification_report(y_test, y_pred_lr)
print(cr)

# making a confusion matrix
cm = confusion_matrix(y_test, y_pred_lr)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()

ld=LinearDiscriminantAnalysis()
ld.fit(X_train,y_train)
cross_val(ld)

y_pred_ld=ld.predict(X_test)

print("Training Accuracy: ", ld.score(X_train, y_train))
print('Testing Accuarcy: ', ld.score(X_test, y_test))

# making a classification report
cr = classification_report(y_test, y_pred_ld)
print(cr)

# making a confusion matrix
cm = confusion_matrix(y_test, y_pred_ld)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()

from sklearn import svm
svm = svm.SVC()
svm = svm.fit(X_train,y_train)
cross_val(svm)

y_pred_svm=svm.predict(X_test)
print("Training Accuracy: ", svm.score(X_train, y_train))
```

```
print('Testing Accuarcy: ', svm.score(X_test, y_test))

# making a classification report
cr = classification_report(y_test, y_pred_svm)
print(cr)

# making a confusion matrix
cm = confusion_matrix(y_test, y_pred_svm)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()

kn = KNeighborsClassifier(n_neighbors=5)
kn = kn.fit(X_train,y_train)
#cross_val(kn)

y_pred_kn=kn.predict(X_test)
print("Training Accuracy: ", kn.score(X_train, y_train))
print('Testing Accuarcy: ', kn.score(X_test, y_test))

# making a classification report
from sklearn.metrics import classification_report
cr = classification_report(y_test, y_pred_kn)
print(cr)

# making a confusion matrix
cm = confusion_matrix(y_test, y_pred_kn)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()

df1=df.copy()

df=df1

#df=df.drop(columns=['feature_7','feature_10',])
df=df.drop(columns=['feature_2','feature_8','feature_9','feature_10','feature_12',])
X=df.drop(columns=["Churn"],axis=1)
y=df["Churn"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30,
random_state=7)
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)



pca = PCA(n_components = 14)
```

```
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)

explained_variance = pca.explained_variance_ratio_
np.cumsum(explained_variance)

'''# feature selection
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import mutual_info_regression
def select_features(X_train, y_train, X_test):
    # configure to select all features
    fs = SelectKBest(score_func=mutual_info_regression, k=13)
    # learn relationship from training data
    fs.fit(X_train, y_train)
    # transform train input data
    X_train_fs = fs.transform(X_train)
    # transform test input data
    X_test_fs = fs.transform(X_test)
    return X_train_fs, X_test_fs, fs'''

xgb_parameters={'n_estimators':300,
                'max_depth':7,
                'subsample':0.7,
                'learning_rate':0.05,
                'random_state':5}

#X_train_fs, X_test_fs, fs = select_features(X_train, y_train, X_test)
xgb=XGBClassifier(**xgb_parameters)
xgb.fit(X_train, y_train)

# evaluate the model
yhat = xgb.predict(X_test)
# evaluate predictions
evaluate(y_test, yhat)

xgb = XGBClassifier(**xgb_parameters)
xgb.fit(X_train, y_train)
#cross_val(xgb)

y_pred_train=xgb.predict(X_train)
print("Training Accuracy: ", xgb.score(X_train, y_train))

# making a classification report
cr = classification_report(y_train, y_pred_train)
print(cr)

y_pred_xgb=xgb.predict(X_test)
```

```
print("Training Accuracy: ", xgb.score(X_train, y_train))
print('Testing Accuarcy: ', xgb.score(X_test, y_test))

# making a classification report
cr = classification_report(y_test, y_pred_xgb)
print(cr)

# making a confusion matrix
cm = confusion_matrix(y_test, y_pred_xgb)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()

xgb.feature_importances_

feature_importance=xgb.feature_importances_

sorted_idx=np.argsort(feature_importance)
pos = np.arange(sorted_idx.shape[0]) + 0.5
fig = plt.figure(figsize=(8, 8))
plt.subplot(1, 1, 1)
plt.barh(pos, feature_importance, align="center")
plt.yticks(pos, np.array(feature_names)[sorted_idx])
plt.title("Feature Importance")

# plot
pyplot.bar(range(len(xgb.feature_importances_)), xgb.feature_importances_)
pyplot.show()
explainer = shap.TreeExplainer(xgb)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test, plot_type="bar")
shap.summary_plot(shap_values, X_test)

# make predictions for test data and evaluate
xgb = XGBClassifier()
xgb.fit(X_train, y_train)
predictions = xgb.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
# Fit model using each importance as a threshold
thresholds = sort(xgb.feature_importances_)
for thresh in thresholds:
# select features using threshold
   selection = SelectFromModel(xgb, threshold=thresh, prefit=True)
   select_X_train = selection.transform(X_train)
# train model
   selection_model = XGBClassifier()
   selection_model.fit(select_X_train, y_train)
# eval model
```

```
    select_X_test = selection.transform(X_test)
    predictions = selection_model.predict(select_X_test)
    accuracy = accuracy_score(y_test, predictions)
    print("Thresh=%.3f, n=%d, Accuracy: %.2f%%" % (thresh,
select_X_train.shape[1], accuracy*100.0))

xgb.feature_importances_

plt.barh(['feature_0', 'feature_1', 'feature_2', 'feature_3', 'feature_4',
      'feature_5', 'feature_6', 'feature_7', 'feature_8', 'feature_9',
      'feature_10', 'feature_11', 'feature_12', 'feature_13', 'feature_14',
      'feature_15'] , xgb.feature_importances_)

eval_set = [(X_train, y_train), (X_test, y_test)]
xgb.fit(X_train, y_train, eval_metric=["error", "logloss"], eval_set=eval_set,
verbose=True)

results = xgb.evals_result()
print(results)

predictions = [round(value) for value in y_pred_xgb]

# evaluate predictions
accuracy = metrics.accuracy_score(y_test, y_pred_xgb)
print("Accuracy: %.2f%%" % (accuracy * 100.0))

# retrieve performance metrics
results = xgb.evals_result()
epochs = len(results['validation_0']['error'])
x_axis = range(0, epochs)

# plot log loss
fig, ax = pyplot.subplots()
ax.plot(x_axis, results['validation_0']['logloss'], label='Train')
ax.plot(x_axis, results['validation_1']['logloss'], label='Test')
ax.legend()
pyplot.ylabel('Log Loss')
pyplot.title('XGBoost Log Loss')
pyplot.show()

# plot classification error
fig, ax = pyplot.subplots()
ax.plot(x_axis, results['validation_0']['error'], label='Train')
ax.plot(x_axis, results['validation_1']['error'], label='Test')
ax.legend()
pyplot.ylabel('Classification Error')
pyplot.title('XGBoost Classification Error')
pyplot.show()
```

```
eval_set = [(X_train, y_train), (X_test, y_test)]
eval_metric = ["auc","error"]

xgb = XGBClassifier()
xgb.fit(X_train, y_train, eval_set=eval_set, eval_metric=eval_metric, verbose=True)

results = xgb.evals_result()
epochs = len(results['validation_0']['error'])
x_axis = range(0, epochs)

fig, ax = pyplot.subplots()
ax.plot(x_axis, results['validation_0']['auc'], label='Train')
ax.plot(x_axis, results['validation_1']['auc'], label='Test')
ax.legend()
pyplot.ylabel('AUC')
pyplot.title('XGBoost AUC')
pyplot.show()


xgb=XGBClassifier(n_estimators =100,model__max_depth=8,booster='gbtree', random_state
= 42, class_weight="balanced")
xgb.fit(X_train, y_train)
output = cross_validate(xgb, X, y, cv=2, scoring = 'f1', return_estimator =True)

for idx,estimator in enumerate(output['estimator']):
   print("Features sorted by their score for estimator {}:".format(idx))
   feature_importances = pd.DataFrame(xgb.feature_importances_,
                                index = ['feature_0', 'feature_2', 'feature_3', 'feature_4',
                                      'feature_6', 'feature_8','feature_11', 'feature_12',
                                      'feature_13', 'feature_14','feature_15'],
                                columns=['importance']).sort_values('importance', ascending=False)
   print(feature_importances)

xgb.feature_importances_

param_grid = {
   'model__max_depth': [3,5,6,7,8],
   'model__n_estimators': [50, 100, 500,1000],
   #'model__learning_rate': [0.1, 0.01, 0.001],
   #"model__subsample": [0.6, 0.8, 1.0],
   #"model__colsample_bytree": [0.6, 0.8, 1.0]
}

xgb_parameters={'n_estimators':500,'max_depth':8,'random_state':40,'learning_rate':0.05}
xgb=XGBClassifier(**xgb_parameters)
xgb.fit(X_train, y_train)
```

```
import eli5
from eli5.sklearn import PermutationImportance
permuter = PermutationImportance(xgb, scoring='f1', cv='prefit', n_iter=2, random_state=42)#instantiate permuter object
permuter.fit(X_test.values, y_test)

feature_names = X_test.columns.tolist()
eli5.show_weights(permuter, top=None, feature_names=feature_names)

mask = permuter.feature_importances_ > 0.01
features = X_train.columns[mask]

X_train = X_train[features]
X_test = X_test[features]
print('Shape after removing features:', X_train.shape)
print('Shape after removing features:', X_test.shape)

df.columns

X_test.head(30)

data_for_prediction = X_test[X_test.index==27520]
data_for_prediction

import shap
shap.initjs()
explainer = shap.TreeExplainer(xgb)
shap_values = explainer.shap_values(data_for_prediction)
shap.force_plot(explainer.expected_value, shap_values, data_for_prediction)

yhat=xgb.predict_proba(X_test[X_test.index==27520])

result = (yhat[0][1])*100;print("This customer will churn with probability",float("{0:.2f}".format(result)),"%")

# evaluate the model
yhat = xgb.predict(X_test)
yhat1 = xgb.predict(X_train)
# evaluate predictions
evaluate(y_train, yhat1)

evaluate(y_test, yhat)

features=['feature_2', 'feature_3', 'feature_5',
      'feature_11', 'feature_13', 'feature_14', 'feature_15']

from sklearn.inspection import permutation_importance
```

```
perm_importance = permutation_importance(xgb, X_test, y_test,random_state=40,scoring='f1',n_jobs=-1)
plt.barh(features, perm_importance.importances_mean)
plt.xlabel("Permutation Importance")
perm_importance.importances_mean

y_pred_xgb=xgb.predict(X_test)
print("Training Accuracy: ", xgb.score(X_train, y_train))
print('Testing Accuarcy: ', xgb.score(X_test, y_test))

# making a classification report
cr = classification_report(y_test, y_pred_xgb)
print(cr)

# making a confusion matrix
cm = confusion_matrix(y_test, y_pred_xgb)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()
cm = confusion_matrix(y_test, y_pred_xgb);cm

# roc curve and auc
ns_probs = [0 for _ in range(len(y_test))]
xgb_probs = xgb.predict_proba(X_test)
# keep probabilities for the positive outcome only
xgb_probs = xgb_probs[:, 1]
# calculate scores
ns_auc = roc_auc_score(y_test, ns_probs)
xgb_auc = roc_auc_score(y_test, xgb_probs)
# summarize scores
print('No Skill: ROC AUC=%.3f' % (ns_auc))
print('XGB: ROC AUC=%.3f' % (xgb_auc))
# calculate roc curves
ns_fpr, ns_tpr, _ = roc_curve(y_test, ns_probs)
xgb_fpr, xgb_tpr, _ = roc_curve(y_test, xgb_probs)
# plot the roc curve for the model
pyplot.plot(ns_fpr, ns_tpr, linestyle='--', label='No Skill')
pyplot.plot(xgb_fpr, xgb_tpr, marker='.', label='XGB')
# axis labels
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
# show the legend
pyplot.legend()
# show the plot
pyplot.show()
```

## Π3 Πηγαίος κώδικας σε Python για την 3η Εφαρμογή

```
def grab_col_names(dataframe, cat_th=10, car_th=10):
    # cat_cols, cat_but_car
    cat_cols = [col for col in dataframe.columns if dataframe[col].dtypes == "O"]
    num_but_cat = [col for col in dataframe.columns if dataframe[col].nunique() < cat_th and
                   dataframe[col].dtypes != "O"]
    cat_but_car = [col for col in dataframe.columns if dataframe[col].nunique() > car_th and
                   dataframe[col].dtypes == "O"]
    cat_cols = cat_cols + num_but_cat
    cat_cols = [col for col in cat_cols if col not in cat_but_car]

    # num_cols
    num_cols = [col for col in dataframe.columns if dataframe[col].dtypes != "O"]
    num_cols = [col for col in num_cols if col not in num_but_cat]

    dict= {"Observations": dataframe.shape[0],
           "Variables": dataframe.shape[1],
           "Categorical": len(cat_cols),
           "Descrete": len(num_but_cat),
           "cardinal": len(cat_but_car),
           "Numeric": len(num_cols)}
    variable_info=pd.DataFrame.from_dict(dict,orient="index",columns=["Variable_Info"])
    display(variable_info)
    return cat_cols, num_cols, cat_but_car

def change_width(ax, new_value) :
    for patch in ax.patches :
        current_width = patch.get_width()
        diff = current_width - new_value

        # we change the bar width
        patch.set_width(new_value)

        # we recenter the bar
        patch.set_x(patch.get_x() + diff * .5)

def chisquare_test(rows,columns,alpha=0.05):

    import scipy.stats as stats

    data_crosstab = pd.crosstab(rows,columns,margins=True, margins_name="Total")
    # Calcualtion of Chisquare
    chi_square = 0
    rows = rows.unique()
    columns = columns.unique()
    for i in columns:
```

```
        for j in rows:
            O = data_crosstab[i][j]
            E = data_crosstab[i]['Total'] * data_crosstab['Total'][j] / data_crosstab['Total']['Total']
            chi_square += (O-E)**2/E

    # The p-value approach
    print("The p-value approach to hypothesis testing in the decision rule")
    p_value = 1 - stats.chi2.cdf(chi_square, (len(rows)-1)*(len(columns)-1))
    result = "Failed to reject the null hypothesis."
    if p_value <= alpha:
        result = "Null Hypothesis is rejected."

    print("chisquare-score is:", chi_square, " and p value is:", p_value)
    print(result)
    return chi_square, p_value

df = pd.read_excel('sample.xlsx')
print('From historical data intrested in Vehicle insurance is',round(df["Response"].value_counts()[1]/sum(df["Response"].value_counts())*100,2),'% from all customers')

df=df.drop(columns=["Unnamed: 0","id","Region_Code"])
df.isnull().sum()

cat_cols, num_cols, cat_but_car = grab_col_names(df,cat_th=55, car_th=55)

my_dict={"categorical":cat_cols,"discrete":cat_but_car,"numeric":num_cols}
table = pd.DataFrame.from_dict(my_dict, orient='index')
table=table.transpose()

numeric=pd.DataFrame(df,columns=["Age","Annual_Premium","Policy_Sales_Channel","Vintage"])
numeric.describe()

categorical=pd.DataFrame(df,columns=list(cat_cols))
for i in cat_cols:
    df[i]=df[i].astype(object)

df[cat_cols].describe()

#Normality test
from scipy import stats
for col in numeric:
    print(f"ks test for {col}")
    print(stats.kstest(numeric[col], 'norm'))
    print("============================")
```

```
#Label encoding
df["Driving_License"]=df["Driving_License"].replace((0,1),("No","Yes"))
df["Previously_Insured"]=df["Previously_Insured"].replace((0,1),("No","Yes"))
df["Response"]=df["Response"].replace((0,1),("No","Yes"))

#Graphs
sns.set(style='whitegrid')

f=plt.figure(figsize=(12, 5))
ax = f.add_subplot(121)
ax = sns.distplot(df['Age'], kde = True, color = 'c',ax=ax)
ax.set_title('Distribution of Age')
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )

ax = f.add_subplot(122)
ax = sns.distplot(df['Annual_Premium'], kde = True, color = 'r',ax=ax)
ax.set_title('Distribution of Annual_Premium')

ax = f.add_subplot(121)
ax = sns.distplot(df['Policy_Sales_Channel'], kde = True, color = 'g',ax=ax)
ax.set_title('Distribution of Policy_Sales_Channel')

ax = f.add_subplot(122)
ax = sns.distplot(df['Vintage'], kde = True, color = 'k',ax=ax)
ax.set_title('Distribution of Vintage')

plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )

ax = f.add_subplot(121)
ax = sns.countplot(df['Gender'],color = 'c',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=12)
ax.yaxis.set_tick_params(labelsize=12)
ax.set_title('Gender')
change_width(ax, .35)

ax = f.add_subplot(122)
ax = sns.countplot(df['Vehicle_Age'], color = 'r',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=12)
ax.yaxis.set_tick_params(labelsize=12)
ax.set_title('Vehicle Age')
change_width(ax, .35)
```

```
ax = f.add_subplot(121)
ax = sns.countplot(df['Vehicle_Damage'], color = 'g',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=12)
ax.yaxis.set_tick_params(labelsize=12)
ax.set_title('Vehicle Damage')
change_width(ax, .35)

ax = f.add_subplot(122)
ax = sns.countplot(df['Driving_License'],color = 'r',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=12)
ax.yaxis.set_tick_params(labelsize=12)
ax.set_title('Driving Licence')
change_width(ax, .35)

f=plt.figure(figsize=(12, 5))
ax = f.add_subplot(121)
ax = sns.countplot(df['Previously_Insured'], color = 'g',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=12)
ax.yaxis.set_tick_params(labelsize=12)
ax.set_title('Previously Insured')
change_width(ax, .35)

ax = f.add_subplot(122)
ax = sns.countplot(df['Response'],color = 'c',ax=ax)
ax.xaxis.label.set_visible(False)
ax.yaxis.label.set_visible(False)
ax.xaxis.set_tick_params(labelsize=12)
ax.yaxis.set_tick_params(labelsize=12)
ax.set_title('Customer is interested in Vehicle Insurance')
change_width(ax, .35)

data = pd.crosstab(df['Gender'], df['Vehicle_Damage'])
#colors = plt.cm.Blues(np.linspace(0, 1, 5))
data.div(data.sum(1).astype(float), axis = 0).plot(kind = 'bar',
                                                   stacked = False,
                                                   figsize = (15, 7))

plt.title('Vehicle Damage per Gender', fontsize = 17)
plt.xlabel('Gender', fontsize=15)
plt.ylabel('Vehicle Damagr', fontsize=15)
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 15)
```

```
plt.legend()
plt.show()

data = pd.crosstab(df['Gender'], df['Vehicle_Damage']); data

data = pd.crosstab(df['Gender'], df['Response'])
#colors = plt.cm.Blues(np.linspace(0, 1, 5))
data.div(data.sum(1).astype(float), axis = 0).plot(kind = 'bar',
                                                   stacked = False,
                                                   figsize = (15, 7))

plt.title('Interest for Vehicle Insurance per Gender', fontsize = 20)
plt.xlabel('Gender', fontsize=15)
plt.ylabel('Interest for Vehicle Insurance', fontsize=15)
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 15)
plt.legend()
plt.show()

data = pd.crosstab(df['Vehicle_Damage'], df['Response'])
#colors = plt.cm.Blues(np.linspace(0, 1, 5))
data.div(data.sum(1).astype(float), axis = 0).plot(kind = 'bar',
                                                   stacked = False,
                                                   figsize = (15, 7))

plt.title('Interest for Vehicle Insurance * Vehicle Damage', fontsize = 20)
plt.xlabel('Vehicle Damage', fontsize=15)
plt.ylabel('Interest for Vehicle Insurance', fontsize=15)
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 15)
plt.legend()
plt.show()

data = pd.crosstab(df['Vehicle_Age'], df['Response'])
#colors = plt.cm.Blues(np.linspace(0, 1, 5))
data.div(data.sum(1).astype(float), axis = 0).plot(kind = 'bar',
                                                   stacked = False,
                                                   figsize = (15, 7))

plt.title('Interest for Vehicle Insurance * Vehicle Age', fontsize = 20)
plt.xlabel('Vehicle_Age', fontsize=15)
plt.ylabel('Interest for Vehicle Insurance', fontsize=15)
plt.xticks(fontsize = 15)
plt.yticks(fontsize = 15)
plt.legend()
plt.show()

chisquare_test(df['Vehicle_Age'],df['Response'])
```

```
pd.crosstab(df['Vehicle_Age'],df['Response'])

f, ax = plt.subplots(1, 1, figsize=(8, 5))
ax = sns.barplot(x='Response', y='Age', data=df, palette='Set1')
ax.set_ylabel('Age', fontsize = 12.0) # Y label
ax.set_xlabel('Response', fontsize = 12) # X label
plt.xticks(fontsize=12 )
plt.yticks(fontsize=12 )
change_width(ax, .35)
plt.show()

#Label encoding
df["Driving_License"]=df["Driving_License"].replace(("No","Yes"),(0,1))
df["Previously_Insured"]=df["Previously_Insured"].replace(("No","Yes"),(0,1))
df["Response"]=df["Response"].replace(("No","Yes"),(0,1))
df["Gender"]=df["Gender"].replace(("Male","Female"),(1,2))
df["Vehicle_Age"]=df["Vehicle_Age"].replace(("< 1 Year","1-2 Year","> 2 Years"),(1,2,3))
df["Vehicle_Damage"]=df["Vehicle_Damage"].replace(("No","Yes"),(0,1))

def cross_val(model):
    cv = KFold(n_splits=10, random_state=2, shuffle=True)
    pred1 = cross_val_score(model, X, y, scoring='accuracy',cv=cv)
    pred2 = cross_val_score(model, X, y, scoring='precision',cv=cv)
    pred3 = cross_val_score(model, X, y, scoring='recall',cv=cv)
    pred4 = cross_val_score(model, X, y, scoring='f1',cv=cv)
    pred5 = cross_val_score(model, X, y, scoring='roc_auc',cv=cv)
    return pred1.mean(),pred2.mean(),pred3.mean(),pred4.mean(),pred5.mean()

def evaluate(true, predicted):
    Accuracy = metrics.accuracy_score(true, predicted)
    Precision = metrics.precision_score(true, predicted)
    Recall = metrics.recall_score(true, predicted)
    f1=metrics.f1_score(true, predicted)
    auc=metrics.roc_auc_score(true, predicted,average='weighted')
    return Accuracy, Precision, Recall,f1,auc

#Original dataset
X=df.drop(columns=["Response"],axis=1)
y=df["Response"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30,
random_state=7,shuffle=True)

#XGboost
xgb=XGBClassifier()
xgb.fit(X_train, y_train)
cross_val(xgb)
```

```
#Dataset with oversampling method (SMOTE)
X=df.drop(columns=["Response"],axis=1)
y=df["Response"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30,
random_state=7,shuffle=True)
sm = SMOTE(random_state=42)
X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)

#XGboost
xgb=XGBClassifier()
xgb.fit(X_train_sm, y_train_sm)
cross_val(xgb)

import shap
explainer = shap.TreeExplainer(xgb)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test, plot_type="bar")

#Final_model
X=df.drop(columns=["Response",axis=1)
y=df["Response"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30,
random_state=7,shuffle=True)
sm = SMOTE(random_state=42)
X_train_sm, y_train_sm = sm.fit_resample(X_train, y_train)

xgb_parameters={'n_estimators':600,
            'max_depth':5,
            'subsample':0.7,
            'learning_rate':0.05,
            'random_state':4}
#XGboost
xgb=XGBClassifier(**xgb_parameters)
xgb.fit(X_train_sm, y_train_sm)

# evaluate the model
yhat = xgb.predict(X_test)
yhat_train = xgb.predict(X_train)
# evaluate predictions
evaluate(y_test, yhat)

print("Training Accuracy: ", xgb.score(X_train_sm, y_train_sm))
print('Testing Accuarcy: ', xgb.score(X_test, y_test))

# making a classification report
cr = classification_report(y_test, yhat)
print(cr)
```

```
# making a confusion matrix
cm = confusion_matrix(y_test, yhat)
sns.heatmap(cm, annot = True, cmap = 'Purples')
plt.show()

explainer = shap.TreeExplainer(xgb)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test, plot_type="bar")
```

# Βιβλιογραφία

## Ελληνική

Maria Spiteri, George Azzopardi (2021). Customer Churn Prediction for a Motor Insurance Company

https://repository.kallipos.gr/bitstream/11419/2128/2/04_chapter03.pdf

https://repository.kallipos.gr/bitstream/11419/1237/2/Kef._10.pdf

http://apothesis.teicm.gr/xmlui/handle/123456789/859

## Ξένη

Trevor Hastie, Robert Tibshirani and Jerome Friedman (2008). The Elements of Statistical Learning

Mauricio Henao Madrigal, Diego Restrepo Tob_on, Henry Laniado (2020). CUSTOMER CHURN PREDICTION IN INSURANCE INDUSTRIES: A MULTIPRODUCT APROACH

Katharina Morik and Hanna K¨opcke (2004). Analysing Customer Churn in Insurance Data – A Case Study

Akashdeep Bhardwaj (2020). Health Insurance Claim Prediction Using Artificial Neural Networks

José M. Maisog, Wenhong Li, Yanchun Xu1, Brian Hurley, Hetal Shah, Ryan Lemberg, Tina Borden, Stephen Bandeian, Melissa Schline1, Roxanna Cross, Alan Spiro, Russ Michael, Alexander Gutfraind. Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach

Cieslak, D., Chawla, N.: Learning Decision Trees for Unbalanced Data. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 241–256. Springer, Heidelberg (2008)

Liu, W., Chawla, S., Cieslak, D., Chawla, N.: A Robust Decision Tree Algorithms for Imbalanced Data Sets. In: Proceedings of the Tenth SIAM International Conference on Data Mining, pp. 766–777 (2010)

Hossein Joudaki, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, and Mohammad Arab, (2015).Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature

https://www.researchgate.net/publication/338292264_Using_massive_health_insurance_claims_data_to_predict_very_high-cost_claimants_a_machine_learning_approach

Jose M. Maisog, Yanchun Xu, Wenhong Li,(2019).Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach

Wei Liu & Sanjay Chawla (2011). Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets

Melih Kirlidog, Cuneyt Asuk (2012).A fraud detection approach with data mining in health insurance

Fotios Mourdoukoutas, Tim J. Boonen, Bonsoo Koo, Athanasios A. Pantelous, (2021). Pricing in a competitive stochastic insurance market

Giorgio Alfredo Spedicato, Christophe Dutang, Leonardo Petrini. Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs. Variance, Casualty Actuarial Society, (2018). Machine Learning Methods to Perform Pricing Optimization. A Comparison with Standard GLMs

Cunningham, P., Cord, M., Delany, S.J. (2008). Supervised Learning. In: Cord, M., Cunningham, P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

Elhassan, Tusneem & M, Aljourf & F, Al-Mohanna & Shoukri, Mohamed. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. Global Journal of Technology and Optimization. 01. 10.4172/2229-8711.S1111

https://www.jmlr.org/papers/volume18/16-365/16-365.pdf

https://bmjopen.bmj.com/content/bmjopen/9/6/e028409.full.pdf

https://policyadvice.net/insurance/insights/health-insurance-statistics/

https://www.igi-global.com/article/health-insurance-claim-prediction-using-artificial-neural-networks/257242

https://ieeexplore.ieee.org/abstract/document/7039220

https://www.valueinhealthjournal.com/article/S1098-3015(18)33095-X/fulltext