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

ΤΜΗΜΑ Σ Τ Α Τ Ι Σ Τ Ι Κ Η Σ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ Ε Π Ι Σ Τ Η Μ Η Σ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ
ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ

ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΑΝΑΛΥΣΗΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΑΝΑΛΟΓΙΣΜΟ

Ανέστης Τσέκος

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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Πειραιάς
Σεπτέμβριος 2021

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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου.

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Αναπληρωτής Καθηγητής Π. Τήνιος
- Αναπληρωτής Καθηγητής Ν. Πελέκης

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

U N I V E R S I T Y O F P I R A E U S

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN ACTUARIAL SCIENCE AND RISK MANAGEMENT

USE OF BIG DATA ANALYTICS IN ACTUARY

By

Anestis Tsekos

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 Actuarial Science and Risk Management

Piraeus, Greece
September 2021

Περίληψη

Στην σημερινή εποχή υπάρχει ανάγκη επεξεργασίας τεράστιων ποσοτήτων πολυδιάστατων δεδομένων με μεγάλη διάσταση (high dimensional big data) και υψηλή πολυπλοκότητα. Η επιστημονική περιοχή της ανάλυσης μεγάλων δεδομένων (big data analytics) έχει πλέον επηρεάσει θετικά τελευταία και τον Ασφαλιστικό χώρο αφού ήδη στις μεγάλες ασφαλιστικές εταιρείες γίνεται εκτεταμένη χρήση τέτοιων τεχνικών στην τιμολόγηση ασφαλίστρων και τη διαχείριση απαιτήσεων. Για παράδειγμα στην τιμολόγηση είναι ιδιαίτερα επωφελής η εφαρμογή τεχνικών κατάτμησης (segmentation) και προβλεπτικής μοντελοποίησης (predictive modelling) για να αποτιμηθεί ακριβέστερα ο κίνδυνος και να γίνει ασφαλής/επωφελής τιμολόγηση ασφαλίστρων. Έτσι, οι ασφαλιστικές εταιρείες χρησιμοποιούν πλέον προχωρημένα εργαλεία π.χ. μοντέλα συμπεριφοράς βασισμένα σε δεδομένα προφίλ πελατών - με συνεχή ροή δεδομένων πραγματικού χρόνου - π.χ. δορυφορικά δεδομένα, αναφορές καιρού, αισθητήρες οχημάτων - για να δημιουργηθεί λεπτομερής και εξατομικευμένη αξιολόγηση του κινδύνου.

Στα πλαίσια της εργασίας αυτής πραγματοποιείται συστηματική παρουσίαση στατιστικών τεχνικών ανάλυσης μεγάλων δεδομένων που έχουν χρησιμοποιηθεί ή έχουν αναπτυχθεί αποκλειστικά για το χώρο του αναλογισμού και της διοίκησης κινδύνου. Ειδικότερα περιγράφονται τα χαρακτηριστικά των Μεγάλων Δεδομένων και οι τεχνικές που τα διέπουν, σύμφωνα με τη βιβλιογραφία. Στη συνέχεια, παρουσιάζονται οι τεχνικές ταξινόμησης και τα δέντρα αποφάσεων. Τέλος, παρουσιάζονται συγκεκριμένα παραδείγματα εφαρμογής σχετικά με μερικές από τις τεχνικές αυτές για να καταδειχθεί ο τρόπος χρήσης τους και η αποτελεσματικότητά τους.

Abstract

Nowadays there is a need to process huge amounts of high dimensional big data with high complexity. The scientific field of big data analytics has now had a positive impact on the insurance industry, since already in large insurance companies there is an extensive use of such techniques in premium pricing and claims management. For instance, in pricing, the application of segmentation techniques and predictive modeling is particularly beneficial to assess the risk more accurately and to make safe / beneficial insurance premiums. Thus, insurance companies, now, use advanced tools e.g. Behavioral models based on customer profile data - with a continuous flow of real-time data - e.g. satellite data, weather reports, vehicle sensors - to create a detailed and personalized risk assessment.

In this thesis, a systematic presentation of statistical techniques of big data analysis that have been used or developed exclusively for the field of actuarial and risk management is carried out. In particular, the characteristics of Big Data and the techniques that govern them are described, according to the literature. Next, the classification techniques and decision trees are presented. Finally, specific examples of application some of these techniques are presented to demonstrate how they are used and their effectiveness.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1: Τα Μεγάλα Δεδομένα

1.1 Εισαγωγή

Στη παρούσα ενότητα πραγματοποιείται μια πρώτη επαφή με την έννοια των μεγάλων δεδομένων, τα οποία πλέον είναι διαθέσιμα σε διάφορους τομείς, ως απόρροια της εξέλιξης της τεχνολογίας. Πιο συγκεκριμένα, η ανοδική τάση του όγκου των διαθέσιμων δεδομένων συνδέεται άμεσα με την αυξημένη χρήση και δημοτικότητα των διαφόρων πηγών προέλευσης τους, όπως τα μέσα κοινωνικής δικτύωσης αλλά και οποιαδήποτε συσκευή που συνδέεται στο διαδίκτυο. Ως εκ τούτου, εταιρείες και οργανισμοί οφείλουν να αναλύσουν τον μεγάλο όγκο δεδομένων, ώστε να αντλήσουν χρήσιμες πληροφορίες και να επιφέρουν νέες τάσεις στην αγορά FinTech. Ο όρος FinTech περιγράφει τον κλάδο παροχής χρηματοοικονομικών υπηρεσιών αποκλειστικά μέσω καινοτόμων τεχνολογιών πληροφορικής και επικοινωνιών. Τα μεγάλα δεδομένα (big data) έρχονται και προσδένονται σε αυτή λογική των επιχειρήσεων.

Το 2013, το λεξικό της αγγλικής γλώσσας *Oxford English Dictionary* εισήγαγε τον όρο «Μεγάλα Δεδομένα» στα περιεχόμενα του. Αυτό το γεγονός, σηματοδοτεί την αξία αυτού του πεδίου σε συνδυασμό με την απότομη αύξηση ενδιαφέροντος γύρω από το πεδίο των Μεγάλων Δεδομένων, παρ' όλο που η παρουσία τους είχε προηγηθεί αρκετά χρόνια πριν πραγματοποιηθεί η σημερινή αλματώδης εφαρμογή τους.

Το 1944, ο Αμερικάνος βιβλιοθηκονόμος *Fremont Rider*, υπολόγισε ότι το μέγεθος των βιβλιοθηκών στα πανεπιστήμια της Αμερικής θα διπλασιάζεται κάθε δεκαέξι χρόνια. Δεδομένου αυτού του ρυθμού ανάπτυξης, κατέληξε στην θεωρία ότι το 2040, η βιβλιοθήκη του πανεπιστημίου Yale θα έχει να διαχειριστεί 200 εκατομμύρια τόμους (Rider, 1944). Προφανώς, δεν προέβλεψε την ψηφιοποίηση των βιβλίων αλλά κατάφερε να εκτιμήσει με ακρίβεια την έκρηξη του όγκου των πληροφοριών. Η μεγάλη ποσότητα είναι ένα από τα κύρια χαρακτηριστικά των Μεγάλων Δεδομένων. Βέβαια, δεν είναι το μόνο χαρακτηριστικό ώστε να θεωρηθεί ένα σύνολο δεδομένων «μεγάλο». Σαφώς, η πρόβλεψη του Rider δεν καθορίζει τα Μεγάλα Δεδομένα, αλλά κατέχει ιστορική θέση ως η πρώτη εισήγηση που σχετιζόταν με αυτά.

Ο όρος «Μεγάλα Δεδομένα» χρησιμοποιήθηκε για πρώτη φορά το 1998 από τον μηχανικό ηλεκτρονικών υπολογιστών John R. Mashey. Λίγα χρόνια αργότερα, το 2001, η αμερικάνικη εταιρεία συμβουλευτικής *Gartner*, διατύπωσε έναν αρκετά διαδεδομένο ορισμό που ισχύει μέχρι και σήμερα. Τα Μεγάλα Δεδομένα είναι ένα σύνολο δεδομένων μεγάλης ποσότητας, υψηλής ταχύτητας, με εκτενή ποικιλομορφία που απαιτούν οικονομικά αποδοτικούς και καινοτόμους τρόπους επεξεργασίας πληροφοριών. Έτσι, επιτυγχάνονται, η διαδικασία εξόρυξης γνώσης, η λήψη αποφάσεων και η αυτοματοποίηση διαδικασιών.

Ένας πιο ευρύς ορισμός δόθηκε από την γνωστή εταιρεία τεχνολογίας IBM. Ερμηνεύει τα Μεγάλα Δεδομένα ως ένα φαινόμενο που βασίζεται από την γέννηση υπέρογκων δεδομένων στην εποχή της πληροφόρησης, που διανύεται. Τα δεδομένα συλλέγονται και παράγονται τόσο γρήγορα που έχουν κατακλύσει τις κυβερνήσεις, την οικονομία και την κοινωνία. Επομένως, δημιουργείται μια πρόκληση αλλά και μια ευκαιρία. Πιο συγκεκριμένα, η πρόκληση σχετίζεται με τον τρόπο αξιοποίησης του μεγάλου όγκου δεδομένων. Παράλληλα, η ευκαιρία αντικατοπτρίζεται στη θετική κατάληξη, που ίσως επιφέρει στην υπό εξέταση κοινωνία, η κατάλληλη ανάλυση αυτών των δεδομένων.

Τα Μεγάλα Δεδομένα αντλούνται από εσωτερικές και εξωτερικές πηγές όπως για παράδειγμα από τα social media, τις συναλλαγές, τους αισθητήρες και τα κινητά. Οι εταιρίες μπορούν να αξιοποιήσουν τα δεδομένα με σκοπό να προσαρμόσουν τα προϊόντα τους καλύτερα στις ανάγκες των πελατών τους ή να ανακαλύψουν νέες πηγές εισοδήματος.

Οι προαναφερόμενοι ορισμοί προσδίδουν νόημα στον όρο «Μεγάλα Δεδομένα», όχι απλά περιγραφή του υπέρογκου μεγέθους ενός συνόλου δεδομένων. Στη συνέχεια του κεφαλαίου αναλύονται τα βασικά χαρακτηριστικά που πρέπει να έχει ένα σύνολο δεδομένων, ώστε να θεωρηθεί ως «Μεγάλα Δεδομένα».

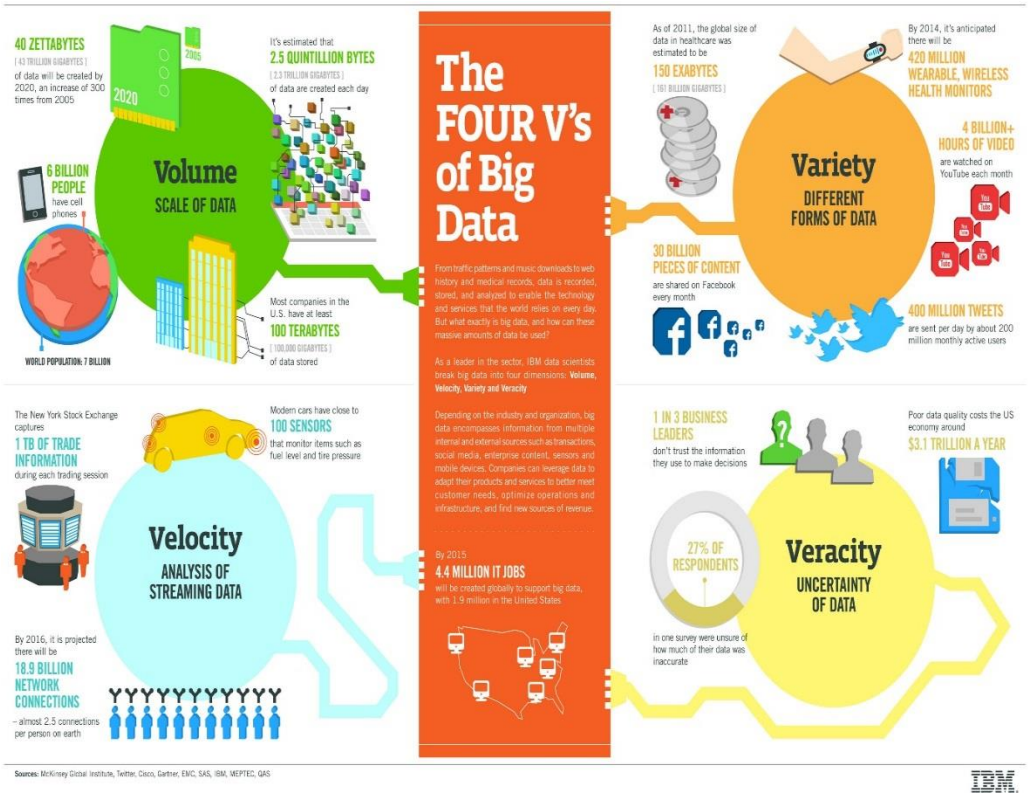
1.2 Χαρακτηριστικά των Μεγάλων Δεδομένων

Στο σημείο αυτό αναφέρονται και αναλύονται τα χαρακτηριστικά που περιγράφουν τα «Μεγάλα Δεδομένα». Ακολουθώντας το Σχήμα 1.1, η εταιρεία IBM ορίζει τέσσερα βασικά χαρακτηριστικά για τα Μεγάλα Δεδομένα, γνωστά ως τα «4 V's», τα οποία είναι:

α. Ποσότητα (Volume)
β. Ποικιλομορφία (Variety)
γ. Ταχύτητα (Velocity)
δ. Ακρίβεια (Veracity)

Δίνουμε στη συνέχεια τον ορισμό αυτών των τεσσάρων χαρακτηριστικών αλλά και κάποια παραδείγματα.

Σχήμα 1.1. *Τα 4 V's σύμφωνα με την IBM (IBM, 2016)*



α. Ποσότητα

Το κύριο χαρακτηριστικό που κάνει τα δεδομένα μεγάλα, είναι το απόλυτο μέγεθός τους. Από ότι γνωρίζουμε το συνολικό μέγεθος των δεδομένων που παράγονται αυξάνεται εκθετικά κάθε χρόνο. Σύμφωνα με την πολυεθνική εταιρία IT και δικτύωσης Cisco, η συνολική αποθηκευμένη ποσότητα δεδομένων εν έτει 2020 υπολογίζεται μεταξύ 10 με 50 zettabytes[1].

Η συνολική αποθηκευμένη ποσότητα δεδομένων ορίζεται ως το γενικό σύνολο δεδομένων της Διαδικτυακής Κυκλοφορίας αλλά και των συνδεδεμένων συσκευών στο διαδίκτυο. Το 2025, εκτιμάται ότι το εύρος θα ανέρχεται μεταξύ 150 με 250 zettabytes. Επίσης, υπολογίζεται ότι καθημερινά παράγονται 2.5 τρισεκατομμύρια gigabytes δεδομένων.

Εξετάζοντας την εξέλιξη του όγκου των δεδομένων, το 1999 τα δεδομένα ανέρχονταν στα 1.5 exabytes. Το 2006, τα συνολικά δεδομένα είχαν αυξηθεί κατά 1000% σε σχέση με το 1999. Συνεπώς, είχαμε μια μεγάλη αύξηση και στα Μεγάλα Δεδομένα. Η ραγδαία αύξηση του συνολικού όγκου των δεδομένων χρόνο με το χρόνο είχε ως αποτέλεσμα να φτάσουμε σε ένα τεράστιο όγκο δεδομένων να διαχειριστούμε σε μορφές εικόνας, κειμένου, βίντεο και άλλων μορφών.

Ένα μεγάλο ποσοστό όγκου δεδομένων αντλείται από το Διαδίκτυο των πραγμάτων (Internet of things – IoT). Η εταιρεία αναλυτών Gartner εκτιμάει ότι το 2020 υπάρχουν πάνω από 26 δισεκατομμύρια συνδεδεμένες συσκευές. Η βασική ιδέα του IoT είναι η σύνδεση οποιασδήποτε ηλεκτρονικής συσκευής στο ίντερνετ ή μεταξύ τους ώστε να υπάρχει η δυνατότητα της ανταλλαγής δεδομένων.

Ως ηλεκτρονικές συσκευές θεωρούνται διάφορες συσκευές όπως smartphones, φορητές συσκευές, κλιματιστικά, αυτοκίνητα καθώς και κάθε άλλη ηλεκτρονική συσκευή που μπορεί να σκεφτεί κανείς, η οποία διαθέτει αισθητήρες, σύνδεση στο διαδίκτυο και λογισμικό. Το Διαδίκτυο των πραγμάτων είναι ένα γιγαντιαίο δίκτυο συνδεδεμένων πραγμάτων και ανθρώπων, μέσα στο οποίο συλλέγονται και διαμοιράζονται δεδομένα. Ένα παράδειγμα είναι τα self-driving αυτοκίνητα, των οποίων οι περίπλοκοι αισθητήρες εντοπίζουν αντικείμενα στην πορεία τους με σκοπό

[1] Το byte είναι η βασική μονάδα μέτρησης χώρου και πληροφορίας στα υπολογιστικά συστήματα. Ένα byte ισοδυναμεί με 8 bit. Πολλαπλάσια του byte είναι τα Kilobyte, Megabyte, Gigabyte, Terabyte, Petabyte, Exabyte, Zettabyte και Yottabyte. Γνωρίζουμε ότι $1\ zettabyte = 1024^7\ bytes$

την αποφυγή τους. Συσκευές και πράγματα με ενσωματωμένους αισθητήρες συνδέονται σε μια πλατφόρμα, η οποία ενσωματώνει τα δεδομένα από τις διάφορες συσκευές. Αυτές οι ισχυρές πλατφόρμες του Διαδικτύου των πραγμάτων μπορούν να ξεχωρίσουν ποιες πληροφορίες είναι σημαντικές και να αναλύσουν τα δεδομένα τους με σκοπό τη βελτιστοποίηση. Συνήθως, μπορεί να χρησιμοποιηθούν για την ανίχνευση μοτίβων ή τον εντοπισμό προβλημάτων πριν καν εμφανιστούν.

Συμπερασματικά, παρατηρούμε ότι η ποσότητα των δεδομένων αυξάνεται καθημερινά με γοργούς ρυθμούς. Το ζήτημα δεν θα είναι πια η αναζήτηση και η ποσότητα των δεδομένων που υπήρχε στο παρελθόν αλλά η εύρεση σημαντικών και χρήσιμων πληροφοριών μέσω διαφόρων αναλύσεων της μεγάλης ποσότητας δεδομένων.

β. Ποικιλομορφία

Ο όρος ποικιλομορφία χαρακτηρίζεται ως το εύρος των διαφορετικών τύπων δεδομένων που υπάρχουν. Ο βασικός διαχωρισμός γίνεται σε τρείς τύπους δεδομένων: τα δομημένα (structured), τα μη δομημένα δεδομένα (unstructured) και τα ημι-δομημένα (semi-structured).

Ο όρος **δομημένα δεδομένα** περιγράφει τους παραδοσιακούς τύπους δεδομένων που μπορούν να αναλυθούν, να αποθηκευτούν, να αναζητηθούν και να επεξεργαστούν με ευκολία από τα υπολογιστικά συστήματα. Παραδείγματα δομημένων δεδομένων αποτελούν ο χρόνος, στοιχεία απογραφής (ημερομηνία γέννησης, εισόδημα, status εργαζόμενου), τηλεφωνικοί αριθμοί κτλ.

Τα **μη δομημένα δεδομένα** δεν καθορίζονται από κάποιο σύνολο κανόνων με αποτέλεσμα να είναι δύσκολο να αναλυθούν, να οργανωθούν και να συνδεθούν μεταξύ τους. Για παράδειγμα, μια φωτογραφία, οι HTML σελίδες, τα αρχεία ήχου, οι ιστοσελίδες, ένα tweet στο twitter από τα 400 εκατομμύρια που ανεβαίνουν την ημέρα ή το κυρίως κείμενο ενός e-mail, μπορεί να έχουν κάποια συγκεκριμένη αλλά όχι εγγενή δομή. Μέσω διάφορων τεχνικών αντιμετώπισης των μη δομημένων δεδομένων όπως είναι η ανάλυση κειμένων, μπορούμε αρκετές φορές να εξάγουμε διάφορα χρήσιμα μοτίβα και συμπεράσματα.

Τέλος τα **ημι-δομημένα δεδομένα** χαρακτηρίζονται ως δεδομένα που διαθέτουν κάποιου τύπου δομή αλλά αυτή η δομή δεν είναι όμοια για όλα τα δεδομένα. Τα δεδομένα αυτά βρίσκονται μεταξύ δομημένων και αδόμητων. Τέτοια παραδείγματα

είναι συγκεκριμένες ιστοσελίδες (DTD – XML Schema) ή οι πίνακες σε υπολογιστικά φύλλα.

γ. Ταχύτητα

Στο άκουσμα του όρου Μεγάλα Δεδομένα, η πλειονότητα του κόσμου αντιλαμβάνεται ένα τεράστιο όγκο δεδομένων πολλών γραμμών και στηλών που χρήζουν ανάλυσης. Όμως εκτός του μεγάλου όγκου και της ποικιλομορφίας των δεδομένων, ένα πολύ σημαντικό χαρακτηριστικό είναι η ταχύτητα.

Η ταχύτητα δύναται να περιγραφεί ως η συχνότητα των εισερχομένων δεδομένων στο σύστημα και παράλληλα ο χρόνος που απαιτείται για την επεξεργασία και ανάλυση των δεδομένων αυτών, ώστε να εξαχθούν κάποια συμπεράσματα. Τα υπολογιστικά συστήματα και το ανθρώπινο δυναμικό κάθε εταιρίας θα πρέπει να είναι σε θέση να διαχειριστεί, να αποθηκεύσει, να επεξεργαστεί αλλά και να αναλύσει ταχύτατα τον υψηλό όγκο δεδομένων που εισέρχεται.

Στο σημείο αυτό, αξίζει να επισημανθεί μια πολύ σημαντική διαδικασία που είναι η ανάλυση των δεδομένων σε πραγματικό χρόνο. Για παράδειγμα, η Walmart είχε την δυνατότητα να παρακολουθεί σε πραγματικό χρόνο τις πωλήσεις ενός δημοφιλούς μπισκότου την περίοδο του Πάσχα. Παρατήρησε ότι το συγκεκριμένο μπισκότο έκανε υψηλές πωλήσεις εκτός από δύο τοποθεσίες όπου δεν πωλούσε καθόλου. Έπειτα από μια ενδελεχή έρευνα στις δύο αυτές περιοχές, διαπιστώθηκε ότι τα μπισκότα δεν είχαν φτάσει ακόμη στα ράφια των μαγαζιών. Αυτή η πληροφορία δεν θα είχε αξία εάν η Walmart εντόπιζε το πρόβλημα μετά την περίοδο του Πάσχα.

Συμπερασματικά, η ταχύτητα παίζει σημαντικό ρόλο στη σωστή λήψη των αποφάσεων για τις εταιρίες. Σύμφωνα με μια επιστολή του CEO της Amazon Jeff Bezos προς τους μετόχους, η «λήψη αποφάσεων υψηλής ταχύτητας» έχει μεγάλη σημαντικότητα στον επιχειρηματικό κλάδο. Μάλιστα, τόνισε ότι οι περισσότερες αποφάσεις πρέπει να παρθούν όταν μια επιχείρηση θα έχει το 70% των πληροφοριών που θα επιθυμούσε. Στις περισσότερες των περιπτώσεων εάν περίμενε το 90% των πληροφοριών θα είχε μείνει πίσω και πιθανότατα θα έπαιρνε μια κακή επιχειρηματική απόφαση. Για αυτό τον λόγο, οι εταιρίες θα πρέπει να εντοπίζουν και να διορθώνουν τις «κακές» αποφάσεις σε πραγματικό χρόνο με στόχο να ελαχιστοποιούν το κόστος.

δ. Ακρίβεια

Ο όρος ακρίβεια περιγράφει το πόσο αξιόπιστο, φερέγγυο και ακριβές είναι ένα σύνολο δεδομένων. Δεν παίζει ρόλο μόνο η ποιότητα των δεδομένων ή το πόσο αξιόπιστες είναι οι πηγές των δεδομένων καθώς είναι σύνηθες να προέρχονται από μη αξιόπιστους παρόχους.

Σύμφωνα πάλι με το Σχήμα 1.1, ένας στους τρείς επιχειρηματικούς ηγέτες για να πάρει τις σωστές επιχειρηματικές αποφάσεις δεν εμπιστεύεται το σύνολο των πληροφοριών που έχει. Μάλιστα, υπολογίζεται ότι η κακή ποιότητα των δεδομένων κοστίζει στην αμερικάνικη οικονομία 3.1 τρισεκατομμύρια δολάρια ετησίως.

Εκτός των τεσσάρων κυρίων «Vs» που αναφέρθηκαν, έχει προστεθεί και ένα άλλο χαρακτηριστικό, η **Αξία (Value)**. Σκοπός κάθε εταιρίας είναι να αξιοποιήσει τα δεδομένα που κατέχει ώστε οι πληροφορίες που θα αντλήσει εξ΄ αυτών να έχουν κάποια αξία και όφελος όπως, για παράδειγμα, για την μελλοντική αύξηση εσόδων.

Ο Γενικός Διευθυντής της ασφαλιστικής εταιρίας MetLife Iberia ισχυρίζεται ότι μέσα σε μια περίοδο δύο μηνών γνωρίζουν πότε υπάρχει υψηλή πιθανότητα ένας πελάτης να ακυρώσει το συμβόλαιό του ή να αγοράσει κάποιο καινούριο. Μέσω της διαδικασίας ανάλυσης των μεγάλων δεδομένων η ασφαλιστική εταιρία μπορεί να εντοπίσει μέσα στο συγκεκριμένο σύνολο πελατών ποιοι είναι οι σημαντικοί πελάτες και ποιοι πελάτες δεν έχουν μεγάλη αξία, είτε επειδή προβαίνουν συχνά σε ακύρωση των συμβολαίων τους για να επωφεληθούν κάποιας έκπτωσης, είτε γιατί είναι ύποπτοι για απάτη. Ωστόσο, μπορεί οι συγκεκριμένοι πελάτες να έχουν παρόμοιο προφίλ αλλά με μια περαιτέρω ανάλυση διαπιστώνεται ότι έχουν διαφοροποιήσεις από τους υπόλοιπους.

Συμπερασματικά, οι επιχειρήσεις που θα αξιοποιήσουν με την σωστή μέθοδο τα Μεγάλα Δεδομένα με χαρακτηριστικά την υψηλή ποσότητα, ταχύτητα, ακρίβεια, αξία και μεγάλη ποικιλομορφία, θα έχουν τη δυνατότητα να γνωρίζουν καλύτερα τις ανάγκες των πελατών. Έτσι, θα μπορούν γρήγορα να προσαρμόσουν τα προϊόντα τους με απώτερο σκοπό την αύξηση των εσόδων τους.

1.3 Μεγάλα Δεδομένα και Ανοιχτά Δεδομένα

Σε γενικά πλαίσια παρατηρείται μια σύγχυση στις έννοιες των Μεγάλων Δεδομένων και των Ανοιχτών Δεδομένων. Μπορεί τα Μεγάλα Δεδομένα και τα Ανοιχτά Δεδομένα να έχουν κάποια κοινά χαρακτηριστικά και να μοιάζουν εκ πρώτης όψεως, όμως διαφέρουν σημαντικά. Σε αυτό το σημείο κρίνεται απαραίτητο να γίνει η διάκριση των δύο εννοιών.

Τα **Ανοιχτά Δεδομένα (Open Data)** αποτελούν δεδομένα στα οποία ο κάθε πολίτης, οργανισμός και εταιρεία έχει την δυνατότητα ελεύθερης πρόσβασης και διαχείρισης οποιαδήποτε στιγμή από οπουδήποτε και αν βρίσκεται. Επίσης, είναι δυνατή η κοινή χρήση τους.

Από σκοπιάς ανάλυσης των δεδομένων, για να χαρακτηριστούν κάποια δεδομένα ανοιχτά πρέπει να υπάρχει η δυνατότητα επαναχρησιμοποίησής τους χωρίς περιορισμούς. Ταυτόχρονα πρέπει να έχουν τέτοια μορφή ώστε να μπορούν να επεξεργαστούν και να αναλυθούν εύκολα από τα υπολογιστικά συστήματα.

Η κύρια διαφοροποίηση των δύο ορισμών εντοπίζεται στο ότι τα Ανοιχτά Δεδομένα ορίζονται συναρτήσει της χρήσης τους, ενώ τα Μεγάλα Δεδομένα ορίζονται από τα χαρακτηριστικά τους, δηλαδή τα «4 V's» (ποσότητα, ποικιλομορφία, ταχύτητα, ακρίβεια). Η χρήση των Ανοιχτών Δεδομένων επιφέρει αρκετά πλεονεκτήματα σε διάφορους τομείς όπως την οικονομία, την κοινωνία αλλά και στην διαφάνεια των κυβερνήσεων.

Τα Μεγάλα Δεδομένα και τα Ανοιχτά Δεδομένα συσχετίζονται και με μια άλλη έννοια, την **Ανοιχτή Διακυβέρνηση (Open Government)**. Η Ανοιχτή Διακυβέρνηση περιγράφεται ως η ελεύθερη πρόσβαση κάθε ενδιαφερόμενου στα αρχεία και τις διαδικασίες της κυβέρνησης με σκοπό την πληροφόρηση αλλά κυρίως την διαφάνεια.

Σύμφωνα με το Σχήμα 1.2, παρατηρείται ότι δημιουργούνται έξι υποενότητες μεταξύ Μεγάλων Δεδομένων, Ανοιχτών Δεδομένων και Ανοιχτής Διακυβέρνησης με κοινά αλλά και ξένα τμήματα. Περαιτέρω επεξήγησης χρήζουν οι παρακάτω τρείς υποενότητες:

Η πρώτη υποενότητα είναι τα Μεγάλα αλλά όχι Ανοιχτά Δεδομένα. Μεγάλο ποσοστό αυτών των δεδομένων αποτελείται από τις πληροφορίες που έχουν οι μεγάλες

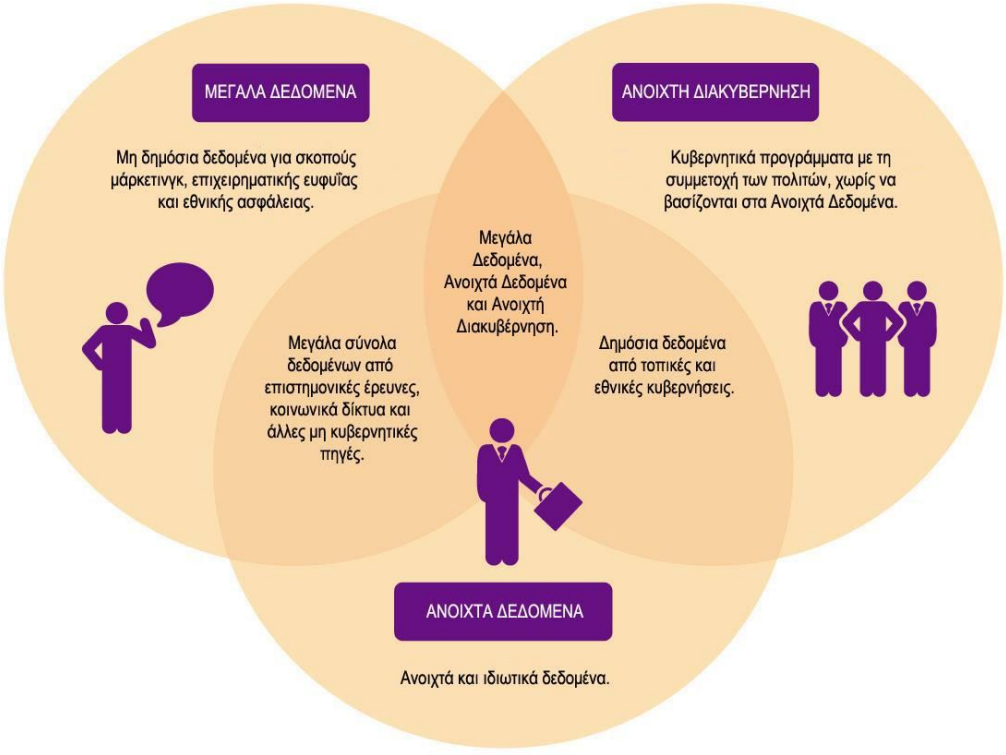
επιχειρήσεις για τους πελάτες τους. Χαρακτηρίζονται ως μη προσιτά σε δημοσιότητα καθώς μπορούν να επωφεληθούν μόνο όσοι έχουν πρόσβαση σε αυτά.

Η επόμενη υποενότητα αποτελείται από τα Μεγάλα και Ανοιχτά Δεδομένα, αλλά όχι Ανοικτής Διακυβέρνησης. Συνήθως, είναι δεδομένα που προέρχονται από ακαδημαϊκές και επιχειρηματικές πηγές για ερευνητικούς, κυρίως, σκοπούς.

Τέλος, η τρίτη υποενότητα αποτελείται από τα συγχρόνως Μεγάλα, Ανοιχτά Δεδομένα και Ανοιχτής Διακυβέρνησης. Αυτός ο συνδυασμός μπορεί να προσφέρει μεγάλα οικονομικά και κοινωνικά οφέλη. Σύμφωνα με την *Πύλη Δημοσίων Δεδομένων της Ευρωπαϊκής Ένωσης (European Data Portal)*, η επαναχρησιμοποίηση Ανοικτών Δεδομένων θα μπορούσε να εξοικονομήσει 7.000 ζωές ετησίως ή να εξοικονομήσει έως και 629 δισεκατομμύρια ώρες στους δρόμους.

Σχήμα 1.2 *Μεγάλα, Ανοιχτά Δεδομένα και Ανοιχτή Διακυβέρνηση*

(Open data, big data and open government: six subtypes of data, 2017)



Συμπερασματικά, τα Μεγάλα και τα Ανοιχτά Δεδομένα είναι δύο μεγάλα «όπλα» για τις κυβερνήσεις, τις επιχειρήσεις και τον κάθε πολίτη. Οι επιχειρήσεις με την ελεύθερη πρόσβαση σε μεγάλα σύνολα δεδομένων, έχουν την δυνατότητα να

βελτιώσουν τις στρατηγικές μάρκετινγκ αλλά και να αυξήσουν τα μελλοντικά τους έσοδα. Αντίστοιχα, οι κυβερνήσεις, έχοντας πρόσβαση σε μεγάλα σύνολα δεδομένων, δύναται να επωφεληθούν οικονομικά αλλά και διοικητικά. Από πλευράς γενικού πληθυσμού, η ελεύθερη πρόσβαση σε όλα αυτά τα δεδομένα από τους πολίτες, ενισχύει την διαφάνεια και την αξιοκρατία σε οικονομικά, πολιτικά και κοινωνικά ζητήματα. Τέλος, μειώνει αισθητά την εμφάνιση κινδύνου κακόβουλης χρήσης των Μεγάλων Δεδομένων είτε από τον επιχειρηματικό κλάδο είτε από τις εκάστοτε κυβερνήσεις.

ΚΕΦΑΛΑΙΟ 2: Τεχνικές ανάλυσης Μεγάλων Δεδομένων

2.1 Εισαγωγή

Στη σύγχρονη πραγματικότητα, σε παγκόσμιο επίπεδο, κάθε μικρός ή μεγάλος οργανισμός εστιάζει στην εύρεση υψηλής ποιότητας δεδομένων με σκοπό την εξαγωγή γνώσης που θα οδηγήσει στις βέλτιστες επιχειρηματικές αποφάσεις. Γενικότερα τα δεδομένα, στην περίπτωση μας τα Μεγάλα Δεδομένα, δεν παράγουν από μόνα τους γνώση. Αντιθέτως, η αξία τους βρίσκεται στην γνώση που μπορεί να αντληθεί μέσω αυτών. Όλη αύτη η διαδικασία πραγματοποιείται μέσω των τεχνικών της **Αναλυτικής Μεγάλων Δεδομένων (Big Data Analytics).**

Η Αναλυτική Μεγάλων Δεδομένων, αναφέρεται ως μια περίπλοκη και πολυσύνθετη διαδικασία διαχείρισης μεγάλων και ποικίλων δεδομένων, με σκοπό τον εντοπισμό χρήσιμων και σημαντικών πληροφοριών, όπως κρυμμένα μοτίβα, άγνωστες συσχετίσεις, προτιμήσεις των πελατών, σχέσεις μεταξύ των μεταβλητών με στόχο την πρόβλεψη μελλοντικών τιμών αλλά και την κατανόηση συμπεριφορών. Με αυτό τον τρόπο, η αξιολόγηση των πληροφορίων, κατευθύνει διάφορους οργανισμούς σε επιχειρηματικές αποφάσεις που βελτιστοποιούν τη χρησιμότητα τους. Πιο αναλυτικά, η μέθοδος περιλαμβάνει την συλλογή, τον καθαρισμό, την οργάνωση, την αποθήκευση και την ανάλυση των δεδομένων μέσα από χρήσιμα εργαλεία και τεχνικές.

Η συνδρομή της υψηλής τεχνολογίας υπολογιστικών συστημάτων, καθώς και, εξειδικευμένων λογισμικών έχει προσφέρει στους αναλυτές τη δυνατότητα επεξεργασίας και εξόρυξης από μεγάλες ποσότητες δομημένων, αδόμητων ή ημι-δομημένων δεδομένων από αρκετές πηγές. Αυτές οι διεργασίες αποφέρουν ένα πλήθος από επιχειρησιακά οφέλη, όπως:

- Πιο αποτελεσματικό Μάρκετινγκ
- Καλύτερη εξυπηρέτηση πελατών
- Νέες ευκαιρίες εσόδων
- Ελαχιστοποίηση κοστών
- Πιο ορθολογικές αποφάσεις, τεκμηριωμένες με στατιστική ισχύ .

2.2 Μεθοδολογίες της Αναλυτικής Μεγάλων Δεδομένων

Η Αναλυτική Μεγάλων Δεδομένων αναφέρεται ως μια σειρά μεθόδων οι οποίες συνήθως έχουν αρκετή δυσκολία στην εκτέλεσή της. Το γεγονός αυτό οφείλεται, κυρίως, στα χαρακτηριστικά των Μεγάλων Δεδομένων που εντοπίζονται αρκετά σε πλήθος αλλά και ταυτοχρόνως διαφορετικά μεταξύ τους. Για την καλύτερη κατανόηση της μεθοδολογίας της Αναλυτικής Μεγάλων Δεδομένων, περιγράφεται σε αυτή την υποενότητα ο προσδιορισμός των σταδίων της.

Το πρώτο στάδιο αποτελεί η λεγόμενη **Επιχειρηματική Αξιολόγηση (Business Case Evaluation)**. Πριν την έναρξη των διαδικασιών ανάλυσης, χρειάζεται σε κάθε περίπτωση να έχουν διευκρινιστεί οι στόχοι, τα κίνητρα της ανάλυσης αλλά και ο σκοπός της μελέτης για τη συγκεκριμένη υπόθεση. Επιπλέον, η Επιχειρηματική Αξιολόγηση συνεισφέρει στο σχεδιασμό ολόκληρης της μεθοδολογίας που θα ακολουθηθεί στα επόμενα βήματα και κυρίως στο αν συμφέρει οικονομικά έναν οργανισμό να προχωρήσει στην υλοποίησή της.

Το δεύτερο στάδιο είναι ο **Εντοπισμός Δεδομένων (Data Identification).** Εκεί διαφαίνονται οι διαδικασίες εύρεσης των συνόλων των δεδομένων αλλά και των πηγών τους (προφανώς χρειάζεται να είναι όσο το δυνατόν πιο έγκυρες). Ακόμη, η ανακάλυψη ποικίλων πηγών δεδομένων αυξάνει κατά πολύ την πιθανότητα παρατήρησης συσχετίσεων και κρυφών προτύπων σε κάποιο ή κάποια από τα χαρακτηριστικά των δεδομένων που έχουν επιλεχθεί.

Το τρίτο στάδιο είναι η **Λήψη και το Φιλτράρισμα Δεδομένων** (**Data Acquisition and Filtering**). Στο στάδιο αυτό, συμπεριλαμβάνονται όλες οι διαδικασίες που αποσκοπούν στη συλλογή των δεδομένων από τις πηγές που λήφθηκαν σε προηγούμενο στάδιο. Κατά αυτό τον τρόπο, τα δεδομένα που έχουν εξορυχθεί υπόκεινται στο απαραίτητο φιλτράρισμα. Το τελευταίο συνδράμει στην απομάκρυνση των αλλοιωμένων, καταστραμμένων δεδομένων, καθώς και, των δεδομένων που δεν έχουν κάποια «σημαντική» αξία για την συγκεκριμένη ανάλυση που πρόκειται να υλοποιηθεί.

Στη συνέχεια, υλοποιείται το στάδιο της **Εξαγωγής Δεδομένων (Data Extraction).** Πολλά από τα δεδομένα που έχουν συλλεχθεί δεν είναι έτοιμα για ανάλυση, καθώς απουσιάζουν από αυτά τα γνωστά χαρακτηριστικά των Μεγάλων

Δεδομένων. Επομένως, τα δεδομένα μετασχηματίζονται και μορφοποιούνται με το βέλτιστο τρόπο, ώστε να είναι έτοιμα για την ανάλυση που απαιτείται.

Το πέμπτο στάδιο αποτελεί η **Επικύρωση και ο Καθαρισμός των Δεδομένων (Data Validation and Cleansing)**. Η πλειονότητα των δεδομένων, και κυρίως των μη δομημένων, είναι αρκετό δύσκολο να προκαθοριστούν και να γίνουν αποδεκτά από κάποιο στάδιο επικύρωσης. Εξαιτίας αυτού του γεγονότος, καθορίζονται πολύπλοκοι κανόνες επικύρωσης και διαγράφονται τυχόν μη έγκυρα δεδομένα, ύστερα από μια πολύ ευαίσθητη και λεπτομερειακή διεργασία. Επιπροσθέτως, η ίδια διαδικασία ακολουθείται για τις τιμές που θεωρούνται λανθασμένες, διπλότυπες ή ελλειπούσες τιμές.

Έπειτα ακολουθεί η **Συγκέντρωση και Αναπαράσταση των Δεδομένων (Data Aggregation and Representation)**. Είναι αντιληπτό πως ένα μεγάλο ποσοστό δεδομένων εξαπλώνεται σε πολλαπλά σύνολα δεδομένων. Επιπλέον, σε διάφορα σύνολα δεδομένων παρατηρείται και η ύπαρξη κοινών πεδίων δεδομένων. Σκοπός αυτού του σταδίου είναι η ενοποίηση των πολλαπλών συνόλων δεδομένων, προκειμένου να είναι διαθέσιμα για οποιαδήποτε επεξεργασία. Η διαδικασία αυτή απαιτεί αρκετό χρόνο και ποικιλότροπες διαδικασίες.

Ένα από τα πιο θεμελιώδη στάδια είναι η **Ανάλυση Δεδομένων (Data Analysis).** Η Ανάλυση Δεδομένων αξιοποιεί πληθώρα τεχνικών ανάλυσης Μεγάλων Δεδομένων και εργαλείων, με σκοπό τον εντοπισμό κρυμμένων μοτίβων και συσχετίσεων, μετατρέποντας έτσι τα δεδομένα σε χρήσιμες πληροφορίες, κομβικές για την επίλυση ζητημάτων. Μια πιο διεξοδική περιγραφή αυτών των τεχνικών ανάλυσης θα γίνει σε επόμενα κεφάλαια.

Ένα, ακόμη, στάδιο πριν το τελικό είναι η **Οπτικοποίηση Δεδομένων (Data Visualization).** Η ικανότητα ανάλυσης τεράστιων ποσοτήτων δεδομένων και εύρεσης χρήσιμων πληροφοριών δεν έχει κάποια αξία εάν οι μόνοι που μπορούν να ερμηνεύσουν τα αποτελέσματα είναι οι αναλυτές. Για αυτό τον λόγο, τα αποτελέσματα από την ανάλυση των δεδομένων, με την χρήση διάφορων τεχνικών και εργαλείων απεικόνισης δεδομένων, παρουσιάζονται σε μορφές δισδιάστατων ή τρισδιάστατων γραφημάτων, ώστε να γίνουν κατανοητά και από άλλους χρήστες ή συμβούλους σε συνέδρια εταιρειών.

Η τελευταία φάση είναι η **Αξιοποίηση των Αποτελεσμάτων της Ανάλυσης (Utilization of Analysis Results)**. Ύστερα από την ανάλυση των αποτελεσμάτων, μπορεί να προκύψουν περαιτέρω ευκαιρίες αξιοποίησης των αποτελεσμάτων της ανάλυσης. Επίσης, όλη η μελέτη που διενεργήθηκε μπορεί να εφαρμοστεί σε νέα προϊόντα μιας εταιρείας ή στην πρόβλεψη οικονομικών δεικτών που θα ωφελήσουν κάποιους επενδυτές είτε γενικότερα στη βελτιστοποίηση του αρχικού πλάνου και γιατί όχι και στην δημιουργία καλύτερων προοπτικών.

2.3 Μηχανική Μάθηση στον αναλογισμό

Η μηχανική μάθηση έχει προσφέρει εξαιρετικές εφαρμογές σε πάρα πολλούς τομείς. Από αυτό δεν θα μπορούσε να διαφύγει η επιστήμη του αναλογισμού που βρίσκει πρόσφορο έδαφος σε πολλές εφαρμογές του, διακρίνοντας τη μηχανική μάθηση σε τρεις θεμελιώδεις συνιστώσες. Αυτές είναι η εποπτευόμενη μάθηση (supervised learning), η μάθηση χωρίς επίβλεψη (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning). Ουσιαστικά η εποπτευόμενη μάθηση αναφέρεται σε εκπαίδευση μέσω επισημάνσεων του μοντέλου ενώ στη μη επιβλεπόμενη η εκπαίδευση υλοποιείται μέσα από μοτίβα των τιμών της μεταβλητής. Η ενισχυτική χρησιμοποιεί τα πλεονεκτήματα και των δύο και τις επικαλείται αρκετά συχνά. Αξίζει να επισημανθεί η προτίμηση των ειδικών αναλογιστών στην επιβλεπόμενη μάθηση.

Μερικά παραδείγματα της επιβλεπόμενης μάθησης που εμφανίζεται σε μεγάλη συχνότητα στον αναλογιστικό τομέα είναι η προσαρμογή (fitting) γραμμικών υποδειγμάτων πάνω σε ανάλογα δεδομένα. Συχνά, αποτελούν αντικείμενο μελέτης και τα γενικευμένα γραμμικά μοντέλα που φαίνεται να ταιριάζουν καλύτερα σε σύνολα δεδομένων αποτελούμενα από αξιώσεις (claims). Ο λόγος που συμβαίνει αυτό, συνήθως, οφείλεται στην ανάγκη για πρόβλεψη των ποσοστών των αξιώσεων αυτών ή του ύψους της σοβαρότητας τους. Ένα ακόμη πλεονέκτημα αντικατοπτρίζεται στην απλότητα του τρόπου λειτουργίας των μοντέλων αυτών είτε με την κλασσική γραμμική παλινδρόμηση είτε με πολυσύνθετες διαδικασίες.

Η μη επιβλεπόμενη μάθηση είναι ο κλάδος της μηχανικής μάθησης που έχει ως αντικείμενο τον εντοπισμό σχημάτων, προτύπων, κρυφών δομών. Σκοπός της διεργασίας αυτής είναι η εύρεση «έξυπνων» μοτίβων με τη χρήση μόνο των δεδομένων της υπό μελέτη μεταβλητής. Με αυτή τη λογική είναι εύκολη η εξόρυξη γνώσης ή ακόμη και να τυποποιηθούν τα δεδομένα για την εύρεση ενός κανόνα πρόβλεψης. Αναφορικά με την μάθηση της ενίσχυσης, για αυτήν δεν θα αναπτυχθούν πολλά καθώς ο αναλογισμός δεν είναι στενά συνδεδεμένος μαζί της.

Σε αυτό το σημείο, τονίζεται η γενική άποψη πως η τεχνητή νοημοσύνη δεν ήρθε να αντικαταστήσει το ρόλο του αναλογιστή. Αντιθέτως, αυτή η πανίσχυρη γνώση μπορεί να λειτουργήσει θετικά στην προσπάθεια και το έργο των αναλογιστών. Ο Nicholas Yeo (2017), ιδρυτής της Nicholas Actuarial Solutions, ισχυρίζεται ότι «η αναλογιστική εργασία θα ολοκληρωθεί γρήγορα, αποτελεσματικά και με ακρίβεια» και ότι «οι αναλογιστές θα ελευθερωθούν από τον περιορισμό των υπολογιστικών διαδικασιών και την παραγωγή συχνών αναφορών (αυτοματοποίηση «έξυπνων» διεργασιών) , ελευθερώνοντας έτσι χρόνο για να επικεντρωθούν σε υψηλής αξίας και ποιότητας δραστηριότητες όπως διορατικές προτάσεις, ανάπτυξη επιχειρήσεων και διαχείριση κινδύνων». Ακόμη σημειώνει ότι, αν και η αυτοματοποίηση των εργασιών έχει ελαχιστοποιήσει τους χρόνους υλοποίησης αναλύσεων και μοντελοποιήσεων συνεισφέροντας και στην ανάδειξη της ποιότητας των αποτελεσμάτων τους, «η αναλογιστική κρίση εξακολουθεί να εφαρμόζεται σε κάθε βήμα της διαδικασίας είτε πρόκειται για χειρισμό δεδομένων, ρύθμιση υπόθεσης ή επιλογή μεθοδολογίας».

Ο αναλογισμός απαιτεί ένα επαρκές σύνολο πληροφοριών και ιδιοτήτων ώστε οι αναλογιστές να είναι σε θέση να λάβουν τις σωστές αποφάσεις και στις κατάλληλες χρονικές στιγμές. Για αυτό και ο ρόλος των αναλογιστών παραμένει αναλλοίωτος και σημαντικός. Όμως, η επαναστατική ιδέα συνδυασμού του αυτοματισμού με την τεχνητή νοημοσύνη μπορεί να αποφέρει εξαιρετικά αποτελέσματα στον τομέα της αναλογιστικής. Πιο συγκεκριμένα, για έναν αναλογιστή φαντάζει απαρχαιωμένο και χρονοβόρο να ελέγχει όλες τις αναλογιστικές διαδικασίες και να υλοποιεί επεμβάσεις σε κάθε βήμα των μεθόδων αυτών. Εδώ ακριβώς εντοπίζεται η αναγκαιότητα της μηχανικής μάθησης, η οποία θα υλοποιεί άμεσα αυτά τα στάδια και ο αναλογιστής μπορεί να συλλέξει το αποτέλεσμα εγκρίνοντάς το ή μη στο τέλος όλων των διεργασιών και αξιολογώντας το με βάση τη δική του εμπειρία και γνώση.

Σύμφωνα με το λόγο του Yeo σε ένα από τα άρθρα του (Yeo, 2017) είχε αναφέρει το εξής: «Με τον αυτοματισμό της διαδικασίας και την τεχνητή νοημοσύνη, ο χρόνος των αναλογιστών δεν θα αφιερώνονταν πλέον άσκοπα στην επεξεργασία δεδομένων και τον περιορισμό των αριθμών. Οι αναλογιστές θα μπορούν πλέον να αφιερώσουν περισσότερο χρόνο εκτελώντας αναλύσεις και να κάνουν συστάσεις σε τομείς που κατανοούν καλά, όπως η διαχείριση χρηματοοικονομικών προϊόντων και η μετρίαση του επιχειρηματικού κινδύνου».

Μια ειδική εφαρμογή της Τεχνητής Νοημοσύνης αποτελεί το σκάκι. Πολλοί κορυφαίοι της επιστημονικής κοινότητας της Τεχνητής Νοημοσύνης εκτιμούσαν πως οι υπολογιστές θα μπορούσαν με την εξέλιξη της τεχνολογίας να νικούν κορυφαίους παίχτες του σκακιού. Αρκετά χρόνια αργότερα διατυπώθηκε ένα αξιοσημείωτο αποτέλεσμα στο οποίο εκφέρεται η άποψη ότι ο βέλτιστος παίχτης σκακιού δεν είναι ούτε άνθρωπος ούτε μηχανή αλλά ένας βέλτιστος συνδυασμός τους. Η επιβεβαίωση ήρθε με την αναπάντεχη νίκη δύο άπειρων παιχτών που με τη βοήθεια της τεχνητής νοημοσύνης και τις δικές τους επιλογές κατάφεραν να αναδειχθούν νικητές. Μάλιστα, ένα εκ των νικητών είχε εκφράσει την αποτελεσματικότητα της στρατηγικής άλλοτε να ακολουθούν τη βέλτιστη επιλογή του υπολογιστή και άλλοτε να προσφεύγουν στην προσωπική τους άποψη, κάτι που τους ώθησε στο βέλτιστο αποτέλεσμα, δηλαδή στις νίκες.

Η παραπάνω διαπίστωση βοηθά στην κατανόηση του συσχετισμού της ανθρώπινης συνείδησης με την τεχνητή νοημοσύνη. Ίδιες προοπτικές προβλέπονται και για τον αναλογισμό. Ουσιαστικά, η αναλογιστική επιστήμη είναι εφικτό να μετατραπεί σε πιο απλή υπόθεση και πιο κύρια, καθώς οι αναλογιστές μπορούν να διαχειρίζονται το εργαλείο της μηχανικής μάθησης για την επιτάχυνση και αύξηση της ποιότητας των οπτικοποιημένων ενδείξεών τους κάνοντας χρήση της τεχνολογίας της τεχνητής νοημοσύνης. Είναι σημαντικό να βρεθεί ο σωστός συνδυασμός του αναλογιστή με τη μηχανική γνώση και την τεχνητή νοημοσύνη γενικότερα, με το ρόλο του αναλογιστή να παραμένει σημαντικός και κομβικός.

Γενικώς σε αυτή τη ραγδαία εξέλιξη της αναλογιστικής επιστήμης με χρήση μηχανικής γνώσης, έχουν προταθεί τρεις πιθανές ενδεικτικές καινοτόμες κατευθύνσεις. Σύμφωνα με την πρώτη είναι επιθυμητή η συνδρομή της τεχνητής νοημοσύνης και ειδικότερα της μηχανικής μάθησης με σκοπό τη μείωση του χρόνου εκτέλεσης βασικών λειτουργιών. Η τεχνολογία της μηχανικής μάθησης μπορεί να εκπαιδευτεί και να

πραγματοποιήσει βελτιώσεις γρηγορότερα από οποιοδήποτε ανθρώπινο νου, κάτι που βοηθά σε υψηλό βαθμό τις αναλογιστικές μεθόδους πρόβλεψης. Επίσης, αυτή η μείωση του χρόνου εκτέλεσης διεργασιών αυτομάτως ελαχιστοποιεί και το κόστος.

Μια άλλη οπτική αναφέρεται στη δυνατότητα υποβοήθησης ως προς τις δεξιότητες και το παραγόμενο επιστημονικό τους έργο. Καθώς η μηχανική μάθηση πραγματοποιεί άριστα εργασίες ρουτίνας, ο αναλογιστής λαμβάνει μεγάλο μερίδιο σε αυτή τη διαδικασία και παρεμβαίνει όποτε χρειαστεί. Ουσιαστικά, η όλη τεχνική επιχειρεί να βελτιώσει και να εκσυγχρονίσει το αξίωμα του αναλογιστή.

Η τρίτη πτυχή είναι η βελτιστοποίηση της αναλογιστικής με ορίζοντα τις τεχνικές της μηχανικής μάθησης και της τεχνητής νοημοσύνης. Αυτή η ιδέα, δηλαδή ότι οι αναλογιστές μπορούν να λειτουργούν παράλληλα με τις νέες τεχνικές αγγίζει πιο πολύ τις προσωπικές τους επιδιώξεις. Οι αναλογιστές θα διαθέτουν αρκετό χρόνο για να ενασχοληθούν με πιο ουσιαστικά γεγονότα. Όλα αυτά βοηθούν όμως στην εγκαιρότητα και εγκυρότητα των αναμενόμενων αποτελεσμάτων που κάθε ασφαλιστική εταιρεία θα ήθελε να επιτύχει.

Αξίζει να σημειωθεί μια σχετική έρευνα που είχε διενεργηθεί με σκοπό την εύρεση ποσοστών χρήσης τέτοιων τεχνικών από αναλογιστές. Ένα αντιπροσωπευτικό δείγμα αποτελούμενο από ειδικούς της αναλογιστικής επιστήμης ρωτήθηκε αν η δουλειά τους συσχετίζεται στενά με την ιδέα της μηχανικής μάθησης. Το 54% απάντησε θετικά ενώ οι υπόλοιποι δεν φαίνεται να είχαν εκσυγχρονιστεί ακόμη με τις μεθόδους της μηχανικής μάθησης (βλ. Riley (2020)).

ΚΕΦΑΛΑΙΟ 3: Τεχνικές ταξινόμησης

3.1 Εισαγωγή

Η ταξινόμηση αποτελεί βασική διενέργεια στην Ανάλυση Δεδομένων αλλά και στη Μηχανική Μάθηση. Χρησιμοποιείται σε δομημένα ή μη δομημένα δεδομένα. Στην ουσία, η ταξινόμηση αποτελεί μια τεχνική όπου κατηγοριοποιούμε τα δεδομένα σε κλάσεις, οι οποίες έχουν εκ των προτέρων καθοριστεί. Ο κύριος στόχος ενός προβλήματος ταξινόμησης είναι ο προσδιορισμός της κατηγορίας σε νέες παρατηρήσεις, στις οποίες δεν γνωρίζουμε σε ποια κλάση ανήκουν εκ των προτέρων. Στις επόμενες υποενότητες αναλύονται βασικές τεχνικές στην ταξινόμηση δεδομένων, αφού πρώτα περιγράφονται τα μέτρα απόστασης που λειτουργούν ως μετρικές στην πλειονότητα των μεθόδων, κυρίως για σκοπούς ακρίβειας του μοντέλου και ελαχιστοποίησης των σφαλμάτων τους.

3.2 Μέτρα απόστασης

Στο συγκεκριμένο υποκεφάλαιο επιχειρείται η επισκόπηση βασικών τρόπων μέτρησης «αποστάσεων» των δεδομένων. Η έννοια της μετρικής πηγάζει από την επιστήμη των Μαθηματικών, όπως η Γεωμετρία αλλά και η Άλγεβρα. Ο σκοπός αυτών των τύπων είναι να δημιουργηθεί ένα μέτρο που θα συνδράμει στην προσπάθεια να αντιμετωπιστούν τα δεδομένα στη συνέχεια της εργασίας με κάποιο κοινό συγκριτικό κανόνα. Τα ακόλουθα μέτρα έχουν χρησιμοποιηθεί εκτενώς σε πολλές περιπτώσεις, και είναι πολύ συνηθισμένα σε περιπτώσεις ταξινόμησης αλλά και παλινδρόμησης (classification, regression):

- *Ευκλείδεια απόσταση*, είναι ένα μέτρο που ορίζεται σε ευκλείδειο χώρο, τον R^N, και προσδιορίζεται με τον παρακάτω τύπο στην περίπτωση δύο παρατηρήσεων i, j

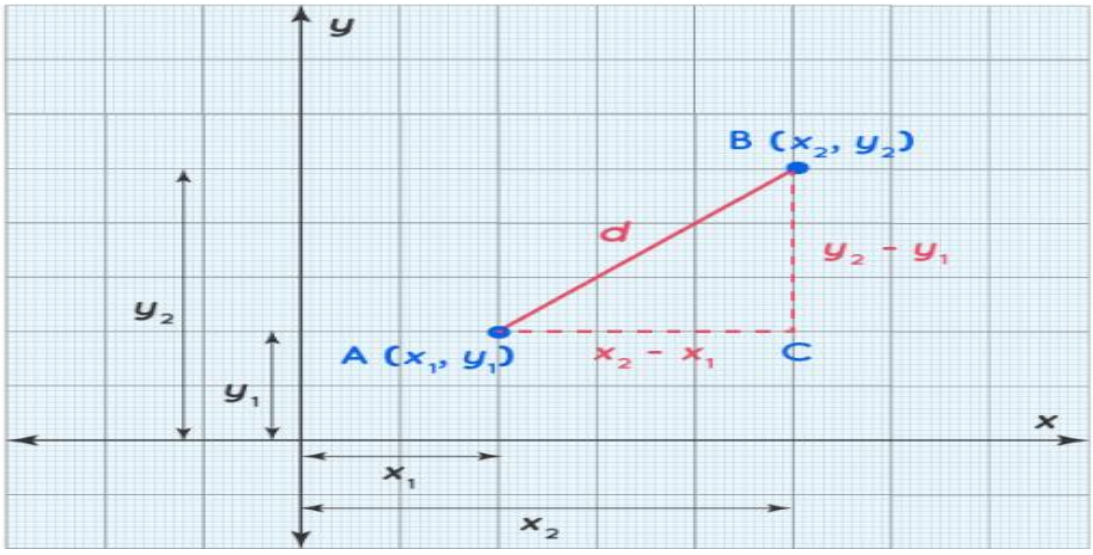
$$D(\boldsymbol{y_i}, \boldsymbol{y_j}) = \left\{\sum_{n=1}^{N}\left(y_n^i - y_n^j\right)^2\right\}^{\frac{1}{2}},$$

όπου το Ν είναι το πλήθος των χαρακτηριστικών και τα $\boldsymbol{y_i, y_j}$ είναι διανύσματα του R^N. Σύμφωνα με τον τύπο, για σταθερό n κάθε φορά η σύγκριση πραγματοποιείται ανά διάσταση, δηλαδή μεταξύ των y_n^i και του y_n^j.

Η Ευκλείδεια απόσταση έχει την απαρχή της στην Ευκλείδεια Γεωμετρία. Στο συγκεκριμένο πεδίο των Μαθηματικών, τα διανύσματα είναι δύο διαστάσεων, μήκος και πλάτος. Έστω $\alpha = (x_1, y_1)$ $και$ $\beta = (x_2, y_2)$ και ο τύπος καθορίζεται από την ακόλουθη σχέση :

$$D(\alpha, \beta) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Ο συγκεκριμένος τύπος αποδεικνύεται εύκολα με χρήση κανόνων Ευκλείδειας Γεωμετρίας. Με τα σημεία στο χώρο α = $A(x_1, y_1)$, β = $B(x_2, y_2)$ και γ = $C(x_2, y_1)$, δημιουργείται ένα ορθογώνιο τρίγωνο και στη συνέχεια αφού τηρούνται οι υποθέσεις του Πυθαγορείου θεωρήματος, εφαρμόζεται και το αποτέλεσμα είναι άμεσο.



Σχήμα 3.1 *Ευκλείδια απόσταση στον χώρο 2-D («Euclidean Distance Formula», n.d)*

Στα προβλήματα με πολλές διαστάσεις και χαρακτηριστικά φαίνεται πως η ευκλείδεια απόσταση βρίσκει εφαρμογή συνεχώς, καθώς αποτελεί μια σημαντική μετρική της ομοιότητας δύο αντικειμένων με ποσοτικά χαρακτηριστικά.

Βέβαια, δεν είναι λίγες οι φορές που η Ευκλείδεια απόσταση εφαρμόζεται και σε κατηγορικά δεδομένα αλλά συνήθως σε αυτές τις περιπτώσεις η ερμηνεία είναι διαφορετική. Για παράδειγμα, σε δυαδικά δεδομένα, είναι εύκολο να χαρακτηριστούν

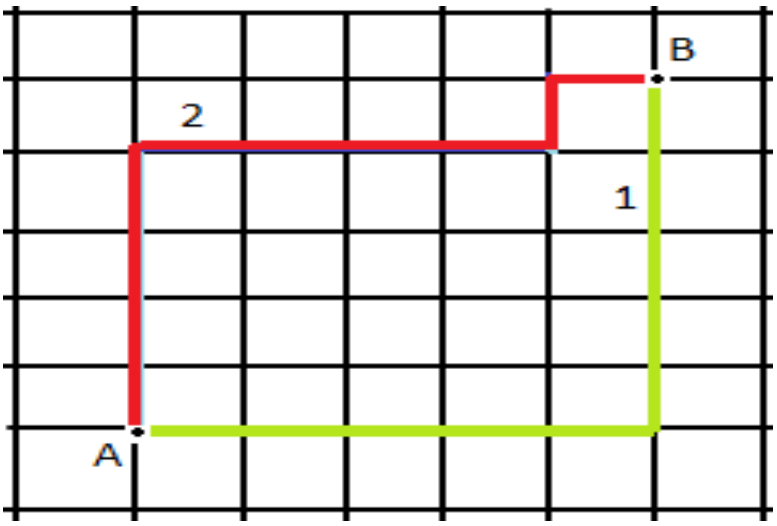
όμοια. Στην περίσταση αυτή, αν η διαφορά των τιμών τους είναι 0 αναφέρονται ως όμοια ενώ το αντίθετο ισχύει αν η διαφορά των τιμών τους είναι 1. Αλλά στην πλειονότητα των πειραμάτων οι μεταβλητές μπορεί να είναι αριθμητικές, κατηγορικές, δυαδικές ή ονομαστικές.

- *Απόσταση Manhattan,* αποτελεί, όπως και η Ευκλείδεια απόσταση, ένα μέτρο που ορίζεται σε ευκλείδειο χώρο, τον R^N, και υπολογίζεται με τον παρακάτω τύπο στην περίπτωση δύο παρατηρήσεων *i, j* :

  $$D(\boldsymbol{y_i}, \boldsymbol{y_j}) = \sum_{n=1}^{N} |y_n^i - y_n^j|,$$

  όπου το N είναι το πλήθος των χαρακτηριστικών, και τα $\boldsymbol{y_i}, \boldsymbol{y_j}$ διανύσματα N-διαστάσεων.

Η διαισθητική ερμηνεία της αποστάσεως Manhattan είναι πιο αλγεβρική σε σχέση με την Ευκλείδεια απόσταση. Αυτό γίνεται αντιληπτό μέσα από ένα απλό παράδειγμα. Στο χώρο των πραγματικών αριθμών, δηλαδή στον R, όταν δίδονται δύο αριθμοί *x, y* η απόσταση τους ορίζεται ως η απόλυτη τιμή της διαφοράς τους. Αυτό το αποτέλεσμα εμπλουτίζει η Manhattan μετρική, καθώς σε δύο πολυδιάστατα διανύσματα λαμβάνει το άθροισμα των απόλυτων αποκλίσεων της κάθε συντεταγμένης τους με την αντίστοιχη τιμή του άλλου ανύσματος.



Σχήμα 3.2 *Απόσταση Manhattan σε διαδρομές (Gohrani, 2019)*

Στην παραπάνω εικόνα, φαίνεται ένα πλέγμα διαχωρισμένο με γραμμές. Σε κάθε κελί αντιστοιχεί ένα κτίριο και οι γραμμές πλέγματος θεωρούνται ως μονοπάτια - δρόμοι. Τώρα, σε περίπτωση που επιθυμεί κάποιος να μετακινηθεί από το σημείο Α στο σημείο Β μπορεί να ακολουθήσει την κόκκινη ή την κίτρινη διαδρομή. Είναι προφανές ότι η διαδρομή δεν είναι ευθεία και υπάρχουν στροφές για να μεταφερθεί στο σημείο τερματισμού. Τότε, συνιστάται η μέτρηση απόστασης με τη Manhattan για να υπολογιστεί η απόσταση που διανύθηκε.

Η απόσταση Minkowski αποτελεί μια γενίκευση των δύο προηγούμενων μετρικών αποστάσεων και χρησιμεύει με τη σειρά της ως μέτρηση απόστασης. Ο τύπος για δύο N-διάστατα διανύσματα $\boldsymbol{y_i}, \boldsymbol{y_j}$ και διάστασης p είναι ο εξής :

$$d(\boldsymbol{y_i}, \boldsymbol{y_j}) = \left(\sum_{i=1}^{n} |y_i - y_j|^p \right)^{\frac{1}{p}}$$

Από την απόσταση Minkowski λαμβάνουμε ως ειδικές περιπτώσεις για $p = 1$ την απόσταση Manhattan και για $p = 2$ την Ευκλείδεια μετρική αποστάσεων.

Υποθέτουμε ένα πρόβλημα με υψηλή διάσταση d των χαρακτηριστικών – ανεξάρτητων μεταβλητών ενός δείγματος του. Δύναται να είναι προτιμότερο να χρησιμοποιούνται χαμηλότερες τιμές p στην απόσταση Minkowski. Αυτό υποδηλώνει ότι η μέτρηση απόστασης Manhattan είναι η πλέον προτιμώμενη για εφαρμογές υψηλής διάστασης. Έτσι, προτιμάται σε σχέση με τη μετρική Ευκλείδειας απόστασης καθώς αυξάνεται η διάσταση των δεδομένων. Αυτό συμβαίνει λόγω ενός φαινομένου, που είναι ευρέως γνωστό ως η «κατάρα της διάστασης» και δημιουργεί προβλήματα στη σύγκλιση των μεθόδων κάτι που επιφέρει επιπτώσεις και στην ακρίβεια των εκτιμήσεων τους.

Το κοινό των δύο αποστάσεων είναι ότι χρησιμοποιούνται ευρέως στην ανάλυση δεδομένων για τον έλεγχο σφαλμάτων ή τη δημιουργία κριτηρίων ομαδοποίησης παρατηρήσεων ως προς κάποια προκαθορισμένα χαρακτηριστικά τους.

3.3 Η τεχνική k-nearest Neighbor

Η μέθοδος του πλησιέστερου γείτονα (KNN) είναι πολύ απλή και διαισθητική. Αρχικά, θεωρούμε σημεία στο χώρο που παρατηρούνται το ένα κοντά στο άλλο. Τα συγκεκριμένα θα πρέπει να έχουν παρόμοια απόκριση ή τιμή ως προς κάποιο χαρακτηριστικό (output), το οποίο ορίζεται μέσα από υπό μελέτη μεταβλητή-χαρακτηριστικό. Η μέθοδος KNN εκπαιδεύεται μέσα από αυτή την πληροφόρηση δηλαδή τις σωστές διατάξεις των παρατηρούμενων και έπειτα ταξινομεί μια νέα παρατήρηση, σύμφωνα με τις τιμές που είναι γνωστές από τα σημεία των k πλησιέστερων γειτόνων του, όπου το k είναι το πλήθος των γειτονικών σημείων.

Η δυσκολία είναι στην προσέγγιση των γειτόνων της νέας παρατήρησης. Αυτό μπορεί να επιτευχθεί μέσω ενός κριτηρίου που να αποφασίζει αν μια παρατήρηση αποτελεί γείτονα ή όχι της νέας παρατήρησης. Προφανώς μια τέτοια λειτουργία προϋποθέτει μια έννοια απόστασης. Σε αυτό βοηθάει η επιλογή της κατάλληλης μετρικής όπως αναπτύχθηκε στην Ενότητα 3.2 η οποία θα αποφασίζει αν ένα σημείο είναι γειτονικό, κάτι που ικανοποιείται όταν απέχει ελάχιστα ως προς την υποκείμενη μετρική. Ιδιαίτερη προσοχή συνιστάται στο πως ορίζεται μια γειτονιά, δηλαδή με χρήση ποιας μετρικής. Η πιο δημοφιλής μέτρηση απόστασης είναι η Ευκλείδεια απόσταση. Η περιοχή της γειτονιάς συνήθως ελέγχεται από μια παράμετρο, k, η οποία μπορεί να επιλεγεί με διασταυρούμενη επικύρωση (cross validation), δηλαδή διάφορες δοκιμές ώστε να καταλήξουμε στον αριθμό γειτόνων που μεγιστοποιεί την ακρίβεια του μοντέλου. Έτσι, δύναται να προβλέπει την τιμή κάποιου χαρακτηριστικού για μια νέα παρατήρηση και με το ελάχιστο σφάλμα.

Αν και απλή στις διαδικασίες της, οι k- πλησιέστεροι γείτονες είναι μια πανίσχυρη μέθοδος, ειδικά για εκείνα τα δεδομένα που εισέρχονται σε ορισμένους τύπους ειδικής τοπικής δομής. Από την άλλη πλευρά έχει και κάποια προβλήματα.

Η Ευκλείδεια απόσταση μπορεί να επηρεαστεί σε μεγάλο βαθμό από στατιστικά ασήμαντες επεξηγηματικές μεταβλητές. Αυτό το πρόβλημα είναι ιδιαίτερα σοβαρό σε χώρους υψηλών διαστάσεων, καθώς μη σημαντικά χαρακτηριστικά επηρεάζουν το αποτέλεσμα της Ευκλείδειας απόστασης. Ο Wang (2005) πρότεινε μια μέθοδο SN (Signal-to-Noise ratio), η οποία προτρέπει στην κατασκευή πολλαπλών ταξινομητών από ένα πλήθος υποσυνόλων μεταβλητών. Στη συνέχεια, λαμβάνονται όλες οι

προβλέψεις υπό την επίδραση μιας συνισταμένης τους σε έναν ταξινομητή. Μάλιστα, διατύπωσε ότι τα υποσύνολα που έδρασε η μέθοδος των KNN είναι σε θέση να αντιμετωπίσουν αρκετά ικανοποιητικά πολλούς προγνωστικούς παράγοντες.

Η μέθοδος μπορεί να αποτύχει πλήρως, όταν χρειάζεται να μελετηθούν κατηγορικά δεδομένα ή ένας συνδυασμός τόσο συνεχών όσο και ποιοτικών χαρακτηριστικών. Αυτό το θέμα δύναται να λυθεί μέσω του ορισμού μιας κατάλληλης μετρικής για αυτές τις δύο περιπτώσεις.

Για έναν καθορισμένο αριθμό πλησιέστερων γειτόνων, k, η ακτίνα της γειτνίασης αποτελεί σημαντικό παράγοντα που καθορίζει την αποτελεσματικότητα της μεθόδου. Το γεγονός αυτό, δημιουργεί την αίσθηση ότι η συχνότητα των τιμών των παρατηρήσεων που παρουσιάζονται σε κάθε υποσύνολο των δεδομένων χρειάζεται να ελέγχεται πριν την ανάλυση του αλγορίθμου. Ουσιαστικά, προκύπτει μια σοβαρή ένδειξη, ότι μια γειτονιά εξαρτάται σημαντικά από τον αριθμό των γειτόνων k αλλά και από την ακτίνα της γειτονιάς. Αυτό προσπάθησε να αντιμετωπιστεί στο άρθρο του Wang, (2005). Ο Wang εισήγαγε την ιδέα των δοκιμών σχετικά με την επιλογή του k, όπου θα μεταβάλλεται ανάλογα της πυκνότητας της συγκεκριμένης γειτονιάς αλλά και του υπο-συνόλου εκπαίδευσης των δεδομένων (train set).

Η μέθοδος KNN μπορεί να διαφοροποιηθεί όταν χρησιμοποιείται σε πραγματικά μεγάλα δεδομένα δηλαδή διαφορετικός αλγόριθμος, άλλα βήματα της μεθόδου. Ας θεωρήσουμε ότι έχουμε ένα σύνολο R δεδομένων και ένα σύνολο S από το οποίο πρέπει να προκύψει η επιλογή των k πλησιέστερων γειτόνων κάθε σημείου στο R. Σε πρώτη φάση, επιλέγονται τυχαία σημεία από τα R και S και καθορίζονται οι διακεκριμένες ομάδες R_i και S_i, όπου στοιχεία του ενός συνόλου αντιστοιχίζονται με μοναδικό τρόπο στο άλλο. Στη δεύτερη φάση, τα R_i και S_i χωρίζονται σε μικρότερες υποομάδες και υπολογίζεται ένα υποψήφιο k πλησιέστερο γειτονικό σύνολο για κάθε αντικείμενο $r \in R_i$. Στην τρίτη φάση διατηρούνται οι k πλησιέστεροι γείτονες με την καλύτερη αναλογία από το υποψήφιο σύνολο.

Οι καλύτεροι γείτονες από τις πιθανές k πλειάδες σύνδεσης από το R στο S με τις υψηλότερες βαθμολογίες, όπου η βαθμολογία μιας πλειάδας καθορίζεται από μια συνάρτηση που χρησιμοποιείται από τον ερευνητή και λειτουργεί καλά πάνω στις μεταβλητές των συνόλων R και S. Οι κορυφαίες πλειάδες k γειτόνων για κάθε σημείο του συνόλου R πραγματοποιείται σε τρία στάδια. Σε πρώτο στάδιο, τα δεδομένα

εισόδου αναδιανέμονται με ίσες βαρύτητες ως προς το μέτρο χρησιμότητας που εφαρμόζεται. Αυτό επιτυγχάνεται με τη δειγματοληψία και την εκτίμηση της επιλεκτικότητας, που χρησιμεύει ως χαμηλότερο όριο για τη βαθμολογία (score) της καλύτερης πλειάδας των k-γειτόνων για κάθε σημείο. Στο δεύτερο στάδιο, η πραγματική συσχέτιση υπολογίζεται μεταξύ των νέων διαμερίσεων R_i, S_i που έχουν προκύψει από τις προηγούμενες επιλογές πλειάδων γειτόνων. Τέλος, στο τρίτο στάδιο, τα ενδιάμεσα αποτελέσματα που προκύπτουν από την επιλογή των ομάδων των πλειάδων των k- πλησιέστερων γειτόνων παράγουν το τελικό αποτέλεσμα κορυφαίας πλειάδας k-πλησιέστερων γειτόνων για το κάθε σημείο του συνόλου R από το σύνολο S.

Στο σημείο αυτό επισυνάπτεται ο αλγόριθμος που θα ακολουθηθεί για τη μέθοδο αυτή στη γλώσσα προγραμματισμού R:

Πρόγραμμα 1 : Μέθοδος του κοντινότερου γείτονα

```
## Φορτώνονται τα δεδομένα που θα χρησιμοποιηθούν στην εφαρμογή
df < - data
## Παρατηρείται η δομή των δεδομένων
head(data)
## Γεννιέται ένας τυχαίος αριθμός, ο οποίος είναι το 90% του συνολικού αριθμού των παρατηρήσεων.
ran <- sample(1:nrow(df), 0.9 * nrow(df))
 ## Δημιουργία συνάρτησης κανονικοποίησης των υπό εξέταση δεδομένων
 nor <- function(x) {
(x -min(x))/(max(x)-min(x))
  }
## Εφαρμόζεται η συνάρτηση κανονικοποίησης στις στήλες m που ενδέχεται να αποτελέσουν τα χαρακτηριστικά – μεταβλητές που μας απασχολούν
df_norm <- as.data.frame(apply(df[,c(1,2,…,m)], nor))
```

Εμφάνιση περιγραφικών στατιστικών στοιχείων των κανονικοποιημένων δεδομένων

```
summary(df_norm)
```

Εξαγωγή του συνόλου εκπαίδευσης

```
df_train <- df_norm[ran,]
```

Εξαγωγή του δοκιμαστικού συνόλου

```
df_test <- iris_norm[-ran,]
```

Εξάγεται η στήλη 8 με σκοπό να χρησιμοποιηθεί ως στόχος για την εκτέλεση της μεθόδου knn

```
df_target_category <- df[ran, 8]
```

Αφαιρείται η στήλη 8 και από το σύνολο ελέγχου με σκοπό να μετρηθεί η ακρίβεια του αλγορίθμου.

```
df_test_category <- iris[-ran,8]
```

Κάλεσμα της βιβλιοθήκης class που εμπεριέχει τη συνάρτηση της μεθόδου kNN.

```
library(class)
```

Εφαρμόζεται η συνάρτηση knn για τα προαναφερθέντα σύνολά μας

```
pr <- knn(df_train,df_test,cl=df_target_category,k=10)
```

Δημιουργία πίνακα που παρουσιάζει την εκτίμηση και την πραγματική τιμή σε κάθε παρατήρηση.

```
tab <- table(pr,df_test_category)
```

Η παρακάτω συνάρτηση διαχωρίζει τις σωστές προβλέψεις από το συνολικό αριθμό των προβλέψεων, εκπροσωπώντας την ακρίβεια του μοντέλου υπό τη δράση του kNN

```
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
```

```
accuracy(tab)
```

3.4 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση αποτελεί ένα στατιστικό μοντέλο που στη συνήθη μορφή του χρησιμοποιείται μια γραμμική συνάρτηση για την προσαρμογή ενός μοντέλου στα δεδομένα (fitting model) μιας δίτιμης εξαρτημένης μεταβλητής, αν και υπάρχουν διάφορες παραλλαγές της μεθόδου. Στη μαθηματική ιδέα, ένα δυαδικό λογιστικό μοντέλο έχει μια εξαρτημένη μεταβλητή με δύο πιθανές τιμές, όπως επιτυχία ή αποτυχία που αντιπροσωπεύεται από δύο τιμές "0" και "1" . Στο λογιστικό μοντέλο, ο λογάριθμος λόγω πιθανότητας (log-odds) για την τιμή με την ένδειξη "1" είναι ένας γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών. Για παράδειγμα, αν υπάρχουν δύο επεξηγηματικές μεταβλητές και ορίζεται ως πιθανότητα επιτυχίας το p και αποτυχίας $1 - p$ (απόδοση ή odd) τότε ισχύει ότι :

$$log\left(\frac{p}{1-p}\right) = b_0 + b_1 * X_1 + b_2 * X_2$$

Οι ανεξάρτητες μεταβλητές μπορεί να είναι είτε δυαδικές μεταβλητές (δύο ομάδες, με τιμές κωδικοποιημένες) ή συνεχείς μεταβλητές (οποιαδήποτε τιμή σε ένα διάστημα). Η συνάρτηση που μετατρέπει αυτόν το λογάριθμο σε πιθανότητα είναι η λογιστική συνάρτηση, γι' αυτό και έχει χρησιμοποιηθεί η συγκεκριμένη ορολογία. Το κυριότερο σημείο αναφοράς του λογιστικού μοντέλου ερμηνεύεται ως εξής : η άνοδος της τιμής μίας από τις ανεξάρτητες μεταβλητές αποτυπώνεται ως αύξηση στο λόγο των πιθανοτήτων επιτυχίας - αποτυχίας, δηλαδή στο κλάσμα με αριθμητή την πιθανότητα επιτυχίας p του προσδοκώμενου αποτελέσματος και παρανομαστή την πιθανότητα αποτυχίας $1 - p$ με ένα σταθερό (fixed) ρυθμό.

Στη συνήθη λογιστική παλινδρόμηση, η εξαρτημένη μεταβλητή έχει πάντα δύο κατηγορίες (κατηγορική μεταβλητή). Η περίπτωση υπάρξεως μεταβλητών με περισσότερες από δύο τιμές, αντιμετωπίζεται με χρήση πολυμεταβλητής λογιστικής παλινδρόμησης. Το μοντέλο λογιστικής παλινδρόμησης υπολογίζει την πιθανότητα της πραγματοποίησης δοθέντος ορίσματος εισόδου και δεν πραγματοποιεί ταξινόμηση, παρόλο που μπορεί να εκτελέσει σε κάποιες εφαρμογές τον ρόλο ενός δυαδικού ταξινομητή. Βασική διαφορά με τη μέθοδο γραμμικών ελαχίστων τετραγώνων είναι η μη ύπαρξη κλειστού τύπου για την εύρεση των συντελεστών της παλινδρόμησης. Η λογιστική παλινδρόμηση ως γενικό μοντέλο στατιστικής ανάλυσης, αναπτύχθηκε

κυρίως από τον Joseph Berkson (1944) , ο οποίος κατασκεύασε το "logit", δηλαδή την περίπτωση του λογαρίθμου του λόγου των πιθανοτήτων επιτυχίας-αποτυχίας που προσαρμόζεται με το γραμμικό μοντέλο.

Η λογιστική παλινδρόμηση εφαρμόζεται σε διάφορες πτυχές, με κυριότερες τη μηχανική μάθηση, την ιατρική και τις κοινωνικές επιστήμες (λόγω της συνήθους χρήσεως κατηγορικών δεδομένων στα πεδία των θεωρητικών επιστημών). Η πρόβλεψη της θνησιμότητας σε τραυματίες ασθενείς, ο κίνδυνος ανάπτυξης μιας δεδομένης ασθένειας (λ.χ. διαβήτης , στεφανιαία νόσος), μέσω κάποιων μεταβλητών-χαρακτηριστικών του ασθενούς (όπως ηλικία, φύλο, δείκτης μάζας σώματος, δείκτες αίματος) αποτελούν κλασσικές περιπτώσεις λογιστικής παλινδρόμησης. Στις δημοσκοπήσεις μπορεί να συσχετισθεί η πρόθεση ψήφου των πολιτών με την ηλικία, το εισόδημα, το φύλο, τη φυλή, ψήφους στο παρελθόν και άλλων παραγόντων. Άλλες κατευθύνσεις που χρησιμοποιούνται είναι το μάρκετινγκ, εκτιμώντας τη διάθεση ενός πελάτη να προβεί σε αγορά του υποκείμενου προϊόντος, αλλά και σε γενικότερα πλαίσια.

Στο σημείο αυτό, θα αναλυθεί το πως λειτουργεί η λογιστική παλινδρόμηση, εξετάζοντας ένα μοντέλο για να εκτιμηθούν οι συντελεστές από τα δεδομένα. Θεωρείται ένα μοντέλο με δύο παράγοντες, Α και Β ως τις ανεξάρτητες μεταβλητές, και μία εξαρτημένη Bernoulli Β τυχαία μεταβλητή , μέσω της οποίας ορίζεται η πιθανότητα επιτυχίας :

$$p = P(Y = 1)$$

και η συμπληρωματική της

$$q = 1 - p = P(Y = 0).$$

Εικάζουμε γραμμική σχέση μεταξύ των μεταβλητών των log-odds του συμβάντος και των παραγόντων Α,Β :

$$log\left(\frac{p}{1-p}\right) = b_0 + b_1 * A + b_2 * B$$

και μετά από πράξεις καταλήγουμε ότι

$$p = \frac{e^k}{e^k + 1} \, με \, k = b_0 + b_1 * A + b_2 * B$$

Μέσω αυτής της γραμμικής σχέσης μπορεί να ερμηνευθεί ως εξής : αυξάνοντας την τιμή του παράγοντα Α κατά μια μονάδα διατηρώντας συγχρόνως σταθερές τις υπόλοιπες τιμές έχουμε

$$p_{new} = \frac{e^{k_{new}}}{e^{k_{new}} + 1} \; \mu\varepsilon \; k_{new} = b_0 + b_1 * (A + 1) + b_2 * B = k + b_1$$

Άρα ο λόγος

$$\frac{p_{new}}{1 - p_{new}} = k + b_1,$$

δηλαδή αυξάνεται ή μειώνεται ανάλογα με το εάν το b_1 είναι θετικό ή αρνητικό αντίστοιχα.

Το θεωρητικό υπόβαθρο που περιγράφτηκε στις προηγούμενες παραγράφους σχετικά με τη Λογιστική Παλινδρόμηση θα εφαρμοστεί αλγοριθμικά στην προσπάθεια να υλοποιηθεί ένα πείραμα σε μια βάση δεδομένων με τη συνδρομή της γλώσσας προγραμματισμού Python.

Πρόγραμμα 3 : Μέθοδος Λογιστικής Παλινδρόμησης

```
!pip3 install sklearn  ## κατέβασμα της βιβλιοθήκης sklearn

## καλούμε όλες τις απαραίτητες βιβλιοθήκες και ειδικές συναρτήσεις ή κλάσεις μέσα σε αυτές

import pandas as pd

import numpy as np

import scipy as scp

import sklearn

import statsmodels.api as sm

import matplotlib.pyplot as pl

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import classification_report

from sklearn import metrics

from sklearn.metrics import confusion_matrix
```

ο διαχωρισμός σε X_train, y_train, X_test, y_test υλοποιείται στην αρχή με ρυθμό εκπαίδευσης 35 % μέσω της R (περιγράφεται στην προηγούμενη τεχνική)

διάβασμα των csv files με τα δείγματα εκπαίδευσης και ελέγχου των ανεξάρτητων μεταβλητών και της μεταβλητής στόχου

```
X_train = pd.read_csv('x_train.csv',  delimiter=',')

y_train = pd.read_csv('y_train.csv',  delimiter=',')

X_test = pd.read_csv('x_test.csv',  delimiter=',')

y_test = pd.read_csv('y_test.csv',  delimiter=',')

X_train = X_train.drop(['Unnamed: 0'], axis=1)

X_test = X_test.drop(['Unnamed: 0'], axis=1)

y_train = y_train.drop(['Unnamed: 0'], axis=1)

y_test = y_test.drop(['Unnamed: 0'], axis=1)
```

```
model1 = LogisticRegression(random_state=100, multi_class='auto', penalty='l2', solver='liblinear').fit(X_train, y_train) # εφαρμογή μοντέλου λογιστικής παλινδρόμησης με βήμα εκπαίδευσης του μοντέλου ανά 100 τιμές και προσαρμογή στα σύνολα X_train, y_train, οι τιμές των παραμέτρων είναι οι κατάλληλες για περιπτώσεις δυαδικής λογιστικής παλινδρόμησης

preds = model1.predict(X_test) ## οι προβλέψεις που προκύπτουν δίδοντας ως εισόδους τις τιμές του δείγματος X_test

params = model1.get_params() ## οι παράμετροι του μοντέλου λογιστικής παλινδρόμησης
```

τύπωση παραμέτρων του μοντέλου και πιο συγκεκριμένα της σταθεράς και των συντελεστών των ανεξάρτητων μεταβλητών

```
print(params)

print('Intercept: \n', model1.intercept_)

print('Coefficients: \n', model1.coef_)
```

confusion_matrix(y_test, preds) ## δημιουργία του πίνακα με στήλες τις προβλεπόμενες ταξινομήσεις κάθε κλάσης και γραμμές τις πραγματικές τιμές των κλάσεων αυτών

confmtrx = np.array(confusion_matrix(y_test, preds))

pd.DataFrame(confmtrx, index=['0','1'],

columns=['predicted_0', 'predicted_1'])

print('Accuracy Score:', metrics.accuracy_score(y_test, preds)) ## τύπωση ακρίβειας του μοντέλου δηλαδή το ποσοστό των σωστών ταξινομήσεων σε σχέση με τις συνολικές ταξινομήσεις

ΚΕΦΑΛΑΙΟ 4: Δέντρα Αποφάσεων

4.1 Εισαγωγή

Στη συγκεκριμένη ενότητα περιγράφονται δύο ευρέως γνωστές τεχνικές επεξεργασίας των στοιχείων, που αποτελούν εναλλακτικές λύσεις πολλών μεθόδων, όπως της διακριτικής ανάλυσης και της συσταδοποίησης των στοιχείων. Οι τεχνικές των δέντρων αποφάσεων αναπτύσσουν δένδρα μέσω διακλαδώσεων, στα οποία κάθε μη τελικός κόμβος προσδίδει ένα επίπεδο τμήσης με σκοπό την επίτευξη μιας βέλτιστης πρόβλεψης ή ταξινόμησης.

Μία δενδρική διχοτόμηση είναι άριστη στην περίπτωση που ο αριθμός των τμήσεων είναι τέτοιος που να καλύπτει την πρόβλεψη των δεδομένων στις σωστές κλάσεις μιας εξαρτημένης μεταβλητής. Στη συνέχεια περιγράφονται οι τεχνικές δενδρικής κατηγοριοποίησης ή παλινδρόμησης και τα στοχαστικά δέντρα.

4.2 Δέντρα Κατηγοριοποίησης και Παλινδρόμησης

Τα δέντρα κατηγοριοποίησης και παλινδρόμησης εμφανίζονται σε εφαρμογές πολλών πεδίων είτε σε ποσοτικές μεταβλητές απόκρισης είτε σε ποιοτικές μεταβλητές. Η μέθοδος των δέντρων κατηγοριοποίησης - παλινδρόμησης (CART) εφαρμόζεται ευρέως από μεγάλη μερίδα επιστημόνων χωρίς αυτό να έχει σχέση με την ειδικότητα που έχουν. Μάλιστα, δεν είναι λίγοι εκείνοι που προτιμούν τη χρήση των δέντρων αποφάσεων αντί για κλασσικές μεθόδους πολυδιάστατων παλινδρομήσεων. Πιο συγκεκριμένα αναφέρουμε τα ακόλουθα πλεονεκτήματα :

- Επιτυγχάνεται απλούστευση των αποτελεσμάτων, κυρίως λόγω του απεικονιστικού χαρακτήρα των δέντρων. Συχνά, η ερμηνεία των αποτελεσμάτων είναι απλή διότι ο διαχωρισμός των παρατηρήσεων βασίζεται σε μερικές λογικές διατυπώσεις της μορφής, αν ισχύει μια συνθήκη ή μια τιμή για κάποια μεταβλητή κινείται σε συγκεκριμένα όρια ή ανήκει σε κάποια κατηγορία τότε το δέντρο εξάγει την πιο πιθανή απόκριση. Η απόδοση και εμπέδωση των αποτελεσμάτων είναι ταχύτερη,

με δύο ή τρεις το πολύ, λογικές διατυπώσεις παρά με την παρουσία δυσνόητων και περίπλοκων μαθηματικών εξισώσεων.

- Η ανάπτυξη των λογικών διατυπώσεων, εκ των προτέρων είναι απαραμετρική μέθοδος, έτσι δεν απαιτούνται έλεγχοι των παραμετρικών προϋποθέσεων όπως συμβαίνει στις άλλες τεχνικές. Η μέθοδος των μονοδιάστατων διαχωρίσεων (Discriminant-based univariate split) θεωρείτο η πιο αποδεκτή και βασίζεται στον προσδιορισμό των βέλτιστων τελικών κόμβων που διαχωρίζονται σε ένα δέντρο υπό τη δράση των ανεξάρτητων μεταβλητών. Στη δενδρική ταξινόμηση, και σε κάθε τελικό κόμβο, εφαρμόζονται στατιστικοί έλεγχοι για τη σημαντικότητα της πρόβλεψης των μελών των κλάσεων με βάση τις τιμές κάθε προβλεπτικής μεταβλητής (συνήθως σε κατηγορικές). Αν εντοπίζονται μόνο ανεξάρτητες κατηγορικές μεταβλητές, χρησιμοποιείται ο έλεγχος χ^2 της ανεξαρτησίας των περιπτώσεων (της εξαρτημένης μεταβλητής) και των επιπέδων (κατηγοριών) κάθε ανεξάρτητης μεταβλητής. Με παρούσες μόνο ποσοτικές ανεξάρτητες, πραγματοποιείται ανάλυση διακύμανσης (ANOVA), η οποία ελέγχει τη σημαντικότητα διαχωρισμού των κλάσεων της εξαρτημένης σε σχέση με τις ανεξάρτητες μεταβλητές. Όποια μεταβλητή παράγει την μικρότερη πιθανότητα σφάλματος (p-value), συγκρινόμενη με την πιθανότητα αναφοράς (συνήθως 0,05), αυτή επιλέγεται να διατμήσει τον αντίστοιχο κόμβο.
- Στις ποσοτικές ανεξάρτητες, χρησιμοποιείται ένας αλγόριθμος ο οποίος δημιουργεί συνθετικά δύο υπερκλάσεις στον κόμβο και μέσω δευτεροβάθμιων εξισώσεων, υπολογίζει δύο πιθανές τιμές, επιλέγοντας τη μια από αυτές για τον διαχωρισμό του αντίστοιχου κόμβου.
- Στις κατηγορικές ανεξάρτητες, δημιουργούνται ψευδο-μεταβλητές τόσες, όσες και οι κατηγορίες αυτών και μετά την εφαρμογή μιας διαδικασίας στατιστικών αναλύσεων με μεγάλο υπολογιστικό κόστος, επιλέγεται εκείνη η κατηγορία μίας μεταβλητής που βελτιστοποιεί τον κανόνα διαχωρισμού σε πιθανούς κόμβους.

Η τεχνική των τμήσεων των δένδρων βασίζεται σε πιθανές αποφάσεις που βελτιστοποιούν ένα πρόβλημα. Όμως παρουσιάζει ένα μειονέκτημα, κι αυτό αντικατοπτρίζεται στο γεγονός ότι δεν έχει την ικανότητα να διακόψει τη διαδικασία των τμήσεων χωρίς επιπλέον πληροφόρηση. Επιβάλλεται ο κανόνας να τερματίζει στο σημείο εκείνο όπου διαπιστώνεται ότι οι επόμενες τμήσεις συνδράμουν ελάχιστα έως καθόλου στη βελτίωση της πρόβλεψης της ταξινόμησης. Για παράδειγμα, αν 25

τμήσεις βοηθούν στην ορθή ταξινόμηση του 85% των παρατηρήσεων και 26 τμήσεις οδηγούν στο 85,2% τότε είναι προφανής η αδυναμία αισθητής βελτίωσης. Όμως, η επιλογή των 25 τμήσεων δεν είναι η οριστική καθώς υπό άλλες συνθήκες και περιορισμούς αυτό ποικίλλει. Πιο συγκεκριμένα, ενυπάρχουν μέθοδοι που ελέγχουν την ποιότητα της πρόβλεψης μέχρι το σημείο που διακόπτονται οι τμήσεις και μπορούν να απαλείψουν οπισθοδρομικά ένα μέρος των κλάδων των δένδρων μέχρι να δημιουργηθεί το σωστό μέγεθος που οδηγεί στην άριστη πρόβλεψη της ταξινόμησης των παρατηρήσεων

Οι μέθοδοι που αναφέρθηκαν, λειτουργούν με τον κανόνα ότι ένα μόνο μέρος των στοιχείων θα υποβληθεί στην ανάλυση της δενδρικής παλινδρόμησης ενώ το υπόλοιπο θα χρησιμοποιηθεί ως είσοδος νέων παρατηρήσεων για την εκτίμηση της ποιότητας της πρόβλεψης. Οι μέθοδοι αυτές είναι:

1. Η διασταυρωτική επικύρωση CV (Cross Validation) κατά την οποία το δένδρο δημιουργείται με χρήση μιας ομάδας παρατηρήσεων γνωστής και ως δείγμα εκπαίδευσης (learning sample ή training sample). Η εγκυρότητα του βέλτιστου δέντρου διασταυρώνεται με ένα άλλο,ανεξάρτητο δείγμα παρατηρήσεων, γνωστό ως δείγμα ελέγχου (testing sample).

2. Η διασταυρωτική επικύρωση V φορές (V-fold cross validation) κατά την οποία το αρχικό δείγμα που εξετάζεται, χωρίζεται σε V το πλήθος μικρότερα υποδείγματα. Σε κάθε επανάληψη της διαδικασίας λαμβάνονται τυχαία παρατηρήσεις από τα υποδείγματα και διασταυρώνονται (cross validation) για τον έλεγχο της εγκυρότητας των διακλαδώσεων του δέντρου. Το άριστο δένδρο πρόβλεψης είναι εκείνο που προκύπτει από την άριστη μέση ακρίβεια των διασταυρούμενων παρατηρήσεων ταξινόμησης.

Η διαδικασία των υπολογισμών που απαιτούνται για την ανάπτυξη ενός δένδρου ταξινόμησης ή παλινδρόμησης, περιλαμβάνει 6 βήματα:

1. **Προσδιορισμός των κριτηρίων υπολογισμού της ακρίβειας πρόβλεψης**. Η ακριβής πρόβλεψη ορίζεται εκείνη που παρουσιάζει το ελάχιστο κόστος, δηλαδή το μικρότερο ποσοστό εσφαλμένης ταξινόμησης των παρατηρήσεων στη δενδρική ταξινόμηση. Για την ενίσχυση της αξιοπιστίας και εγκυρότητας της ακρίβειας της πρόβλεψης υπολογίζονται τα κόστη στην κάθε περίπτωση με σκοπό τη μείωση της διακύμανσης. Στη δενδρική ταξινόμηση όπου η εξαρτημένη μεταβλητή είναι

κατηγορική, η ελαχιστοποίηση του κόστους προσδιορίζεται όταν οι εκ των προτέρων πιθανότητες είναι ανάλογες του μεγέθους των κλάσεων (κατηγοριών) και παράλληλα όταν οι πιθανότητες των εσφαλμένων ταξινομήσεων είναι ίσες σε κάθε κλάση. Στη δενδρική παλινδρόμηση, δεν υφίστανται εκ των προτέρων πιθανότητες λόγω απουσίας κατηγοριών στην εξαρτημένη μεταβλητή.

2. **Επιλογή των μεταβλητών που θα καθορίζουν τον τρόπο διαχωρισμού των διακλαδώσεων ανά περίπτωση.** Ουσιαστικά, αφορά την ορθή επιλογή των διατμήσεων που υφίστανται οι ανεξάρτητες μεταβλητές με σκοπό να προβλέψουν τα μέλη των κλάσεων στην κατηγορική εξαρτημένη μεταβλητή (ταξινόμηση) ή να προβλέψουν τιμές της ποσοτικής εξαρτημένης μεταβλητής (παλινδρόμηση) με τέτοιο τρόπο ώστε να επιτυγχάνεται η μέγιστη ορθή πρόβλεψη. Η ακρίβεια της πρόβλεψης μετριέται σε κάθε τελικό κόμβο ως μέτρο της ομοιογένειας των παρατηρήσεων ή του μεγέθους της πολυπλοκότητας του :

$$G(t) = \sum_{i \neq j} p(i|t)p(j|t)$$

$$p(j|t) = \frac{p(j,t)}{p(t)}\,, \qquad p(j,t) = \frac{\pi_j\, N_j(t)}{N_j}$$

όπου η $p(j|t)$ είναι η πιθανότητα να βρεθεί η παρατήρηση στην κατηγορία j δεδομένου ότι εντοπίζεται στον κόμβο t

, η $p(j,t)$ είναι η πιθανότητα να βρεθεί η παρατήρηση στην κατηγορία j και στον κόμβο t

, το π_j είναι η προγενέστερη πιθανότητα για την κατηγορία j

, το $N_j(t)$ είναι το πλήθος των στοιχείων της κατηγορίας j στον κόμβο t

, το N_j είναι το συνολικό πλήθος των στοιχείων της κατηγορίας j.

Αν όλες οι παρατηρήσεις στους τελικούς κόμβους είναι διαφοροποιημένες, δηλαδή η κάθε παρατήρηση να ανήκει σε μία κλάση μόνο, τότε η ομοιογένεια παρουσιάζεται ως μέγιστη και η πρόβλεψη θεωρείται ως άριστη - ιδανική. Η νοθεία (impurity) ορίζεται ως το αντίστροφο της ομοιογένειας και οι βασικοί δείκτες που υπολογίζουν την νοθεία σε μια ταξινόμηση με δέντρα αποφάσεων είναι :

- Ο δείκτης του Gini (Gini measure of impurity) του κόμβου, ο οποίος προσεγγίζει το μηδέν όταν είναι παρούσα σε ένα κόμβο μόνο μία (καθαρή) κλάση. Όταν οι εκ των προτέρων πιθανότητες λαμβάνονται ως ανάλογες του μεγέθους των κλάσεων και το κόστος εσφαλμένης ταξινόμησης ίσο εντός των κλάσεων, ο δείκτης Gini προκύπτει ως το άθροισμα του γινομένου όλων των ζευγών των συνδυασμών των κλάσεων που είναι παρούσες στον συγκεκριμένο κόμβο. Ο δείκτης λαμβάνει τη μέγιστη τιμή όταν τα μεγέθη των κλάσεων στον κόμβο είναι ίσα και μηδέν, όταν όλες οι παρατηρήσεις στον κόμβο ανήκουν στην ίδια κλάση.

- Ο δείκτης χ^2, που μπορεί να υπολογιστεί με χρήση της απόκλισης των παρατηρούμενων συχνοτήτων από τις αναμενόμενες συχνότητες ταξινόμησης των παρατηρήσεων.

Στην παλινδρόμηση των δέντρων, ως δείκτης νοθείας εφαρμόζεται το κριτήριο της απόκλισης των ελαχίστων τετραγώνων R(t) όπως ακολούθως :

$$R(t) = \frac{1}{N_w(t)} \sum_{i \in t} w_i f_i \left(y_i - \overline{y(t)}\right)^2$$

όπου f_i είναι η εκτίμηση της τιμής της παρατήρησης i που ανήκει στον κόμβο t, y_i είναι η πραγματική τιμή της παρατήρησης i, $\overline{y(t)}$ είναι η μέση τιμή των παρατηρήσεων που ανήκουν στον κόμβο t, $N_w(t)$ είναι το πλήθος των παρατηρήσεων στον κόμβο t που έχουν βαρύτητα w.

3. **Προσδιορισμός του σημείου διακοπής των επαναληπτικών διαμερίσεων**. Δύο κριτήρια προσδιορισμού για την περάτωση της διαδικασίας είναι :

α) Επιλογή του ελάχιστου αριθμού παρατηρήσεων που απαιτούνται στον βέλτιστο τελικό κόμβο.

β) Επιλογή ενός συγκεκριμένου ποσοστού των πραγματικών δεδομένων που ανήκουν σε μία ή περισσότερες κλάσεις για τον κάθε κόμβο και ένα άνω όριο τιμών από όμοια κλάση. Η επιλογή των κριτηρίων ενίοτε καθορίζεται από το είδος της έρευνας και του τρόπου που επιθυμεί ο ερευνητής να τη συσχετίσει (εξαρτάται από τη φύση του προβλήματος).

4. **Επιλογή του δένδρου του ορθού μεγέθους που προκύπτει με τη σωστή οπισθοδρόμηση και την επιλεκτική αφαίρεση περιττών κλάδων.** Το μέγεθος του

δένδρου ταξινόμησης ή παλινδρόμησης έχει βασικό ρόλο στην αποτελεσματικότητα της ανάλυσης, καθώς ένα ογκώδες δέντρο, δυσκολεύει σε μεγάλο βαθμό την ερμηνεία των αποτελεσμάτων. Ο προσδιορισμός του κατάλληλου δενδρικού μεγέθους διακλαδώσεων ταυτίζεται με τη λογική να μην υπάρχει μεγάλη πολυπλοκότητα αλλά ούτε και ένα υπερβολικά απλό δέντρο. Ουσιαστικά αυτό που πρέπει να επιτευχθεί είναι το δέντρο να αναγνωρίζει αν μια πληροφορία είναι χρήσιμη, ώστε να αυξήσει την ακρίβεια της πρόβλεψης κι όχι η τελευταία να βελτιώνεται με οποιαδήποτε πληροφορία εισέρχεται.

Στις επόμενες παραγράφους αναπτύσσονται δύο στρατηγικές επιλογής της ακρίβειας του μεγέθους των διακλαδώσεων. Η πρώτη στρατηγική βασίζεται στον προσδιορισμό ενός προκαθορισμένου ποσοστού μεγέθους παρατηρήσεων – στόχου που λειτουργεί ως το ανώτατο όριο για τον τελικό κόμβο. Η άλλη διαδικασία βασίζεται στην απόφαση του μεγέθους των δέντρων με αυτοματοποιημένο τρόπο, δηλαδή τη βοήθεια ενός ειδικού αλγόριθμου για το κατάλληλο μέγεθος παρατηρήσεων. Στην πρώτη κατηγορία, επιλογής ιδανικού μεγέθους, εφαρμόζονται διάφορα κριτήρια με σκοπό τον υπολογισμό της αξιοπιστίας:

α) Το δείγμα ελέγχου. Κατασκευάζεται το δένδρο μέσα από το υπό εξέταση δείγμα και πραγματοποιείται έλεγχος σχετικά με την ακρίβειά του βασισμένο στην τιμή της πρόβλεψη. Σε περίπτωση που τα κόστη του δείγματος ελέγχου ξεπερνούν τα κόστη του δείγματος που διερευνάται, τότε υπάρχει το φαινόμενο χαμηλής διασταύρωσης. Στο τελευταίο ενδεχόμενο, ο αλγόριθμος κατευθύνεται στην επιλογή διαφορετικού μεγέθους κλάδων με σκοπό τη βελτιστοποίηση στο δείγμα ελέγχου.

β) Το δείγμα V υποσυνόλων. Αν το δείγμα ελέγχου δεν είναι καθορισμένο από το πρόβλημα και το δείγμα που μελετάται είναι μικρού μήκους, τότε από το σύνολο των παρατηρήσεων επιλέγεται με τυχαίο τρόπο μια διαμέριση ίσων υποδειγμάτων, δηλαδή η ένωση των υποδειγμάτων ισούται με το δείγμα ελέγχου ενώ μεταξύ τους τα υποδείγματα δεν έχουν κοινά στοιχεία. Με αυτό τον τρόπο σε κάθε επανάληψη δημιουργείται ένα δέντρο με καθορισμένο μέγεθος, όπου την κάθε φορά ένα υπόδειγμα καθορίζεται για έλεγχο της ακρίβειας του μοντέλου ενώ τα υπόλοιπα V-1 υποδείγματα εκπαιδεύουν το δέντρο. Τα κόστη αξιοπιστίας της διασταυρωτικής επικύρωσης είναι ο μέσος όρος των εκτιμήσεων από τα V το πλήθος βήματα της διεργασίας.

Στη δεύτερη στρατηγική, του προκαθορισμένου μεγέθους δέντρου, εφαρμόζεται ολική διασταυρωτική επικύρωση (Global Cross Validation, GCV). Στο

σημείο αυτό, το αρχικό δείγμα διασπάται σε πρότυπα-σύνολα διαμέρισης όπως και στην προηγούμενη στρατηγική σε V το πλήθος στάδια, παρακρατώντας σε κάθε βήμα ένα υπόδειγμα που λειτουργείς ως δείγμα ελέγχου. Η λειτουργία του αλγόριθμου στοχεύει στην ανάπτυξη ενός υπερμεγέθους δένδρου με πολλούς κόμβους. Πιο συγκεκριμένα, επιλέγονται αρκετά δένδρα ως βέλτιστα που να τηρούν τους περιορισμούς, από τα οποία διατηρείται εκείνο που ακολουθεί την εξής αρχή :

«το ιδανικό δένδρο είναι το δένδρο εκείνο που έχει σχετικά λίγα κλαδιά και επιπλέον τα κόστη του δεν υπερβαίνουν τα ελάχιστα κόστη +1 τυπικό σφάλμα εκείνου του δένδρου που φέρει λιγότερα κλαδιά».

Ουσιαστικά, αν επιλεχθεί τυπικό σφάλμα ανώτερο της μονάδος, ενδέχεται να οδηγήσει σε ασθενέστερα δένδρα ενώ αν επιλεχθεί τυπικό σφάλμα που ισούται με το 0, προκύπτει το δένδρο με τον ελάχιστο αριθμό κλαδιών, κάτι που δεν είναι αυτό που ζητείται.

5. **Τρόπος εκτίμησης των κριτηρίων της ακρίβειας πρόβλεψης.** Στη ταξινόμηση με μεθόδους δένδρων αποφάσεων, η ακρίβεια της πρόβλεψης μετρείται ως ποσοστό των ορθών ταξινομήσεων μεταξύ εκτιμηθέντων και πραγματικών. Βέβαια, όπως τονίστηκε και προηγουμένως, η ομοιογένεια του κόμβου εκτιμάται με το δείκτη νοθείας του Gini και το κριτήριο χ^2 . Στη αντίστοιχη παλινδρόμηση με δένδρα, η ακρίβεια υπολογίζεται σαν το μέσο σφάλμα των τετραγώνων της μεταβλητής που συνεισφέρει στην πρόβλεψη την κάθε φορά. Στην ίδια κατηγορία ο ρόλος του δείκτη νοθείας αντικατοπτρίζεται στην απόκλιση των ελαχίστων τετραγώνων.

Όσον αφορά την ακρίβεια πρόβλεψης του κόστους στη δενδρική ταξινόμηση, αξίζει να αναφερθεί ο παρακάτω τρόπος:

Εκτίμηση κόστους επαναντικατάστασης, το οποίο ταυτίζεται με το ποσοστό των παρατηρήσεων που ταξινομήθηκαν λανθασμένα υπό την επιρροή μιας μεταβλητής - ταξινομητή και απευθύνεται στο αρχικό δείγμα. Η επαναντικατάσταση ορίζεται στον ακόλουθο τύπο :

$$R(d) = \frac{1}{N}\sum_{i=1}^{N} X(d(x_n) \neq j_n)$$

όπου d είναι ο ταξινομητής , το x_n είναι η τιμή μιας παρατήρησης για κάποια ανεξάρτητη μεταβλητή, το j_n αποτελεί τις πραγματικές τιμές της εξαρτημένης

μεταβλητής ως προς κάποια ανεξάρτητη. Ουσιαστικά, το X αποτελεί έναν μετρητή που λαμβάνει την τιμή 1 όταν αληθεύει η έκφραση του ορίσματος $d(x_n) \neq j_n$ ενώ σε αντίθετη περίπτωση αποδίδει τιμή 0.

6. Επιλογή της μεταβλητής με τη μέγιστη σημαντικότητα πρόβλεψης ως η επικρατούσα μεταβλητή διαχωρισμού κλάδων. Η σημαντικότητα της πρόβλεψης των μεταβλητών υπολογίζεται αθροίζοντας ξεχωριστά κάποιες ποσότητες σε κάθε μεταβλητή. Σχετικά με την περίπτωση της ταξινόμησης μέσω δένδρων, προστίθενται όλες οι σωστές ταξινομήσεις για τους αντίστοιχους κόμβους ενώ αν διεξάγεται παλινδρόμηση εξετάζονται οι τιμές επαναντικατάστασης, όπως ορίστηκε παραπάνω. Ιδιαίτερα, στην τελευταία περίπτωση μετά την άθροιση ακολουθεί μετατροπή σε ποσοστά ως προς το μεγαλύτερο άθροισμα που προέκυψε από τις ανωτέρω τιμές. Έτσι, επιλέγεται εκείνη που μεγιστοποιεί το συγκεκριμένο ποσοστό.

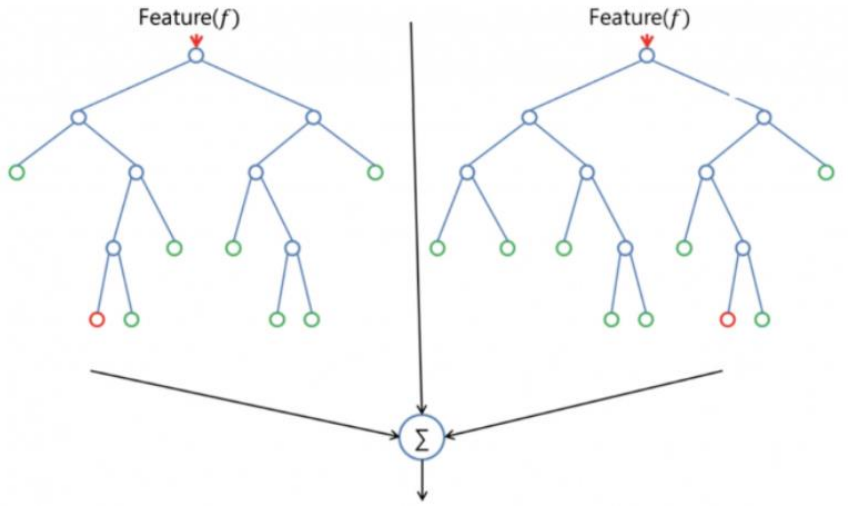
Βασικό πρόβλημα της μεθόδου είναι η επιλογή μιας μεταβλητής που σε άλλες τεχνικές δεν θα προοριζόταν ως βέλτιστη, δηλαδή ένα είδος σφάλματος εκχώρησης μη σημαντικής μεταβλητής. Όμως, αυτό διορθώνεται με βάση την, συνήθως, υψηλή προβλεπτική ικανότητα που διαθέτει.

4.3 Η τεχνική των Στοχαστικών Δέντρων

Το τυχαίο ή στοχαστικό δάσος (Random Forest) είναι ένας εποπτευόμενος αλγόριθμος μάθησης . Το «δάσος» που δημιουργείται, είναι ένα σύνολο δένδρων αποφάσεων, συνήθως εκπαιδευμένο με τη μέθοδο bagging. Η γενική ιδέα της τεχνικής αυτής αντικατοπτρίζεται στο γεγονός ότι ένα σύνολο μοντέλων εκμάθησης αυξάνει το συνολικό αποτέλεσμα μιας πρόβλεψης.

Με λίγα λόγια το τυχαίο δάσος δημιουργεί δέντρα πολλαπλών αποφάσεων και τα συγχωνεύει για να πάρει μια πιο ακριβή και σταθερή πρόβλεψη. Ένα μεγάλο πλεονέκτημα του τυχαίου δάσους είναι ότι μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για παλινδρόμηση, τα οποία αποτελούν την πλειονότητα των τρεχόντων συστημάτων μηχανικής εκμάθησης. Η περίπτωση του τυχαίου δάσους στην ταξινόμηση, θεωρείται μερικές φορές το θεμέλιο στοιχείο της μηχανικής μάθησης. Στην παρακάτω εικόνα παρατηρείται γραφικά η περίπτωση που περιλαμβάνει δύο δένδρα απόφασης στο τυχαίο δάσος. Είναι προφανές ότι μετά την

παλινδρόμηση ή κατηγοριοποίηση σε κάθε δένδρο ξεχωριστά λαμβάνονται οι τελικοί κόμβοι ανά περίπτωση. Έπειτα με τη χρήση κάποιων σταθμίσεων καθορίζεται η τελική λύση που εξαρτάται από τις μονοδιάστατες λύσεις κάθε δένδρου.



Σχήμα 4.1 *Τυχαίο δάσος με δύο δένδρα (Donges, 2019)*

Αξίζει να σημειωθεί, η σημαντική υπόσταση του τυχαίου δάσους στην τυχαιότητα ενός μοντέλου με την ταυτόχρονη αύξηση του μεγέθους των δένδρων. Αντί να αναζητά το πιο σημαντικό χαρακτηριστικό ενώ διαχωρίζει έναν κόμβο, αναζητά το καλύτερο χαρακτηριστικό σε ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτό οδηγεί σε μια μεγάλη ποικιλία που γενικά προσδιορίζει το καλύτερο μοντέλο. Επομένως, σε ένα τυχαίο δάσος, μόνο ένα τυχαίο υποσύνολο των χαρακτηριστικών λαμβάνεται υπόψη από τον αλγόριθμο για τη διάσπαση ενός κόμβου. Η τυχαιότητα των δένδρων γίνεται ολοένα και καλύτερη με τη χρήση τυχαίων κατώτατων ορίων για κάθε κόμβο αντί την διαδικασία εύρεσης των καλύτερων ελάχιστων ορίων, όπως πραγματοποιείται σε ένα απλό δέντρο αποφάσεων.

Παρ’ όλο που το τυχαίο δάσος είναι μια συλλογή δέντρων αποφάσεων, εντοπίζονται κάποιες διαφορές που θα φανούν στη συνέχεια. Έστω ένα σύνολο δεδομένων εκπαίδευσης με μεταβλητές και κλάσεις σε ένα δέντρο αποφάσεων. Με αυτό τον τρόπο διαμορφώνεται κάποιο σύνολο κανόνων, οι οποίοι θα χρησιμοποιηθούν για την πραγματοποίηση των προβλέψεων. Συγκριτικά, ο αλγόριθμος τυχαίων δασών επιλέγει τυχαία τις παρατηρήσεις και τα χαρακτηριστικά με σκοπό τη δημιουργία

πολλών δένδρων αποφάσεων και στη συνέχεια υπολογίζει τα αποτελέσματα και αποφασίζει. Μια άλλη κύρια διαφορά παρατηρείται στο ότι τα ογκώδη δέντρα αποφάσεων μπορεί να παρουσιάζουν το γνωστό πρόβλημα της υπερβολικής εκπαίδευσης ή υπερ-εκπαίδευσης, όπου το δείγμα εκπαίδευσης υπερβαίνει κατά πολύ το δείγμα ελέγχου με αποτέλεσμα να προκύπτει λανθασμένη υψηλή ακρίβεια στο μοντέλο. Τις περισσότερες φορές, το τυχαίο δάσος το αποτρέπει αυτό δημιουργώντας τυχαία υποσύνολα των χαρακτηριστικών και κατασκευάζοντας μικρότερα δέντρα χρησιμοποιώντας αυτά τα υποσύνολα. Στη συνέχεια, συνδυάζει τα υποσύνολα των δέντρων. Βέβαια, αυτή η τεχνική δεν λειτουργεί πάντα και καθιστά τον υπολογισμό πιο αργό, ανάλογα με το πλήθος των δέντρων που αναπτύσσει το τυχαίο δάσος.

Ένα από τα μεγαλύτερα προβλήματα στη μηχανική μάθηση είναι η υπερβολική τροφοδότηση. Αυτό το γεγονός αντιμετωπίζεται με τον ταξινομητή δασών. Ο κύριος περιορισμός του τυχαίου δάσους είναι ότι ένας μεγάλος αριθμός δέντρων μπορεί να κάνει τον αλγόριθμο πολύ αργό και αναποτελεσματικό για προβλέψεις σε πραγματικό χρόνο. Σε γενικές γραμμές, αυτοί οι αλγόριθμοι είναι ταχείς στην εκπαίδευση, αλλά σχετικά χαμηλού ρυθμού στην δημιουργία προβλεπτικών εκτιμήσεων. Μια πιο ακριβής πρόβλεψη απαιτεί περισσότερα δέντρα, με αποτέλεσμα, ακόμη πιο αργό μοντέλο. Στις περισσότερες εφαρμογές, ο τυχαίος αλγόριθμος δασών είναι αρκετά γρήγορος, αλλά σίγουρα μπορεί να υπάρχουν καταστάσεις, όπου ο ρυθμός του χρόνου εκτέλεσης είναι σημαντικός και θα υποτιμούνταν με άλλες προσεγγίσεις.

Πρέπει να επισημανθεί, πως το τυχαίο δάσος είναι ένα εργαλείο πρόβλεψης μοντελοποίησης και όχι ένα περιγραφικό εργαλείο. Αυτό βασίζεται στην ερμηνεία ότι αν αναζητείται μια περιγραφή των σχέσεων στα δεδομένα , άλλες τεχνικές θα ήταν καλύτερες. Η σημαντικότητα των τυχαίων δασών είναι εμφανής, αν συνυπολογίσει κανείς τη ευρεία χρήση τους σε πολλές επιστήμες. Ο αλγόριθμος του στοχαστικού δάσους χρησιμοποιείται σε πολλά διαφορετικά πεδία, όπως τραπεζικές συναλλαγές, χρηματιστήριο, φάρμακα και ηλεκτρονικό εμπόριο.

Στα οικονομικά, για παράδειγμα, χρησιμοποιείται για τον εντοπισμό πελατών που είναι πιο πιθανό να εξοφλήσουν εγκαίρως το χρέος τους, ή να χρησιμοποιούν τις υπηρεσίες μιας τράπεζας πιο συχνά. Σε αυτόν τον τομέα χρησιμοποιείται επίσης για τον εντοπισμό διαφθοράς ή εξαπάτησης από την τράπεζα. Στις συναλλαγές, ο αλγόριθμος μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της μελλοντικής συμπεριφοράς μιας μετοχής. Επιπροσθέτως, το τυχαίο δάσος χρησιμοποιείται στο

ηλεκτρονικό εμπόριο για να προσδιορίσει εάν σε έναν πελάτη θα αρέσει πραγματικά ένα συγκεκριμένο προϊόν ή όχι.

ΚΕΦΑΛΑΙΟ 5: Εφαρμογή

5.1 Εισαγωγή

Αυτό το κεφάλαιο περιγράφει αναλυτικά το πρόβλημα που θα χρησιμοποιηθεί στην συγκεκριμένη εφαρμογή της διπλωματικής εργασίας. Έπειτα περιγράφονται τα δεδομένα που έχουν ληφθεί, ο στόχος του πειράματος και οι τεχνικές ταξινόμησης που θα χρησιμοποιηθούν. Στο τελευταίο μέρος πραγματοποιείται μια σύγκριση των ευρημάτων της εφαρμογής με σκοπό την μετέπειτα ανάλυση και ερμηνεία των ακριβειών της κάθε μεθόδου σε σχέση με τη βιβλιογραφία.

5.2 Περιγραφή προβλήματος

Στο σημείο αυτό περιγράφεται το πρόβλημα πάνω στο οποίο θα πραγματοποιηθεί η ανάλυση μας. Το αρχικό σύνολο δεδομένων (dataset) είχε δημιουργηθεί ώστε να χρησιμοποιηθεί σε ένα διαγωνισμό data mining με σκοπό την πρόβλεψη πελατών που θα αγόραζαν ασφαλιστικό συμβόλαιο για τροχόσπιτο («Insurance Company Benchmark (COIL 2000) Data Set», 2000). Μετέπειτα, έγιναν με τα ίδια δεδομένα διάφορες εναλλακτικές μελέτες (Van der Putten et al., 2000).

Στην παρούσα διπλωματική εργασία, το σύνολο δεδομένων χωρίζεται αρχικά σε δύο μέρη, ένα σύνολο εκπαίδευσης (training set), ένα σύνολο ελέγχου (test set) και αντί της εξαρτημένης μεταβλητής πρόβλεψης που ήταν στον διαγωνισμό ο αριθμός των ασφαλιστικών συμβολαίων για τροχόσπιτο (CARAVAN Number of mobile home policies με τιμές 0 - 1), θέτουμε μια νέα ανεξάρτητη μεταβλητή που σχετίζεται με τον αριθμό των συμβολαίων των αυτοκινήτων. Έπειτα, ορίζεται μια νέα μεταβλητή, έστω y, τέτοια ώστε να λαμβάνει την τιμή 1, όταν ο αριθμός συμβολαίων ασφάλειας αυτοκινήτων υπερβαίνει τον μέσο όρο τους, και την τιμή 0, στην περίπτωση που είναι μικρότερος (δίτιμη μεταβλητή):

$$y = \begin{cases} 1, \alpha\nu\ X \geq \mu \\ 0, \alpha\nu\ X < \mu, \end{cases}$$

όπου X είναι ο αριθμός των συμβολαίων αυτοκινήτων και μ, ο μέσος τους.

Ουσιαστικά ο σκοπός είναι να προβλέψουμε την πιθανότητα ο αριθμός των συμβολαίων ασφάλειας να υπερβαίνουν τον μέσο όρο όσο το δυνατόν με μεγαλύτερη

ακρίβεια. Γι’ αυτό το λόγο προτείνεται η χρήση των μεθόδων που αναπτύχθηκαν στις Ενότητες 3, 4 και μια σύγκριση τους ως προς το αποτέλεσμα.

5.3 Περιγραφή δεδομένων

Τα κύρια χαρακτηριστικά των συνόλων δεδομένων περιγράφονται ακολούθως. Το αρχικό σύνολο δεδομένων (dataset) είναι χωρισμένο σε δύο μέρη. Ένα σύνολο εκπαίδευσης (training set) από 5822 εγγραφές πελατών και ένα σύνολο ελέγχου (test set) 4000 εγγραφών πελατών. Ακόμη, το dataset αποτελείται από 86 μεταβλητές που περιγράφουν κάποια δημογραφικά στοιχεία των πελατών (όπως εισόδημα, κατοχή αυτοκινήτου, ιδιοκτήτης σπιτιού) αλλά και τι είδος ασφαλιστικού συμβολαίου κατέχουν (αριθμός συμβολαίων σπιτιού, αυτοκινήτου, μηχανής, contribution στο συμβόλαιο).

Σε πρώτη φάση δημιουργείται ένα νέο σύνολο δεδομένων ώστε να εφαρμοστεί η ανάλυσή μας. Στο αρχικό σύνολο εκπαίδευσης αφαιρείται η μεταβλητή που αφορά την κατοχή συμβολαίων τροχόσπιτων αλλά και η παλιά μεταβλητή που αφορούσε τον αριθμό συμβολαίων αυτοκινήτων για τον κάθε πελάτη. Με αυτό τον τρόπο και τα δύο αρχικά σύνολα έχουν 84 μεταβλητές.

Στη συνέχεια, ενώνονται τα δύο σύνολα δεδομένων με σκοπό τη δημιουργία ενός νέου συνόλου δεδομένων, με το όνομα data. Για την κάθε μέθοδο που θα χρησιμοποιηθεί το ολικό σύνολο δεδομένων θα διαχωρίζεται σε δύο νέα υποσύνολα, ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Αντίστοιχα, στα δύο αυτά σύνολα, υπολογίζεται με το ίδιο τρόπο διαχωρισμού, η εξαρτημένη μεταβλητή. Μια μεταβλητή με την ιδιότητα της εκμάθησης του μοντέλου και μια ανάλογη μεταβλητή για το σκοπό του ελέγχου της εξαρτημένης μεταβλητής ώστε να υπολογιστεί η ακρίβεια του μοντέλου. Βέβαια, σε μεθόδους μη εποπτευόμενης μάθησης αυτός ο διαχωρισμός δεν είναι απαραίτητος, αφού δεν υφίσταται εκπαίδευση του μοντέλου.

Σχετικά με τις μεταβλητές που συμμετέχουν στο αρχικό σύνολο δεδομένων, περιγράφονται παρακάτω με σαφήνεια ως προς τις τιμές του και τον τύπο τους (κατηγορικές ή ποσοτικές) :

1 **MOSTYPE**: τύπος πελάτη, δηλαδή :

1 : Υψηλό εισόδημα

2 : Κάτοικος επαρχίας

3 : Ηλικιωμένοι υψηλού βιοτικού επιπέδου

4 : Εύποροι σε ακριβά διαμερίσματα

5 : Ηλικιωμένοι όχι κάποιας ειδικής κατηγορίας

6 : Γονείς με καριέρα και παιδική μέριμνα

7 : Διπλό εισόδημα χωρίς παιδιά

8 : Οικογένειες μεσαίας τάξης

9 : Σύγχρονες οικογένειες

10 : Σταθερή οικογένεια

11 : Νέοι γονείς

12 : Εύπορες νέες οικογένειες

13 : Νέα αμερικάνικη οικογένεια

14 : Πελάτες με προτίμηση σε μικρά κοσμοπολίτικα σπίτια

15 : Πελάτες με προτίμηση σε ακριβά κοσμοπολίτικα σπίτια

16 : Μαθητές που κατοικούν σε διαμερίσματα

17 : Νέοι δάσκαλοι στην πόλη

18 : Ανύπαντροι νέοι

19 : Νέοι που κατοικούν στα προάστια

20 : Αλλοεθνείς

21 : Μη προνομιούχοι νέοι κάτοικοι αστικής ζώνης

22 : Κάτοικοι σε κτίρια πολλαπλών χρήσεων

23 : Νέοι και ανερχόμενοι

24 : Νέοι, χαμηλού μορφωτικού επιπέδου

25 : Ηλικιωμένοι που μετανάστευσαν στην αστική ζώνη

26 : Ηλικιωμένοι με ιδιόκτητο σπίτι

27 : Ηλικιωμένοι που κατοικούν σε διαμέρισμα

28 : Ηλικιωμένοι που διαμένουν σε γηροκομείο

29 : Ηλικιωμένοι με προτίμηση σε σπίτι με βεράντες, χωρίς μπροστινή αυλή

30 : Ηλικιωμένοι μοναχοί

31 : Καθολικοί χαμηλού εισοδήματος

32 : Ηλικιωμένοι που υπάγονται σε πάνω από δύο κατηγορίες ηλικιωμένων

33 : Μεγάλες οικογένειες χαμηλότερης κοινωνικής τάξης

34 : Μεγάλη οικογένεια, με παιδί που εργάζεται

35 : Οικογένειες στην επαρχία

36 : Ζευγάρια εφήβων με παιδιά

37 : Κάτοικοι μικρής πόλης

38 : Παραδοσιακές οικογένειες

39 : Μεγάλες θρησκευτικές οικογένειες

40 : Μεγάλες οικογενειακές αγροτικές οικογένειες

41 : Αγροτικές οικογένειες

2 **MAANTHUI** : Αριθμός οικείων

3 **MGEMOMV** : Μέσο μέγεθος νοικοκυριού

4 **MGEMLEEF** : Μέση ηλικία:

1 : 20-30 χρονών

2 : 30-40 χρονών

3 : 40-50 χρονών

4 : 50-60 χρονών

5 : 60-70 χρονών

6 : 70-80 χρονών

5 **MOSHOOFD** : Βασικός τύπος πελάτη:

1 : Επιτυχημένοι επιχειρηματίες

2 : Καλλιεργητές

3 : Μέση οικογένεια

4 : Πελάτες που επενδύουν στην καριέρα

5 : Καλή ποιότητα ζωής

6 : Ανώτεροι πλοίαρχοι

7 : Συνταξιούχος, πρώην απασχολούμενος επαγγελματικά με τη θρησκεία

8 : Οικογένεια με μέλη μεγάλων ηλικιών

9 : Συντηρητικές οικογένειες

10 : Αγρότες

6 **MGODRK** : Ρωμαιοκαθολικοί

(το ποσοστό ρωμαιοκαθολικών στην περιοχή κατοικίας του πελάτη):

0 : 0%

1 : 1 - 10%

2 : 11 - 23%

3 : 24 - 36%

4 : 37 - 49%

5 : 50 - 62%

6 : 63 - 75%

7 : 76 - 88%

8 : 89 - 99%

9 : 100%

7 **MGODPR :** Προτεστάντες

8 **MGODOV :** Άλλη θρησκεία

9 **MGODGE :** Χωρίς θρησκεία

10 **MRELGE :** Παντρεμένος

11 **MRELSA** : Συζούν μαζί

12 **MRELOV :** Άλλη σχέση

13 **MFALLEEN :** Ελεύθεροι

14 **MFGEKIND :** Οικία χωρίς παιδιά

15 **MFWEKIND :** Οικία με παιδιά

16 **MOPLHOOG :** Εκπαίδευση υψηλού επιπέδου

17 **MOPLMIDD :** Εκπαίδευση μεσαίου επιπέδου

18 **MOPLLAAG :** Εκπαίδευση κατώτερου επιπέδου

19 **MBERHOOG :** Υψηλό βιοτικό επίπεδο

20 **MBERZELF :** Επιχειρηματίας

21 **MBERBOER :** Αγρότης

22 **MBERMIDD :** Μάνατζερ

23 **MBERARBG :** Ειδικευμένοι εργάτες

24 **MBERARBO :** Ανειδίκευτοι εργάτες

25 **MSKA :** Κοινωνική τάξη Α

26 **MSKB1 :** Κοινωνική τάξη Β1

27 **MSKB2 :** Κοινωνική τάξη Β2

28 **MSKC :** Κοινωνική τάξη Γ

29 **MSKD :** Κοινωνική τάξη Δ

30 **MHHUUR :** Ενοικίαση σπιτιού

31 **MHKOOP :** Ιδιοκτήτες σπιτιού

32 **MAUT1 :** Κάτοχος ενός αυτοκινήτου

33 **MAUT2 :** Κάτοχος δύο αυτοκινήτων

34 **MAUT0 :** Χωρίς αυτοκίνητο

35 **MZFONDS :** Εθνική Υπηρεσία Υγείας

36 **MZPART :** Ιδιωτική ασφάλιση υγείας

37 **MINKM30 :** Έσοδα <30.000 $

38 **MINK3045 :** Έσοδα 30-45.000 $

39 **MINK4575 :** Έσοδα 45-75.000 $

40 **MINK7512 :** Έσοδα 75-122.000 $

41 **MINK123M :** Εισόδημα> 123.000 $

42 **MINKGEM :** Μέσο εισόδημα

43 **MKOOPKLA :** Κατηγορία αγοραστικής δύναμης

44 **PWAPART :** Συνασφάλιση ασφαλιστηρίου αστικής ευθύνης προς τρίτους :

0 : 0

1 : 1 -49

2 : 50 - 99

3 : 100 - 199

4 : 200 - 499

5 : 500 - 999

6 : 1000 - 4999

7 : 5000 - 9999

8 : 10.000 - 19.999

9 : ≥ 20.000

45 **PWABEDR** : Συνασφάλιση ασφαλιστηρίου αστικής ευθύνης (εταιρείες)

46 **PWALAND :** Συνασφάλιση ασφαλιστηρίου αστικής ευθύνης (γεωργία)

47 **PPERSAUT :** Συνασφάλιση ασφαλιστηρίου αυτοκινήτου

48 **PBESAUT :** Συνασφάλιση ασφαλιστηρίου μεταφορικών φορτηγών

49 **PMOTSCO :** Συνασφάλιση ασφαλιστηρίου μοτοσικλέτας

50 **PVRAAUT :** Συνασφάλιση ασφαλιστηρίου φορτηγών

51 **PAANHANG** Συνασφάλιση ασφαλιστηρίου τρέιλερ

52 **PTRACTOR :** Συνασφάλιση ασφαλιστηρίου τρακτέρ

53 **PWERKT :** Πολιτικές γεωργικών μηχανημάτων συνεισφοράς

54 **PBROM :** Συνασφάλιση ασφαλιστηρίου μοτοποδήλατου

55 **PLEVEN :** Συνασφάλιση ασφαλιστηρίου ζωής

56 **PPERSONG :** Συνασφάλιση ασφαλιστηρίου ατυχημάτων

57 **PGEZONG :** Συνασφάλιση ασφαλιστηρίου οικογενειακών ατυχημάτων

58 **PWAOREG :** Συνασφάλιση ασφαλιστηρίου αναπηρίας

59 **PBRAND :** Συνασφάλιση ασφαλιστηρίου πυρκαγιάς

60 **PZEILPL :** Συνασφάλιση ασφαλιστηρίου ιστιοσανίδας

61 **PPLEZIER :** Συνασφάλιση ασφαλιστηρίου σκαφών

62 **PFIETS :** Συνασφάλιση ασφαλιστηρίου ποδηλάτων

63 **PINBOED :** Συνασφάλιση ασφαλιστηρίου περιουσίας

64 **PBYSTAND :** Συνασφάλιση κοινωνικής ασφάλισης

65 **AWAPART :** Αριθμός ασφαλιστηρίων αστικής ευθύνης προς τρίτους

66 **AWABEDR :** Αριθμός ασφαλιστηρίων αστικής ευθύνης (εταιρείες)

67 **AWALAND :** Αριθμός ασφαλιστηρίων αστικής ευθύνης (γεωργία)

68 **APERSAUT :** Αριθμός ασφαλιστηρίων αυτοκινήτων

69 **ABESAUT :** Αριθμός ασφαλιστηρίων μεταφορικών φορτηγών

70 **AMOTSCO :** Αριθμός ασφαλιστηρίων μοτοσικλετών

71 **AVRAAUT :** Αριθμός ασφαλιστηρίων φορτηγών

72 **AAANHANG :** Αριθμός ασφαλιστηρίων τρέιλερ

73 **ATRACTOR :** Αριθμός ασφαλιστηρίων τρακτέρ

74 **AWERKT :** Αριθμός ασφαλιστηρίων γεωργικών μηχανημάτων

75 **ABROM :** Αριθμός ασφαλιστηρίων μοτοποδηλάτων

76 **ALEVEN :** Αριθμός ασφαλιστηρίων ασφαλίσεων ζωής

77 **APERSONG :** Αριθμός ασφαλιστηρίων προσωπικού ατυχήματος

78 **AGEZONG :** Αριθμός ασφαλιστηρίων για οικογενειακά ατυχήματα

79 **AWAOREG :** Αριθμός ασφαλιστηρίων αναπηρίας

80 **ABRAND :** Αριθμός ασφαλιστηρίων πυρκαγιάς

81 **AZEILPL :** Αριθμός ασφαλιστηρίων ιστιοσανίδας

82 **APLEZIER :** Αριθμός ασφαλιστηρίων σκαφών

83 **AFIETS :** Αριθμός ασφαλιστηρίων ποδηλάτων

84 **AINBOED :** Αριθμός ασφαλιστηρίων περιουσίας

85 **ABYSTAND :** Αριθμός ασφαλιστηρίων κοινωνικής ασφάλισης

86 **CARAVAN :** Αριθμός ασφαλιστηρίων για τροχόσπιτα

Στις περισσότερες μεταβλητές που χαρακτηρίζονται από κάποια ιδιότητα είναι δίτιμες παίρνοντας την τιμή 1, αν ικανοποιούν το συγκεκριμένο χαρακτηριστικό και την τιμή 0 σε διαφορετική από την πρώτη περίπτωση.

5.4 Ανάλυση

Στην ανάλυση μας χρησιμοποιούνται τόσο μέθοδοι με εκπαίδευση όσο και μη επιβλεπόμενοι με σκοπό την πρόβλεψη της συχνότητας των συμβολαίων ασφάλειας αυτοκινήτων που υπερβαίνουν τον μέσο όρο. Ακόμη, πραγματοποιείται και μια ανάλυση σχετικά με την ακρίβεια του μοντέλου προσδιορίζοντας τα πλεονεκτήματά του σε σχέση με τα υπόλοιπα.

Προτού, εφαρμόσουμε τις τεχνικές μεθόδους, προβαίνουμε σε τεχνικές προσαρμογής των δεδομένων.

Πρόγραμμα 4

data <- read.table(file = "data_with.txt", header=FALSE) # διάβασμα των δεδομένων του αρχικού συνόλου επαίδευσης

data = data[,1:85] # αφαίρεση της μεταβλητής αριθμού συμβολαίων τροχόσπιτων από το προηγούμενο σύνολο

data1 <- read.table(file = "data_without.txt", header=FALSE) # διάβασμα των δεδομένων του αρχικού συνόλου ελέγχου

data <- rbind(data,data1) # ένωση των δύο συνόλων (αρχικού εκπαίδευσης και αρχικού ελέγχου)

data <- data.frame(data) # προσαρμογή του νέου συνόλου δεδομένων σε dataframe

target <- data$V68 # καθορισμός μεταβλητής στόχου

m <- mean(target) # εύρεση του μέσου της εξαρτημένης μεταβλητής

```
for (i in 1:nrow(data)){
  if (target[i] >= m){
    target[i]=1
  }
```

} # επαναληπτική διαδικασία στην οποία κάθε στοιχείο της μεταβλητής ελέγχου παίρνει την τιμή 1, αν υπερβαίνει τον μέσο όρο και την τιμή 0, σε άλλη περίπτωση

data2 <- data[,1:67] # υποσύνολο νέου συνόλου δεδομένων από τις μεταβλητές 1 έως 67

data3 <- data[, 69:85] # υποσύνολο νέου συνόλου δεδομένων από τις μεταβλητές 69 έως 85

data <- cbind(data2, data3, target) # νέο σύνολο δεδομένων με τα προηγούμενα δεδομένα και τη μεταβλητή στόχου στο τέλος

df <- data # αλλαγή ονόματος για τα δεδομένα που θα χρησιμοποιηθούν στην εφαρμογή

df<-data.frame(df) # εφαρμογή dataframe

δημιουργία συνάρτησης κανονικοποίησης των υπό εξέταση δεδομένων

```
nor <- function(x) {
  (x -min(x))/(max(x)-min(x))
```

```
}
```

Εφαρμόζεται η συνάρτηση κανονικοποίησης στις στήλες m που ενδέχεται να αποτελέσουν τα χαρακτηριστικά – μεταβλητές που μας απασχολούν

```
df_norm <- as.data.frame(sapply(df[,c(1:84)], nor))
```

Η πρώτη μέθοδος που εκτελείται αφορά την τεχνική των k-πλησιέστερων γειτόνων. Λάβαμε 5 πιθανές τιμές για τα k, δηλαδή :

- k =2
- k = 4
- k = 5
- k = 7
- k =9

Ακολουθεί το Output από την R για τις πέντε περιπτώσεις :

k =2

```
  df_test_category
pr      0      1
 0   2649    485
 1    521   2730
```

Φαίνεται ότι η προβλεπτική ικανότητα είναι αρκετά καλή καθώς μόλις 521 τοποθετήθηκαν στην κατηγορία 1 αντί για την ορθή 0 ενώ 485 τοποθετήθηκαν στην 0 αντί για την 1. Συνολική, ακρίβεια του μοντέλου **84.24432 %.**

k = 4

pr	df_test_category 0	1
0	2818	276
1	359	2932

Παρατηρείται ότι η προβλεπτική ικανότητα βελτιώθηκε καθώς 359 τοποθετήθηκαν στην κατηγορία 1 αντί για την ορθή 0 ενώ 276 τοποθετήθηκαν στην 0 αντί για την 1. Συνολική, ακρίβεια του μοντέλου **90.05482 %.**

k = 5

pr	df_test_category 0	1
0	2905	190
1	225	3065

Επιπλέον βελτίωση διακρίνεται στην προβλεπτική ικανότητα αφού μόλις 225 τοποθετήθηκαν στην κατηγορία 1 αντί για την ορθή 0 ενώ 190 τοποθετήθηκαν στην 0 αντί για την 1. Συνολική, ακρίβεια του μοντέλου **93.50039 %.**

k = 7

pr	df_test_category 0	1
0	2947	117
1	186	3135

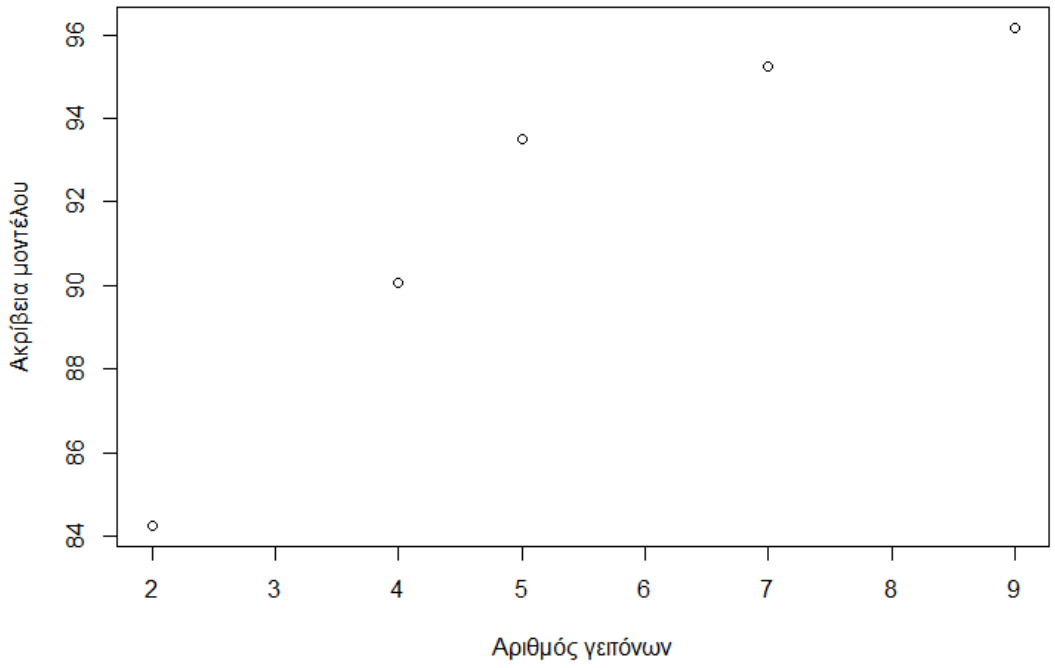
Η προβλεπτική ικανότητα παρουσιάζεται επιπρόσθετα ακόμη καλύτερη σε σχέση με την προηγούμενη, καθώς μόλις 186 τοποθετήθηκαν στην κατηγορία 1 αντί για την ορθή 0 ενώ 117 τοποθετήθηκαν στην 0 αντί για την 1. Συνολική, ακρίβεια του μοντέλου **95.2545 %.**

k =9

Στη δοκιμή για k=9 παρατηρείται περαιτέρω άνοδος της ακρίβειας.

	df_test_category	
pr	0	1
0	2976	88
1	156	3165

Εδώ, η προβλεπτική ικανότητα είναι εξαιρετική καθώς μόλις 156 τοποθετήθηκαν στην κατηγορία 1 αντί για την ορθή 0 ενώ 88 τοποθετήθηκαν στην 0 αντί για την 1. Συνολική, ακρίβεια του μοντέλου **96.17854 %**. Άξιο παρατήρησης είναι ότι ναι μεν παρουσιάζεται βελτίωση αλλά πολύ μικρότερη σε σχέση με προηγούμενα επίπεδα. Ωστόσο, τα επίπεδα βελτίωσης αρχίζουν να φθίνουν και η ακρίβεια να συγκλίνει σε μια τιμή.



Σχήμα 5.1 *Ακρίβεια του μοντέλου σε διαφορετικές τιμές γειτόνων για τη μέθοδο KNN*

Η δεύτερη μέθοδος που χρησιμοποιήθηκε αφορά τη μέθοδο της λογιστικής παλινδρόμησης. Στη λογιστική παλινδρόμηση η εξαρτημένη μεταβλητή μας θα παραμείνει η ίδια, δηλαδή:

$$y = \begin{cases} 1, \alpha\nu\ X \geq \mu \\ 0, \alpha\nu\ X < \mu, \end{cases}$$

όπου X είναι ο αριθμός των συμβολαίων αυτοκινήτων και μ, ο μέσος τους.

Στο διάνυσμα των επεξηγηματικών μεταβλητών χρησιμοποιούμε όλες τις μεταβλητές με ρυθμό εκπαίδευσης 35%.

	predicted_0	predicted_1
0	3120	0
1	0	3265

Παρατηρείται ότι η προβλεπτική ικανότητα είναι άριστη μιας και όλες οι παρατηρήσεις τοποθετήθηκαν στη σωστή κατηγορία με τη συνολική ακρίβεια του μοντέλου να είναι 100 %. Το αποτέλεσμα σίγουρα προβληματίζει ωστόσο έγιναν πειράματα και με μικρότερα δείγματα εκπαίδευσης χωρίς να υπάρχει κάποια διαφορά. Ακόμη, αξίζει να τονιστεί ότι δεν υπάρχει σημαντική βαρύτητα στις δύο κλάσεις καθώς η κλάση «1» υπάρχει στο 56% των συνολικών στοιχείων ενώ τα υπόλοιπα υπάγονται στην κλάση «0» κάτι που σημαίνει πως δεν υπάρχει πιθανότητα «ψευδούς» ακρίβειας στο μοντέλο που εφαρμόστηκε.

Η τρίτη μέθοδος που εφαρμόζεται, σχετίζεται με την τεχνική των στοχαστικών δέντρων. Στο μοντέλο προσαρμόζεται όλο το σύνολο δεδομένων, όπως και στην περίπτωση των k- πλησιέστερων γειτόνων.

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 27

Mean of squared residuals: 0.0001171136

% Var explained: 99.95

Σύμφωνα με τα ανωτέρω εξαγόμενα παρατηρείται ότι η μέθοδος προσαρμόζεται καλά με τη δημιουργία 500 δέντρων και τη συνεισφορά 27 μεταβλητών στο κάθε σημείο τομής με ακρίβεια 96.9 %. Ακόμη, η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος λογίζεται της τάξεως του 10^{-4} , δηλαδή $\sqrt{MSE} \approx C * 10^{-4}$, όπου C κάποιος ακέραιος μικρός αριθμός (μικρότερος από 10)

5.4 Συμπεράσματα

Στην παρούσα εργασία αναλύθηκε τόσο θεωρητικά όσο και πρακτικά το θέμα γύρω από τις τεχνικές της ανάλυσης μεγάλων δεδομένων στον αναλογισμό. Στα πρώτα κεφάλαια δόθηκαν όλοι οι απαραίτητοι ορισμοί σχετικά με τα Μεγάλα Δεδομένα και τις τεχνικές που έχουν χρησιμοποιηθεί δίδοντας μεγαλύτερη βαρύτητα στην αναλογιστική επιστήμη. Στα κεφάλαια 3,4 παρουσιάστηκαν τρεις τεχνικές που χρησιμοποιούνται ευρέως με σκοπό την ταξινόμηση, την τεχνική ανάλυση μεγάλων δεδομένων και τη διαχείριση κινδύνου.

Στην πτυχή που αφορούσε το πειραματικό μέρος, αντλήσαμε πληροφορίες και δεδομένα από έναν παλιό γνωστό διαγωνισμό («Insurance Company Benchmark (COIL 2000) Data Set», 2000). Το σύνολο δεδομένων που χρησιμοποιήθηκε, μπορεί το 2000 να θεωρούταν όντως ένα σύνολο μεγάλων δεδομένων. Πλέον, εν έτει 2021, δύσκολα θεωρείται σύνολο μεγάλων δεδομένων, λόγω του μικρού πλήθους των χαρακτηριστικών του (για την εποχή) και ο λόγος της χρήσης του γίνεται λόγω αδυναμίας εύρεσης πιο αντιπροσωπευτικού συνόλου μεγάλων δεδομένων στο τομέα του αναλογισμού. Με χρήση γνωστών τεχνικών προσαρμογής, προετοιμάστηκαν οι παρατηρήσεις με τέτοιον τρόπο ώστε να είναι επιτρεπτή η οποιαδήποτε ανάλυση τους με τις τεχνικές μας. Αυτό πραγματοποιήθηκε με την τυποποίηση στις τιμές με τον μετασχηματισμό min-max και την ενσωμάτωση dataframe δηλαδή ενός πίνακα με στήλες τις μεταβλητές και γραμμές τις παρατηρήσεις, για να υποδέχονται τις μεθόδους που θα ακολουθούσαν.

Οι τεχνικές που χρησιμοποιήθηκαν ήταν οι k πλησιέστεροι γείτονες με εκμάθηση στο 35% του αρχικού συνόλου και για k=2, 4, 5, 7, 9 , η λογιστική παλινδρόμηση και τα στοχαστικά δέντρα που δεν απαιτούν διαχωρισμό σε εκπαιδευτικό σύνολο και μη. Αυξάνοντας το k διαφάνηκε ότι αυξάνεται η ακρίβεια του μοντέλου αν και η αύξηση αυτή φθίνει από την αλλαγή 7 γειτόνων σε 9 από ότι 5 σε 7. Η λογιστική παλινδρόμηση απέδωσε το μεγαλύτερο ποσοστό ακρίβειας αγγίζοντας το 100%.Επιπροσθέτως, η τελευταία τεχνική, αυτή των στοχαστικών δέντρων, παρατηρήθηκε ότι ερμηνεύει μεγάλο ποσοστό της μεταβλητότητας με αρκετά χαμηλά σφάλματα και αρκετά υψηλό ποσοστό ακρίβειας..

Βιβλιογραφία

Ξένη

Aggarwal1, C., Hinneburg, A., & Keim, D. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. *ICDT 2001*, pp. 420–434.

Baraniuk, C. (2015). The Cyborg Chess Players That Can't Be Beaten. *BBC Future, BBC*.

Bian, J., Tian, D., Tang, Y. & Tao, D. (2019). Trajectory Data Classification: A Review. *ACM Trans. Intell. Syst. Technol.* 10, (4), Article 33, p 33.

Breiman, L., & Cutler, (n.d.). A. *Random Forests*. Ανακτήθηκε από https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Chai, W. (2021). Big data analytics. *TechTarget*. Ανακτήθηκε από https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Charpentier, A. (n.d.). Predictive Modeling in Insurance, in the context of (possibly) Big Data. *Statistical & Actuarial Sciences Joint Seminar & Center of studies in Asset Management (CESAM)*. Ανακτήθηκε από http://freakonometrics.free.fr/slides_MS_stats_louvain.pdf

Clark, J. (2016). What is the Internet of Things (IoT)?. *IBM*. Ανακτήθηκε από https://www.ibm.com/blogs/internet-of-things/what-is-the-iot/

Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory*, 13, pp. 21–27.

Donges, N. (2019). A complete guide to the random forest algorithm. *Built in.* Ανακτήθηκε από https://builtin.com/data-science/random-forest-algorithm .

Dykes, B. (2017). Big Data: Forget Volume and Variety, Focus On Velocity. *Forbes*. Ανακτήθηκε από https://www.forbes.com/sites/brentdykes/2017/06/28/big-data-forget-volume-and-variety-focus-on-velocity/?sh=62c7fbb26f7d

Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis–Nonparametric Discrimination: Consistency Properties. Report Number 4, Project Number 21-49-004, *USAF School of Aviation Medicine*, Randolph Field, Texas.

Gohrani, K. (2019). Different Types of Distance Metrics used in Machine Learning. Ανακτήθηκε από https://medium.com/@kunal_gohrani/different-types-of-distance-metrics-used-in-machine-learning-e9928c5e26c7

Hechenbichler, K., & Schliep, K. P. (2004). Weighted K-Nearest-Neighbor Techniques and Ordinal Classification. *Discussion Paper* 399, SFB 386, Ludwig-Maximilians University Munich.

Hilbert, M. & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*. 332 (6025), 60–65.

Insurance Company Benchmark (COIL 2000) Data Set. (2000). *Machine learning repository, UCI*. Ανακτήθηκε από https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+(COIL+2000)

Lateef, Z. (2020). Comprehensive Guide To Logistic Regression In R. Ανακτήθηκε από https://www.edureka.co/blog/logistic-regression-in-r/

Le, J. (2020). Logistic Regression in R Tutorial. *Datacamp*. Ανακτήθηκε από https://www.datacamp.com/community/tutorials/logistic-regression-R

Looft, M. (2016). Is Machine Learning a Threat to the Actuarial Profession? *Earnix Blog, Earnix*.

Magoulas, R. & Lorica, B. (2009). *Introduction to Big Data*. Release 2.0. Sebastopol CA: O'Reilly Media (11). Ανακτήθηκε από https://www.oreilly.com/data/free/release-2-issue-11.csp

Mashey, J. R. (1998). *Big Data ... and the Next Wave of InfraStress*. Ανακτήθηκε από https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems

Open data, big data and open government: six subtypes of data. (2017). *datos.gob.es*. Ανακτήθηκε από https://datos.gob.es/en/noticia/open-data-big-data-and-open-government-six-subtypes-data

Press, G. (2013). A Very Short History Of Big Data. *Forbes*. Ανακτήθηκε από https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/?sh=48a6c5eb65a1

Richman, R. (2018). AI in Actuarial Science. *SSRN*. Ανακτήθηκε από https://ssrn.com/abstract=3218082

Sutton, R. & Barto, A. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.

Rider, F. (1944). *The Scholar and the Future of the Research Library*. New York City: Hadham Press.

Riley, J. (2020). AI and Machine Learning Usage in Actuarial Science. *The University of Akron – Williams Honors College.* Ανακτήθηκε από https://ideaexchange.uakron.edu/cgi/viewcontent.cgi?article=2186&context=honors_research_projects

Sirohi, K. (2018). K-nearest Neighbors Algorithm with Examples in R (Simply Explained knn). Ανακτήθηκε από https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples-in-r-simply-explained-knn-1f2c88da405c

Su, W. (2008). *Efficient Kernel Methods for Statistical Detection.* Waterloo, Ontario, Canada. Ανακτήθηκε από https://uwspace.uwaterloo.ca/bitstream/handle/10012/3598/wsuthesis.pdf?sequence=1&isAllowed=y

Swaminathan, S. (2018). *Logistic Regression - Detailed Overview*. Ανακτήθηκε από https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Tolles, J. & Meurer, W. J. (2016). Logistic Regression Relating Patient Characteristics to Outcomes. *JAMA.* doi: 10.1001/jama.2016.7653

Tree-Based Models. *Datacamp*. Ανακτήθηκε από https://www.statmethods.net/advstats/cart.html

Van der Putten, P., Ruiter, M., & Someren, M. (2000). CoIL Challenge 2000 Tasks and Results: Predicting and Explaining Caravan Policy Ownership. *Sentient Machine Research, Leiden Institute of Advanced Computer Science, Sociaal Wetenschappelijke Informatica*.

Volume, velocity, variety, veracity and value are the five keys to making big data a huge business. *BBVA*. Ανακτήθηκε από https://www.bbva.com/en/five-vs-big-data/

Walker, S.H. & Duncan, D.B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54 (1/2), 167-179.

Wang, Y. (2005). Statistical Models for High Thoughput Screening Drug Discovery Data (Ph.D. Thesis). *Department of Statistics and Actuarial Science*, University of Waterloo.

Yeo, N. (2017). Actuarial Profession in the Age of Artificial Intelligence and Process Automation. *Innovators & Entrepreneurs, SOA*. Ανακτήθηκε από www.soa.org/news-and-publications/newsletters/innovators-andentrepreneurs/2017/november/ei-2017-iss-61/actuarial-profession-in-the-age-of-artificial-intelligence-and-process-automation/

Yiu, T. (2019). Understanding Random Forest : *How the Algorithm Works and Why it Is So Effective.* Ανακτήθηκε από https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Thomas E., Wajid K., Paul B., (2016). *Big Data Adoption and Planning Considerations, Prentice Hall.*